Received 21 November 2023
Revised 16 December 2023
Accepted 18 January 2024

Portraying passenger travel patterns for Beijing public transit system with user profiling method

Ke Zhang
Faculty of Civil Engineering and Geosciences, TU Delft, Delft, The Netherlands, and
Ailing Huang
School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China

Abstract

Purpose – The purpose of this paper is to provide a guiding framework for studying the travel patterns of PT users. The combination of public transit (PT) users' travel data and user profiling (UP) technology to draw a portrait of PT users can effectively understand users' travel patterns, which is important to help optimize the scheduling of PT operations and planning of the network.

Design/methodology/approach – To achieve the purpose, the paper presents a three-level classification method to construct the labeling framework. A station area attribute mining method based on the term frequency-inverse document frequency weighting algorithm is proposed to determine the point of interest attributes of user travel stations, and the spatial correlation patterns of user travel stations are calculated by Moran's Index. User travel feature labels are extracted from travel data containing Beijing PT data for one consecutive week.

Findings – In this paper, a universal PT user labeling system is obtained and some related methods are conducted including four categories of user-preferred travel area patterns mining and a station area attribute mining method. In the application of the Beijing case, a precise exploration of the spatiotemporal characteristics of PT users is conducted, resulting in the final Beijing PTUP system.

Originality/value – This paper combines UP technology with big data analysis techniques to study the travel patterns of PT users. A user profile label framework is constructed, and data visualization, statistical analysis and K-means clustering are applied to extract specific labels instructed by this system framework. Through these analytical processes, the user labeling system is improved, and its applicability is validated through the analysis of a Beijing PT case.

Keywords User profiling, Big data analysis, Travel pattern, Station area attributes, Public transit

Paper type Case study

1. Introduction

Urban public transit (PT) is a crucial component of the urban transportation system, providing basic travel services for the public. With the advancement of internet technology, passengers are increasingly using electronic payment methods, enabling complete recording of individual travel data, such as boarding and alighting points, time and payment methods.

This work was supported jointly by the National Key Research and Development Program of China (Grant No. 2018YFB1601200) and the National Natural Science Foundation of China (Grant No. 72371021).

Smart and Resilient Transportation
Vol. 6 No. 1, 2024
pp. 50-71
Emerald Publishing Limited
e-ISSN: 2632-0495
p-ISSN: 2632-0487
DOI 10.1108/SRT-11-2023-0014

This data can be used by the government or PT companies to provide better operation plans and services, improve passenger travel efficiency, increase the possibility of passengers using PT and alleviate urban traffic congestion.

In the era of big data, user profiling (UP) has become an essential concept in the field of data analysis and marketing. Profiling is the process of labeling user information and abstracting specific information into a label to describe a user virtually in a network environment. UP technology is widely applied in various fields, with the most common application being in the e-commerce sector (Badriyah *et al.*, 2017; Fijałkowski and Zatoka, 2011). Also, UP assists businesses or systems in gaining a better understanding of user needs and providing personalized services and experiences to achieve precision marketing, including recommendation systems (OUAFTOUH *et al.*, 2019; Yu *et al.*, 2006) and pricing systems (Hu *et al.*, 2022). The combination of PT transit data and UP technology is significant for understanding user travel patterns, providing targeted services to users and more. UP technology can use PT big data to analyze PT user travel patterns, provide reference and guidance for urban planning and PT construction and realize intelligent decision-making for transportation systems.

Therefore, studying public transit user profiling (PTUP) not only has practical significance but also has certain research value in theory. The PTUP constructed in this paper in the spatial and temporal dimensions can help researchers classify users and can be more targeted when using PT travel data to mine user patterns in subsequent studies, making data-based UP more comprehensive. In short, a complete PTUP can greatly enhance the understanding of user needs and preferences, leading to better PT services and policies. PTUP can also assist PT departments in fine-grained demand forecasting. It can precisely identify popular travel times, travel routes, etc. within the spatiotemporal dimensions, facilitating planning departments to respond more effectively to unforeseen events and emergencies.

However, the research on PTUP also faces challenges and difficulties. Due to the diversity and complexity of user data, defining and establishing a unified standardized, accurately depicting UP framework to describe user attributes and patterns is challenging. PT travel involves multiple factors, such as travel purposes, frequency and timing. Obtaining and analyzing these factors accurately and converting them into specific UP information requires comprehensive consideration of data collection, algorithm models and analytical methods.

In this paper, a three-level methodology is presented for developing a PTUP standard framework that includes spatiotemporal feature labels. Corresponding mining methods are proposed or defined for the feature labels within the framework. The framework and label definition methods presented in this paper are universal and can serve as guiding systems for future research on PT user travel profiles in different cities or among various groups. This holds significant research value within the field.

The remaining of this paper is organized as follows. Section 2 reviews previous works related to PT data analysis and UP technology. Section 3 establishes the standard UP framework for PT users. Methods for exploring station spatial distribution and area attributes are proposed in section 4. Section 5 is a case study using Beijing PT users as an example. Section 6 concludes the work and gives potential topics for future research.

2. Related work

Recently, more work and research have been done including mining passenger travel patterns through PT transit data, the application of UP technology and methods used when constructing UP framework, which can be summarized as follows.

A substantial amount of work has been done in using various data sources for the analysis of PT travel behavior. Big data analysis technologies are applied in these fields to

explore the travel patterns of PT users (Zannat and Choudhury, 2019). More specifically, VA Van Oort and Cats (2015) demonstrated, through two cases in The Netherlands and Sweden, how to use emerging big data sources to support the PT industry in addressing significant challenges such as improving efficiency and enhancing satisfaction. Yu *et al.* (2006) used this technology to analyze the spatial-temporal patterns of the Spring Festival travel rush in China and proposed to establish a unified real-time traffic platform to alleviate problems caused by the rush. Li *et al.* (2016) used large-scale taxi trajectory data to discover the passenger travel patterns, also proposed an information framework to preprocess the data, developed a distribution model to explore the rules of citizen travel accurately and presented a parallel clustering approach to obtain the characteristic changes of inhabitant travel at different periods of the day. Xia *et al.* (2021) applied big data visualization in the analysis of passenger flow data within urban rail transit networks, highlighting its effectiveness in characterizing and explaining complex data, providing valuable insights for operational evaluation, planning and design.

UP is gradually being applied nowadays. Many university libraries also use UP technology to analyze and optimize the allocation of book resources, among other aspects (Zhiyuan *et al.*, 2017; Jomsri, 2014; Shirude and Kolhe, 2014). Thompson (Badriyah *et al.*, 2017) applied UP to precision services in libraries, becoming the most suitable library service marketing tool. In the medical area, Thompson *et al.* (2017) collected a large amount of relevant data from elderly patients with chronic diseases and built UP based on this data, assisting medical institutions or companies in developing health-care service systems for elderly patients with chronic diseases. UP is also applied in the field of smart geriatric care to refer to corresponding care measures and support (LeRouge *et al.*, 2013). However, the application of UP technology in the transportation domain is not commonly seen. Only a few scholars have begun to research the application of UP in the field of PT. Gutierrez *et al* LeRouge *et al.* (2013) used data obtained from an automatic fare collection system to study the usage patterns of PT for tourists traveling to Costa Daurada. Li and Tang (2020) used clustering methods to categorize PT users in Portuguese urban areas, mentioning that this could help create UP without delving into a detailed analysis.

The process of constructing UP framework is to categorize or label user information (Gutiérrez *et al.*, 2020). Therefore, the aim of the related methods and techniques is a process of extracting labels. Sugiyama *et al.* (2004) researched adaptive search techniques, using user browsing history to construct user profiles for obtaining search results that cater to user preferences. Amoretti *et al.* (2017) adopted the K-means algorithm to divide the user into distinct classes, while this method is only able to generate keywords for a group of data. Corney *et al.* (2011) took supervised machine learning methods to build Web UP, extracted content feature, pattern feature and term feature. Supervised machine learning was easily scalable in this research, but it was limited to academic data and cannot be applied to general web UP. Tang *et al.* (2010) used a statistical model to build UP based on computer system logs. A statistical-based approach is widely used but only reflects the history or in-time behavior of users.

There is extensive research on user travel patterns. Rafiq and McNally (2021) conducted a study on the activity-travel patterns of PT users using latent class analysis, classifying them into five distinct categories. The characteristics of PT users typically involve spatiotemporal features (Hasan *et al.*, 2013). Verma *et al.* (2021) conducted a case analysis by examining millions of daily operational trajectories of the primary subway services in the Greater London area, revealing diverse spatiotemporal commuting patterns among users. Zhao *et al.* (2017) also used user travel data to define individual travel patterns in terms of three aspects: temporal, spatial and spatiotemporal. Based on these characteristics, they

categorized subway passengers into four groups and identified passenger groups that typically take the subway for one leg of the journey and return by bus in a single trip, among other findings. Liu *et al.* (2020) used spatiotemporal data of users to explore individual variations in travel patterns, revealing the role of age and gender in the variability of public transportation usage patterns.

Using a single method to mine labels has its limitations. In this paper, the statistics model approach and the hierarchical clustering method can be used together to extract PT users' fundamental travel behavior labels. Mining the geographical location and regional attributes of PT stations can analyze the travel purposes of users and, thus, determine the travel characteristics. Although research on regional functional attributes based on stations is relatively limited, many scholars have proposed other types of functional zone division methods, which also have reference significance. Point of interest (POI) data are mostly used to identify urban function areas (Hu and Han, 2019; Wang *et al.*, 2021; Bao *et al.*, 2020). Term frequency-inverse document frequency (TF-IDF) algorithm is commonly used to weigh the importance of different POIs among all (Miao *et al.*, 2021; Mishra and Urolagin, 2019).

To sum up, using big data analytics to study user travel characteristics in the PT field has gained significant attention. However, these studies often lack a unified framework or standardized guidelines for conducting a comprehensive analysis of user travel characteristics, resulting in fragmented research efforts. In addition, while UP technology is well-established and widely used, there exists a noticeable research gap in its application within the PT. Furthermore, the review of UP techniques reveals that current profiling methods tend to rely on single approaches for extracting user labels, leading to limitations in the labeling process. Taking into consideration the conclusions drawn from the literature review, this paper aims to establish a standardized framework for PTUP to guide the utilization of big data analytics in analyzing the spatiotemporal features of user travel. Multiple methods will be used within this framework for comprehensively extracting user labels, addressing the existing gaps in the field.

3. Constructing up framework for public transit users

This section is to construct a user profile of PT users. A three-layer approach has been enhanced to preliminarily construct a PTUP label framework that includes factual labels, model labels and predictive labels. This section explains these labels and introduces relevant computational indicators, which are significant for guiding the subsequent exploration of the travel patterns of Beijing's PT users.

3.1 Public transit user label classification

The primary objective of UP is to develop comprehensive profiles for PT users. A UP comprises symbolic labels that describe various dimensions of user's patterns, addressing the challenge of description and categorization. However, in practical applications, these labels also need to capture the associations between data, leading to the design of a label framework that handles data associations. UP labels can be categorized into basic and behavioral attribute labels. Basic property labels encompass objective attributes, such as gender, age and occupation, which are not self-expressed attributes. On the other hand, behavioral attribute labels capture subjective attributes influenced by personal preferences and style patterns.

In line to construct a UP system for PT users, this section proposed a three-level hierarchical approach. The systematic architecture consists of basic property labels, model labels and predictive labels, each serving a specific purpose. Basic property labels are derived from preliminary statistical analysis of the original database and provide specific information about the PT user, such as smart card (IC card) number, payment method and

payment amount. Model labels, generated through model analysis based on basic property labels, do not have a direct correspondence to the original data. They require the definition of rules and the association of data generated by label instances combined with various mathematical methods such as correlation analysis, and cluster analysis to generate model labels. For PT users, examples of model labels include travel patterns, job stations, residence stations and so on. Predictive labels are generated based on model labels and involve inferring additional attributes from the existing data. For instance, predictive labels may infer a user's economic situation or job type based on their travel preference areas, as identified by the model labels.

Within each label category, further subdivision into primary, secondary and tertiary labels is possible. For example, the model labels focus on mining PT user travel patterns, and this paper specifically constructs model labels for user travel patterns, which will be explored from two dimensions: time and space.

3.2 Public transit user labels of temporal travel pattern

A total of 14 temporal dimension labels are defined and categorized into travel time patterns, travel activity and travel loyalty. The definitions of these labels encompass commonly observed temporal aspects of travel characteristics. The PTUP emphasizes their distinctive travel time patterns, particularly the frequency of trips taken within a specific period and their preference for PT route selection. Travel activity pertains to the average number of PT trips taken by users during different periods, including weekdays and holidays. Travel loyalty represents the level of consistency and reliance of users on PT within a specified period, as indicated by factors like the average number of consecutive travel days in their travel cycle. The period or cycle refers to a continuous segment of time, which can be a week, a month, a year, etc. Table 1 provides a comprehensive list of these labels and their corresponding definitions.

3.3 Public transit user labels of spatial travel pattern

A total of 11 labels are defined and categorized into work/dwell place, travel preference, travel habits and spatial correlation related to stations, highlighting PT users' distinct preferences for travel areas, stations, residence stations and workplaces. The spatial dimension encompasses the attributes of POI in the vicinity of the travel stations, enabling the identification of users' residential and workplace attributes, as well as their travel habit preferences. Table 2 presents the specific labels and definitions for these spatial attributes. The No. is following the Table 1.

3.4 Public transit user labeling framework

In addition to the travel feature labels based on time and spatial dimensions, it is necessary to supplement the basic label framework containing basic attributes, model labels and prediction labels. The preliminary construction of the public transit user labeling (PTUL) framework is illustrated in Figure 1.

The PTUL framework constructed in this section comprehensively builds the profile of PT users from three perspectives. This study specifically focuses on the temporal and spatial patterns of PT. Simple names are proposed for basic attributes and prediction labels to showcase the completeness of the PTUL. Detailed exploration work is not covered in this study.

4. Methods for label extraction

This section aims to introduce the methods used for extracting certain labels from Figure 1. Hierarchical clustering will be applied to find the PT user travel preference. For the spatial

Labels	No.	Name	Definition
Travel time patterns	1	Working day travel time patterns	Onboarding time of working day users
	2	Weekend travel time patterns	Onboarding time of weekend users
Travel activity	3	Travel activity during a cycle	Average number of trips during a cycle
	4	Travel activity on working days	Average number of trips on working days
	5	Travel activity during morning peak on working days	Average number of trips during morning peak on working days
	6	Travel activity during evening peak on working days:	Average number of trips during evening peak on working days
	7	Travel activity during off-peak on working days	Average number of trips during off-peak on working days
	8	Travel activity on weekends and holidays	Average number of trips on weekends and holidays
	9	Travel activity during morning peak on weekends and holidays	Average number of trips during morning peak on weekends and holidays
	10	Travel activity during evening peak on weekends and holidays	Average number of trips during evening peak on weekends and holidays
	11	Travel activity during off-peak on weekends and holidays	Average number of trips during off-peak on weekends and holidays
Travel loyalty	12	Working day loyalty	Average number of consecutive travel days on working days
	13	Weekend and holiday loyalty	Average number of consecutive travel days on weekends and holidays
	14	Travel cycle loyalty	Average number of consecutive travel days in a cycle

Source: Created by authors

Table 1.
PT user travel patterns labels based on time dimension

Labels	No.	Name	Definition
Work/dwell place	15	Residence and workplace station	Inferred based on the boarding and alighting stations within the user's commuting time
	16	Region (administrative area)	The administrative area where the residence and workplace stations are located
	17	POI attributes	Attributes of the area where the residence and workplace stations are located
Travel preference	18	Preferred travel stations	High-frequency travel stations for the user
	19	Preferred travel origin-destination (OD) chains	High-frequency travel OD chains for the user
	20	Preferred travel region	The region where the high-frequency travel station for the user is located
	21	Preferred travel business district	Obtained through cluster analysis of high-frequency travel stations
Travel habits	22	Consumption and entertainment	The POI attribute for user's alighting station is consumption and entertainment
	23	Education and training	The POI attribute for user's alighting station is education and training
	24	Medical services	The POI attribute for user's alighting station is medical services
Spatial correlation	25	Station spatial distribution patterns	Derived from Moran's index analysis of the station

Source: Created by authors

Table 2.
PT user travel patterns labels based on spatial dimension

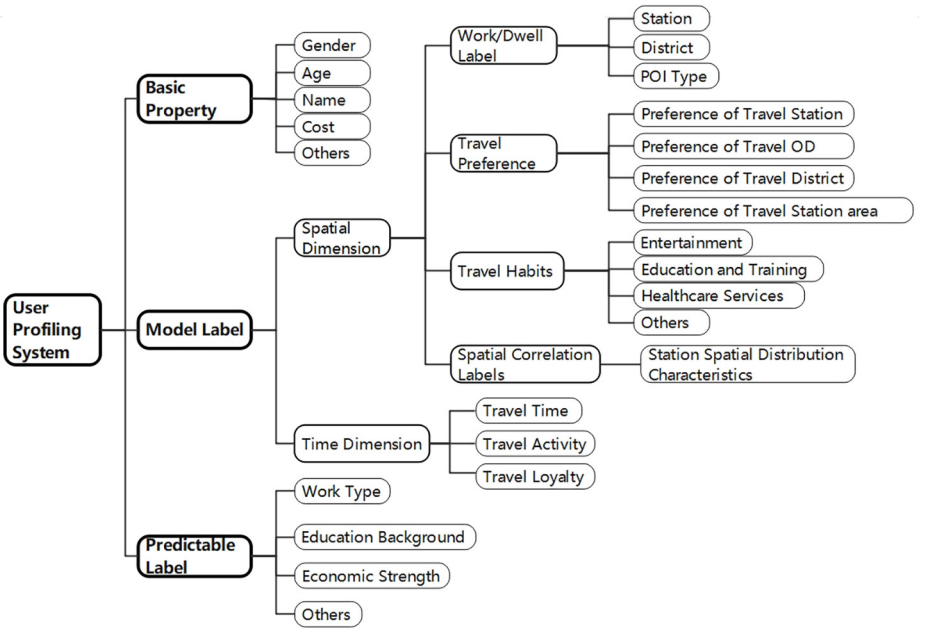


Figure 1.
Public transit user labeling framework

Source: Created by authors

and regional patterns of users' boarding and alighting stations, the correlation of PT stations is analyzed and proposes a method to define the attributes of station areas.

4.1 Travel preference clustering

PT users' travels to different stations can reflect their travel preferences. This paper chooses to use the hierarchical clustering algorithm to cluster and classify users' travel stations. The basic idea of hierarchical clustering is to calculate the similarity between nodes using a certain similarity measure, rank the nodes based on their decreasing similarity and gradually reconnect the nodes. The main steps include data preprocessing, similarity calculation, cluster merging and dendrogram visualization. Compared to other clustering methods, the hierarchical clustering algorithm makes it easier to define distance and rules for similarity, has fewer constraints and offers more advantages. For example, K-means clustering requires setting the number of clusters in advance and setting it too high or too low can affect the clustering results. In contrast, hierarchical clustering can automatically cluster without the need to predefine the number of clusters. In this study, it is not known how many categories of user-traveled station preferences exist before clustering. Therefore, the hierarchical clustering algorithm is chosen.

When using hierarchical clustering to cluster user travel stations, the key is to determine the feature values. This study chooses to define four user travel station features, including user's weekday average station dwell time for work, weekday average station dwell time for off-duty, weekend average station dwell time and daily average station passenger flow. To make the obtained station feature results more comparable, data normalization is performed in the data preprocessing step to transform the feature values to the same scale before

similarity calculation. Objects with higher similarity are merged into a category, and a dendrogram is drawn to visually display the clustering results.

4.2 Spatial correlation analysis for public transit stations

The Moran's Index is a useful tool for evaluating the spatial autocorrelation of geographic spatial data. When analyzing the spatial distribution patterns of PT stations, Moran's Index can help understand the spatial distribution patterns of stations and the organizational form of urban PT. Moreover, applying Moran's Index to analyze the spatial distribution patterns of the boarding and alighting stations that users frequently travel to can enable the identification of spatial clustering or dispersion phenomena and comprehend the spatial correlation of stations in the area. This is of great significance for optimizing the layout of PT networks, improving system efficiency and meeting the travel needs of users.

The Moran's Index consists of the global index and the local index. The significance test in the global Moran's Index is used to indicate the spatial correlation of data in the entire spatiotemporal system of the road network, while the local Moran's Index mainly reflects the spatial correlation of data in a certain area. When interpreting Moran's Index, a significance evaluation test should be done to ensure that the data provide sufficient evidence, usually performed as p-value and z-value. Normally, the global Moran's Index of a region is first calculated to determine whether there is clustering or an outlier in space. If spatial autocorrelation is present in the global index, the local autocorrelation is then calculated and will indicate where the outliers or clustering occur. The mathematical formula for calculating the Moran's Index is as follows:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{1}$$

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{i,j} \tag{2}$$

The formula (1) represents the calculation of the Moran's Index I, while the z_i is the deviation of the attribute of element i from its mean value $\left(x_i - \hat{X}\right)$, $\omega_{i,j}$ is the spatial weight between elements i and j, n is equal to the total element. In the formula (2), the aggregation of all spatial weights S_0 can be obtained. When applying the Moran's Index to determine the spatial correlation analysis for PT stations, the elements i and j refer to stations, and the spatial weight $\omega_{i,j}$ represents the traffic flow between station i and j.

4.3 Mining point of interest attributes of station area

POI data play an important role in mining urban area attributes. Each record of POI data includes four pieces of information: name, category, address and specific location (longitude and latitude). In this study, the TF-IDF weighted algorithm is used to weigh the station blocks based on POI data, and the functional attributes of each station are ultimately determined according to the weight.

TF-IDF is a common weighting technique algorithm used in information retrieval and data mining (Badriyah *et al.*, 2017). The TF-IDF is a statistical method to assess the importance of a word to a document set or one of the documents in a corpus. The

importance of a word increases proportionally with its number of occurrences in a document but decreases inversely with its frequency in a corpus. Therefore, the core idea of TF-IDF is that if a word or phrase appears frequently in an article and rarely in other articles, it is considered to have good category differentiation ability and is suitable for classification.

When using POI data to determine station attributes, each POI data can be considered as a word, all POI data contained in a station range as a document and all stations and their covered POI data as the overall corpus. Thus, the formula for calculating TF-IDF values for station area attributes is shown in the formula (3):

$$TF_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^{k} n_{i,j}} \tag{3}$$

The numerator $n_{i,j}$ is the number of POIs of type i within the functional area of station j. The denominator $\sum_{i=1}^{k} n_{i,j}$ is the number of all type of POI contained at station j, and k represents the POI data type. The following formula (4) indicates the calculation of IDF_i, which uses the logarithmic function:

$$IDF_i = log \frac{A}{|\{j : t_i \in d_j\}| + 1} \tag{4}$$

Regarding the formula (4), first, for words t with particularly high frequency that appears in almost every document, the number of documents containing t in the document set is approximately equal to the total number of documents in the set, i.e. $N/n = 1$ (N/n is constantly greater than 1). When using only TF, the weight of t is large, but it cannot distinguish between important and unimportant words, which does not meet our expectations. However, after applying IDF (with logarithmic function), the weight of t calculated by TF-IDF is 0, as $log(1) = 0$, which satisfies expectations. Second, the use of a logarithmic function prevents weight explosion. If certain words appear in only one or a few documents, the IDF will be very large without the logarithmic function (as the denominator is too small), thus, affecting their weight. Therefore, the logarithmic function is used to mitigate this effect.

Let A represent the number of PT stations, and $|\{j : t_i \in d_j\}|$ represent the number of stations that contain POI data type i. It should be noted that if there is no POI of this type in the entire document, this number is 0, and one needs to be added to the denominator to prevent it from becoming 0. The final weight of the TF-IDF value is calculated as follows, as shown in formula (2):

$$TF - IDF = TF_{i,j} * IDF_i \tag{5}$$

The purpose of adjusting the weights is to highlight important words and suppress minor words, which has its rationale. Essentially, IDF is a type of weighting that tries to suppress noise. For instance, if a particular station has a high frequency of medical-type POI data, i.e. the value of TF is large, and its frequency of occurrence in other stations is small, then this station has obvious medical attributes. Conversely, if the frequency of occurrence in other stations is high, it is difficult to determine the medical attributes of the station.

The distance between PT stations typically falls within the range of 500 meters to 1,500 meters, with an average distance of about 1 kilometer. For PT users, the generally acceptable

walking distance is usually between 500 meters and 800 meters. Considering both factors, this study uses PT stations as points and draws a circle with a radius of 500 meters. The area covered by the circle represents the functional range of the stations. If a certain POI is located within or equal to 500 meters from the center of the station, it belongs to the zone of that station; otherwise, it does not belong to the zone.

Using the application programming interface interface of Baidu Maps to obtain POI data in Haidian District, Beijing, Mingguangcun Station is taken as an example, and the corresponding results are obtained by using the TF-IDF algorithm, as shown in Table 3. According to the table, the top three categories are Real Estat, Companies and Enterprise and Education and Training, accounting for 27.067700%, around 15% and around 15%, respectively. Therefore, it can be inferred that Mingguangcun Station is a comprehensive residential area, with a mix of residential, commercial and educational establishments. It should be noted that if a POI data type is not present in a document, the denominator is zero, which can be prevented by adding one to the numerator. The final weight TF-IDF value is then obtained.

5. Case study: analysis of public transit user profiling in Beijing

In this section, one-week swipe card data from Beijing PT users are collected. Guided by the PTUL framework constructed in Section 3 and the method proposed in Section 4, parts of labels will be selected for practical analysis to uncover Beijing PT user travel patterns and obtain a final PTUP in Beijing.

5.1 Data description

This study collected travel data of Beijing PT users for one consecutive week from March 24 to March 30, 2018, including bus and subway. Beijing PT adopts the entry-exit charging system which can capture information about users, their boarding and alighting stations,

No.	POI type	TF value	IDF value	TF-IDF weight	Weighting %
1	Real estate	0.231735	0.364663	0.084505	27.067700
2	Corporate enterprise	0.134703	0.370277	0.049878	15.976200
3	Education	0.111872	0.405578	0.045373	14.533300
4	Government agencies	0.076484	0.386730	0.029579	9.474300
5	Restaurant	0.057078	0.404266	0.023075	7.391000
6	Life services	0.045662	0.380900	0.017393	5.571000
7	Finance	0.023973	0.521276	0.012496	4.002700
8	Entrance/exit	0.222603	0.041496	0.009237	2.958700
9	Shopping	0.019406	0.409978	0.007956	2.548400
10	Hotel	0.012557	0.528769	0.006640	2.126800
11	Media	0.009132	0.534033	0.004877	1.562100
12	Tourist attractions	0.007991	0.529933	0.004235	1.356400
13	Transportation facilities	0.009132	0.366660	0.003348	1.072500
14	Medical	0.006849	0.463070	0.003172	1.015900
15	Recreation	0.005708	0.469616	0.002680	0.858600
16	Beauty	0.004566	0.520705	0.002378	0.761600
17	Administrative landmark	0.004566	0.480403	0.002194	0.702600
18	Sports	0.013699	0.129088	0.001768	0.566400
19	Gate address	0.001142	0.654678	0.000747	0.239400
20	Auto service	0.001142	0.586512	0.000670	0.214500
21	Natural features	0.000000	0.997256	0.000000	0.000000

Source: Created by authors

Table 3. An example of the TF-IDF value of Mingguangcun station in Beijing

station numbers, travel modes, travel time and travel line numbers. The data source is electronic swipe card data such as user IC card and cell phone payment except for temporary ticketing data through cash payment, and some data fields are shown in Table 4. In the column mode, GJ represents for bus and DT for subway.

5.2 Extraction of public transit user labels and travel patterns

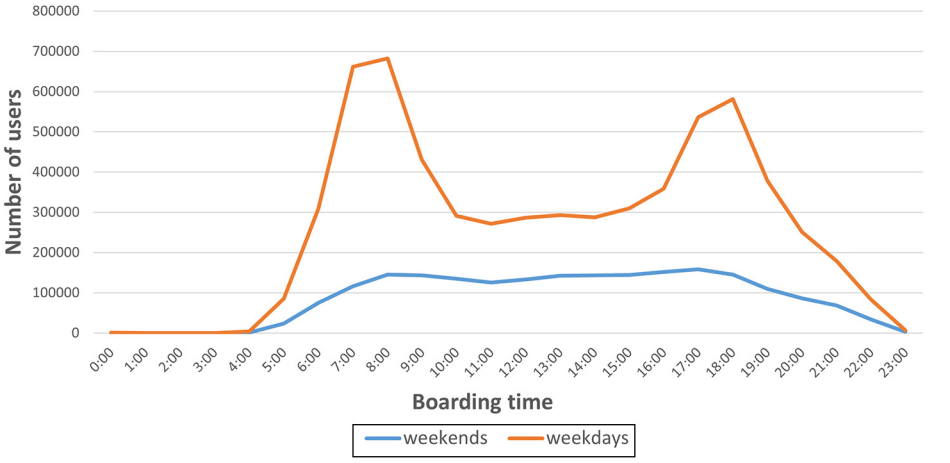
5.2.1 Travel time labels. The boarding periods of PT users on weekends and weekdays are extracted separately, and the results are shown in Figure 2.

It can be observed that there are differences in travel time patterns for PT users between weekends and weekdays. During weekdays, users exhibit distinct peak periods in the morning (7–9 a.m.) and evening (5–7 p.m.), corresponding to their commute to and from work. This pattern reflects the regularity of travel time based on user's work schedules. On the other hand, during weekends, the number of users remains relatively stable throughout different periods without clear peak periods. This indicates that users have more flexibility in choosing travel times during weekends and do not have fixed time requirements for their trips.

User ID	Mode	Line no.	Boarding station no.	Boarding station	Boarding time	Alighting station no.	Alighting station	Alighting time
***85 5f	GJ	666	27	Chen Jialing	2018032 4191501 #GJ	31	Dabei Yao east	20180324 192503
***a b62	DT	13	27	Wu Daokou	2018032 4091400 #DT	33	Haidian Huangzhuang	20180324 093208
***a e77	GJ	804	35	Xidajie Road East	2018032 4092901 #GJ	40	Ritan road	20180324 100320

Note: *** = Omit the number
Source: Created by authors

Table 4. Example of some of the user travel data fields



Source: Created by authors

Figure 2. Distribution of user travel period in a day

5.2.2 Travel activity labels. According to the user ID number information in the data, the frequency of users in a week, every seven times for a travel frequency interval, which corresponding to once a day, seven days a week. results for the number of users in the interval as shown in Figure 3.

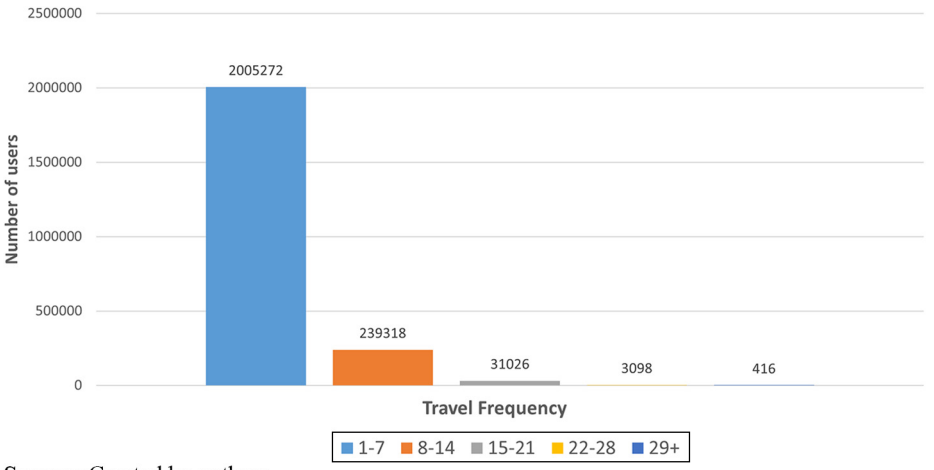
Most PT users have a weekly travel frequency ranging from one to seven trips, with an average of at least one trip per day. The second-highest group consists of users with a travel frequency of more than 8 to 14 trips, averaging at least two trips per day. This indicates that these users have a higher level of travel activity throughout the week. As the frequency range increases, the number of users sharply decreases, indicating that there are very few users who maintain a consistently high level of travel activity across all periods.

5.2.3 Preference of travel station labels. When extracting the user travel station features, the user's transfer stations are not considered, only the boarding and alighting stations and the corresponding latitude and longitude coordinates of the stations are retained, and the user travel stations are visualized. Figure 4 shows the passenger flow distribution of user travel in different periods.

In the visualized results, each bar region represents a station, where the height and color of the bar indicate the passenger flow. The higher the bar and the darker the color, the higher the passenger flow at the station. From the results, whether on weekends or weekdays, the distribution of passenger flow for boarding and alighting stations is remarkably similar. The areas with high passenger flow are primarily located in Chaoyang District, Tongzhou District, Dongcheng District, Xicheng District and Haidian District, indicating that PT users in Beijing are most active in these five districts. Among them, the stations with the highest passenger flow are consistently located in Chaoyang District, specifically at Dawanglu Station, Guomao Station and Qingnianlu Station.

5.2.4 Preference of travel origin-destination labels. Disregarding the interchange stations, extracting the travel origin-destination (OD) of weekday and weekend passengers and the corresponding OD station latitude and longitude data. The results of the top 1,000 travel ODs on weekdays and weekends are shown in Figure 5.

Similar to the distribution of preference of travel station, the travel OD during weekdays and weekends is mostly concentrated in Chaoyang District, Tongzhou District, Dongcheng



Source: Created by authors

Figure 3.
Distribution of travel frequency in a week

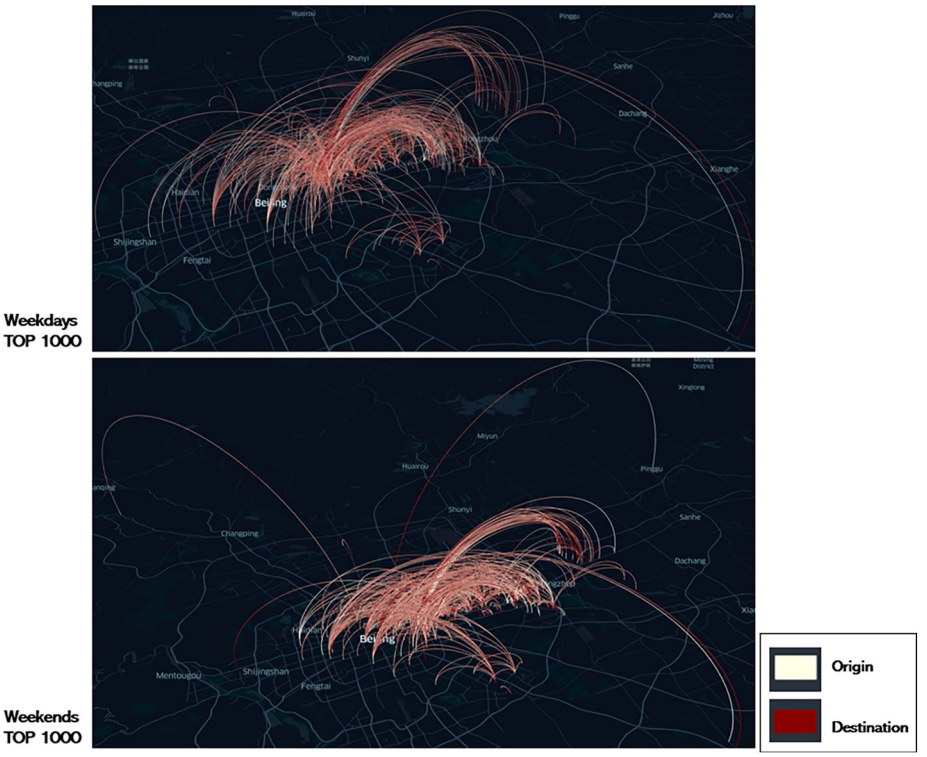
Figure 4.
PT user traffic flow distributions on weekdays or weekends

Source: Created by authors

District, Xicheng District and Haidian District. However, a notable difference is that there is a significant increase in "Urban-Suburban" travel during weekends, indicating that a majority of PT users choose to travel to the suburbs during the weekends. This reflects a change in user travel habits from weekdays to weekends. The most frequent OD during weekdays is from Sihui Hub Station to the intersection of Beiyuan Road in Tongzhou, while during weekends, it is from Changying to the Institute of Materials.

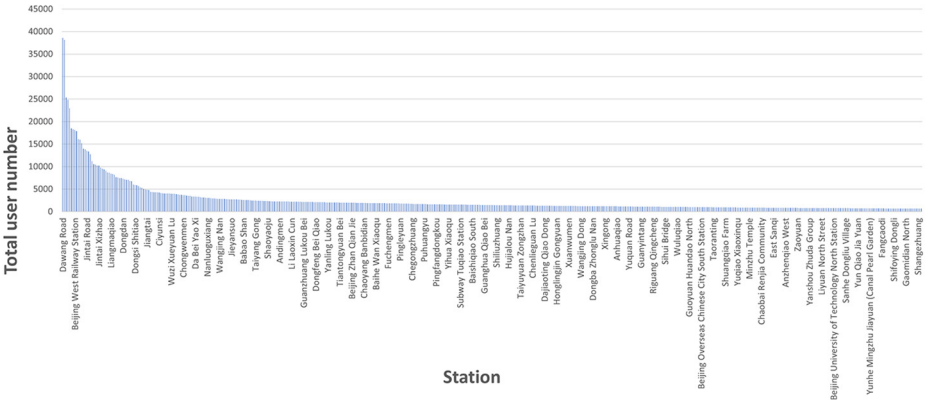
5.2.5 Work and dwell labels. For users with high travel loyalty and activity, their residential and workplace stations can be derived by analyzing the OD stations where they frequently travel on weekdays. Typically, the morning peak period on weekdays is the time when users depart from their homes to their workplaces. From subsection 5.2.1, the morning peak period for Beijing PT users to travel is from 7:00 a.m. to 9:00 a.m., while the evening peak time is from 5:00 p.m. to 7:00 p.m. The evening peak is usually when users depart from work to go home. To obtain information about the user's residence and workplace stations, this section extracts the morning and evening peak travel stations of each user for five consecutive weekdays and sorts them to identify the most frequently traveled stations. The results of the top 500 residential and workplace stations, and also the corresponding trend curves, based on the number of users traveling, are shown in Figures 6 and 7.

Fitting the curves in Figures 6 and 7 results in power-law functions as shown in the following equations (6) and (7), respectively. A significance test for the fitted power-law functions is conducted, resulting in R-squared values of 0.8451 and 0.8426 for the respective equations. These values are close to 1, indicating a high level of reference value for the fitted power-law equations:



Source: Created by authors

Figure 5.
Travel OD on weekdays and weekends, top 1000



Source: Created by authors

Figure 6.
Residential station and user numbers on weekdays

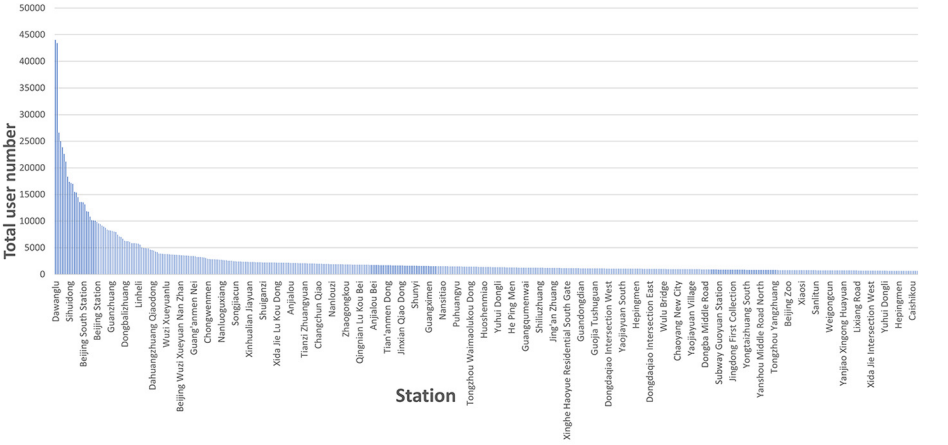


Figure 7.
Workplace station and user numbers on weekdays

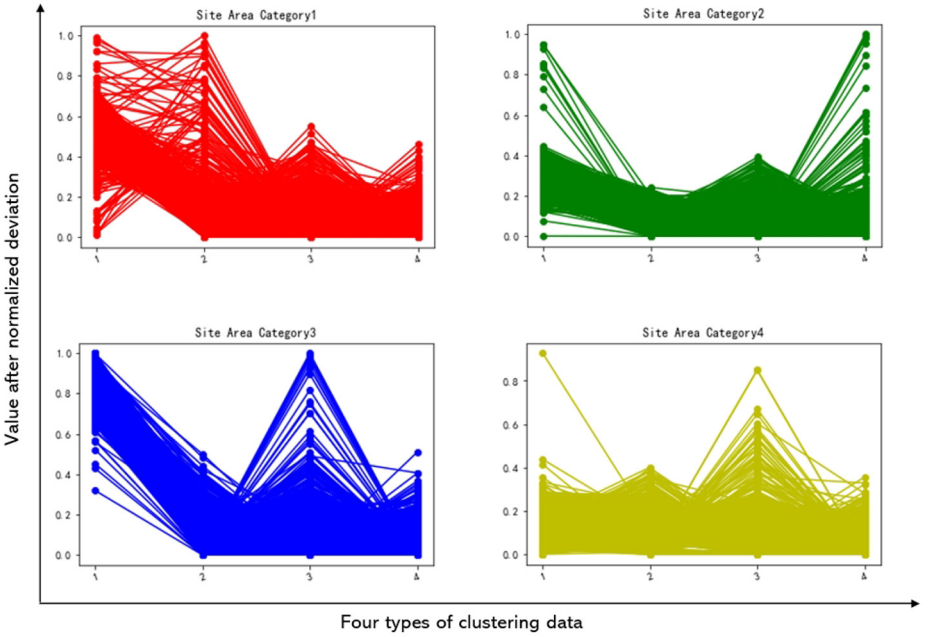
Source: Created by authors

$$y = 101587x^{-0.78} \tag{6}$$

$$y = 116739x^{-.0812} \tag{7}$$

Therefore, the relationship between the residential and workplace stations of Beijing PT users and the flow at each station exhibits a characteristic of a power-law distribution. The overall shape of the distribution is a continuously decreasing curve, starting from a peak and rapidly declining, followed by a long tail. The residential or workplace station as a whole shows significant heterogeneity, with most stations having relatively low user counts, and a vast difference between the minimum and maximum values. One of the properties of a power-law curve is that a few factors often play a decisive role in determining an outcome, while the majority of factors are inconsequential. Therefore, the stations that represent the main residential and workplace areas for Beijing PT users are the top-ranked stations. The top three stations with the most users are Dawang Road, Guomao and Jiulong Shan stations in Chaoyang District. Dawang Road and Guomao, have accumulated over 40,000 users over 5 days as workplace stations, and they also have close to 40,000 users as residence stations, while the third-ranking station, Jiulong Shan, has around 25,000 users for both types of stations. The number of the remaining stations gradually decreased, with the lowest being Dongsi Shitiao Station, which still has over 5,000 users. Meanwhile, the stations with a high volume of users boarding and alighting on weekdays are concentrated in Chaoyang District.

5.2.6 Preference of travel area labels. When studying PT user preferences of travel area types, hierarchical clustering is performed on four types of data including user's weekday average station dwell time for work, weekday average station dwell time for off-duty, weekend average station dwell time and daily average station passenger flow. The clustering results of the four-class travel stations are shown in Figure 8. The site area category represents the area category to which these sites (stations) belong. The vertical axis a represents value after normalized deviation processing, and the Numbers 1, 2, 3 and 4 on the horizontal axis correspond to the four types of data for clustering (in order).



Source: Created by authors

Figure 8. Clustering results of preferred travel areas for PT users

As shown in Figure 8, the clustering result of station area category 1 shows that both the average user's travel time during working hours and off-working hours on weekdays in each station are long, the weekend dwell time is in the middle, but the average daily traffic flow is not high. This station area is neither a residential-class nor a work-class station area, but more inclined to a comprehensive station area. Users who travel to this type of station area do not have specific time restrictions, such as some mobile workers, and have travel flexibility. For station area category 2, the average user's travel time during working hours on weekdays in each station and the average daily traffic flow is very high. It can be classified as a work-class station area that is preferred for travel. The users who travel to this category of station usually travel for work purposes. Station area category 3 indicates that the average user's travel time during working hours on weekdays and the average user's dwell time on weekends are high, which belongs to the work/commercial class station area that is preferred for travel. Users who travel to this station area choose to shop, entertain and consume on weekends and some staff work in the commercial area and travel for work. The result of station area category 4 reveals that the overall travel time patterns and the average daily traffic flow are not high. Only the average user's dwell time on weekends is higher, indicating that users rarely reach this area, which can be classified as a preference for remote areas. For instance, users go to the suburbs for fun on weekends.

In summary, the travel patterns of Beijing PT users include four categories: travel preference for comprehensive station area, travel preference for work-based station area, travel preference for work/commercial station area and travel preference for remote station area.

5.2.7 Spatial correlation labels. In this label, the spatial correlation analysis of PT stations is conducted to understand the spatial distribution patterns of PT user travel behavior. The global Moran's I of the boarding station is calculated, in which the passenger

flow of station is selected as the variable for constructing the spatial weights, and the results are shown in Figure 9.

The global Moran's I of the boarding station is 0.54, in the range of 0 to 1, which indicates that there is a certain linear relationship which reveals that the passenger traffic flow of the stations shows spatial correlation, and objects with similar attributes are clustered together in space, i.e. stations with high travel traffic flow clustered together and with low traffic flow clustered together. The results are evaluated for significance and 999 permutations are selected, which are shown in Figure 10.

As shown in Figure 10, the *p*-value is 0.001 and the *z*-value is 69.8221. It is concluded that the confidence interval is 99%. There is 99% certainty that the passenger flow at the

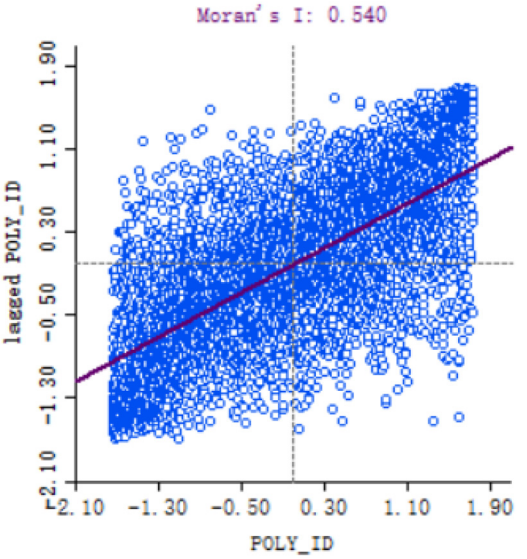


Figure 9.
Global Moran's I of the boarding station

Source: Created by authors

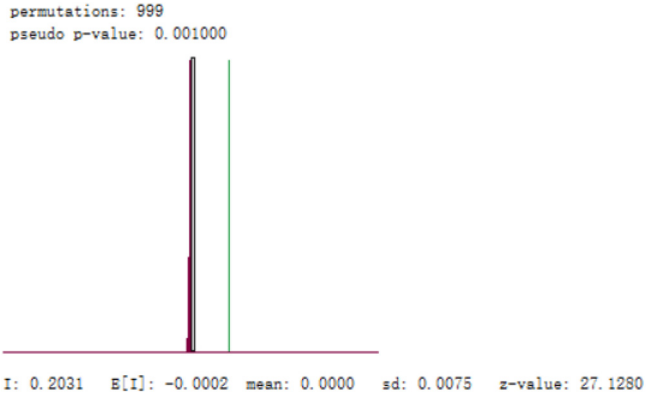


Figure 10.
Significance assessment results

Source: Created by authors

boarding stations is spatially correlated. Performing the local Moran's I calculation, the significance plot and the aggregated distribution plot of Moran scatter are obtained, and the results are shown in Figures 11 and 12.

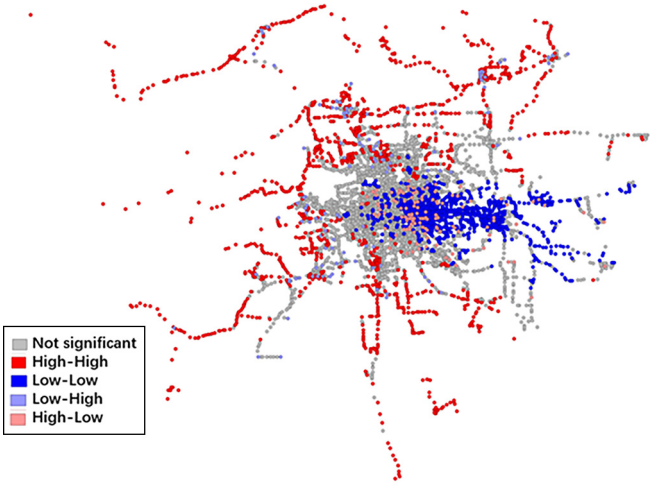
Figure 11 depicts the aggregated distribution of Moran scatter for boarding stations based on four spatial relations corresponding to the four quadrants of the Moran scatter plot, namely, H–H (the region falling in the first quadrant of the Moran scatter plot), L-L (third quadrant), H–L (fourth quadrant) and L–H (second quadrant). Figure 12 shows the corresponding significance intervals, indicating whether the local correlation value for each region is significant or not. For example, stations with a strong passenger flow correlation are primarily concentrated around the central city, while areas with strong prominence are mainly located in downtown areas such as Chaoyang District. The results for alighting stations are similar, but the spatial correlation of passenger traffic flow for these stations is less aggregated, and stations with strong spatial correlation are concentrated in downtown areas, with fewer and more dispersed stations having high confidence.

Overall, when station traffic is set as the spatial weight, both the distribution of user-boarding stations and user-alighting stations in space show a certain spatial correlation. This indicates that stations with high (low) passenger flow tend to gather spatially, reflecting the existence of cluster travel preferences in user's travel habits.

5.3 Portraying travel pattern of Beijing public transit users

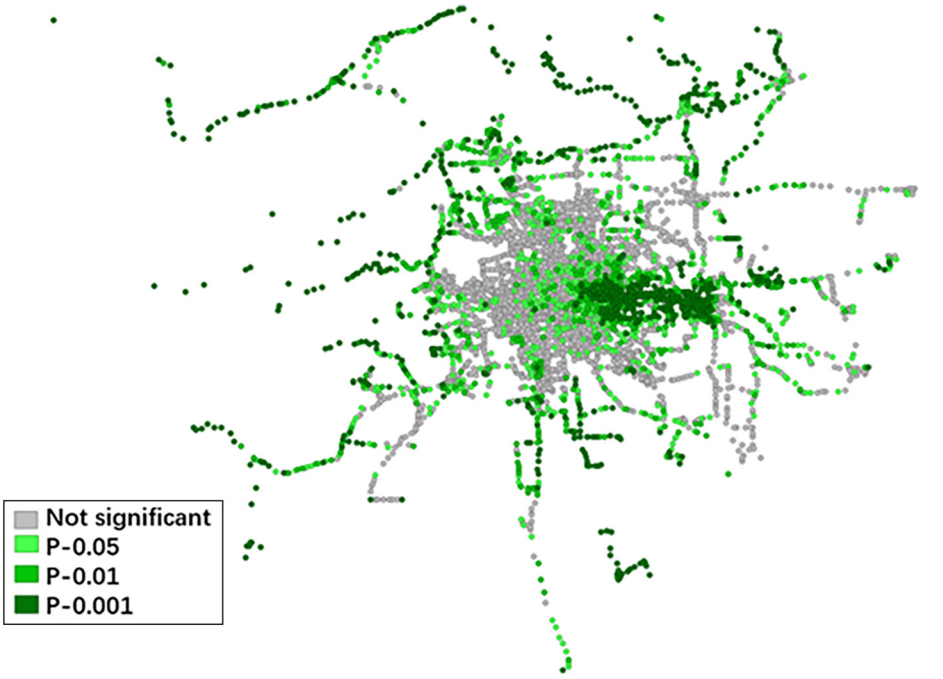
Based on the travel patterns of Beijing PT users obtained from the processing and analysis of user travel data, the PTUL established in subsection 3.4 is further subdivided to improve the model labels. This leads to the development of the final Beijing PT user travel profiling system, as illustrated in Figure 13.

From the Beijing PTUP results, Beijing PT user exhibit distinct spatiotemporal characteristics in their travel patterns. From time dimension, there exist obvious morning and evening peaks on weekdays, while travel on weekends is evenly distributed. On average, each user travel at least once a day and mainly ranges from 1 to 10 times a week in total. In terms of space, Chaoyang, Dongcheng and Haidian are the top three popular travel



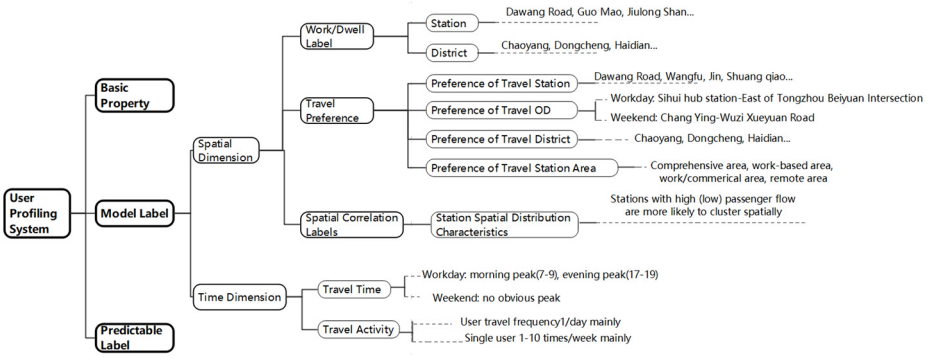
Source: Created by authors

Figure 11. Aggregated distribution of Moran scatter for boarding stations



Source: Created by authors

Figure 12.
Significance distribution for boarding stations



Source: Created by authors

Figure 13.
Beijing PTUP system

districts. The passenger flow of Dawang Road, Guomao, Jiulongshan Station, Wangfujing and Shuangqiao Station is always at a high level. The primary purpose of users' travel is for work, as evidenced by the most popular OD pair, which is from Sihui Hub Station to Tongzhou Beiyuan Road East. Sihui Hub Station serves as a major transportation hub within the Beijing urban area, while Tongzhou is home to a significant residential

population. Furthermore, when examining stations, high and low passenger flow stations exhibit a noticeable clustering pattern in space.

6. Conclusion

This study contributes by integrating UP techniques into the analysis of user travel patterns. It proposes a hierarchical framework for constructing a PTUL framework. Guided by this framework, the study examines the spatiotemporal distribution patterns of user travel, resulting in a comprehensive analytical approach. In addition to establishing the PTUP framework, this study defines methods for mining different labels within the framework. Taking Mingguang Village Station in Haidian District as an example, it introduces a station area attribute mining method based on the TF-IDF weighted algorithm, providing a mathematical model for studying user travel characteristics in subsequent research. Furthermore, the study proposes using the Moran's Index to examine the spatial correlation of stations' distribution. Subsequently, the PTUP framework is applied to Beijing PT users as a case study, resulting in a comprehensive spatiotemporal user travel profile, thus, confirming the feasibility of PTUP. However, his study still needs further research and improvement. Label mining mainly focuses on model labels, without proposing a definition method for predictive labels and without exploring basic property labels, and predictive labels, which lack certain data support.

References

Amoretti, M., Belli, L. and Zanichelli, F. (2017), "UTravel: smart mobility with a novel user profiling and recommendation approach", *Pervasive and Mobile Computing*, Vol. 38, pp. 474-489.

Badriyah, T., Wijayanto, E.T., Syarif, I. and Kristalina, P. (2017), "A hybrid recommendation system for E-commerce based on product description and user profile", *2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, IEEE, pp. 95-100.

Bao, H., Ming, D., Guo, Y., Zhang, K., Zhou, K. and Du, S. (2020), "DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data", *Remote Sensing*, Vol. 12 No. 7, p. 1088.

Corney, M., Mohay, G. and Clark, A. (2011), "Detection of anomalies from user profiles generated from system logs", *Proc. 9th Australas. Inf. Secur. Conf*, Vol. 116, pp. 23-32.

Fijałkowski, D. and Zatoka, R. (2011), "An architecture of a Web recommender system using social network user profiles for e-commerce", *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, pp. 287-290.

Gutiérrez, A., Domènech, A., Zaragozí, B. and Miravet, D. (2020), "Profiling tourists' use of public transport through smart travel card data", *Journal of Transport Geography*, Vol. 88, p. 102820.

Hasan, S., Schneider, C.M., Ukkusuri, S.V. and González, M.C. (2013), "Spatiotemporal patterns of urban human mobility", *Journal of Statistical Physics*, Vol. 151 Nos 1/2, pp. 304-318.

Hu, Y. and Han, Y. (2019), "Identification of urban functional areas based on POI data: a case study of the Guangzhou economic and technological development zone", *Sustainability*, Vol. 11 No. 5, p. 1385.

Hu, L., Zhang, Y., Chung, S.H. and Wang, L. (2022), "Two-tier price membership mechanism design based on user profiles", *Electronic Commerce Research and Applications*, Vol. 52, p. 101130.

Jomsri, P. (2014), "Book recommendation system for digital library based on user profiles by using association rule", *Fourth edition of the International Conference on the Innovative Computing Technology (INTECH 2014)*, IEEE, pp. 130-134.

LeRouge, C., Ma, J., Sneha, S. and Tolle, K. (2013), "User profiles and personas in the design and development of consumer health technologies", *International Journal of Medical Informatics*, Vol. 82 No. 11, pp. e251-e268.

Li, S. and Tang, Y. (2020), "A simple framework of smart geriatric nursing considering health big data and user profile", *Computational and Mathematical Methods in Medicine*, Vol. 2020, p. 5013249.

Li, J., Ye, Q., Deng, X., Liu, Y. and Liu, Y. (2016), "Spatial-temporal analysis on spring festival travel rush in China based on multisource big data", *Sustainability*, Vol. 8 No. 11, p. 1184, doi: 10.3390/su8111184.

Liu, S., Yamamoto, T., Yao, E. and Nakamura, T. (2020), "Exploring travel pattern variability of public transport users through smart card data: role of gender and age", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23 No. 5, pp. 4247-4256.

Miao, R., Wang, Y. and Li, S. (2021), "Analyzing urban spatial patterns and functional zones using Sina Weibo POI data: a case study of Beijing", *Sustainability*, Vol. 13 No. 2, p. 647.

Mishra, R.K. and Urolagin, S. (2019), "A sentiment analysis-based hotel recommendation using TF-IDF approach", *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, IEEE, pp. 811-815.

Ouaftouh, S., Sassi, I., Zellou, A. and Anter, S. (2019), "Flat and hierarchical user profile clustering in an e-commerce recommender system", *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, IEEE, pp. 1-5.

Rafiq, R. and McNally, M.G. (2021), "Heterogeneity in activity-travel patterns of public transit users: an application of latent class analysis", *Transportation Research Part A: Policy and Practice*, Vol. 152, pp. 1-18.

Shirude, S.B. and Kolhe, S.R. (2014), "Measuring similarity between user profile and library book", *2014 international conference on information systems and computer networks (ISCON)*, IEEE, pp. 50-54.

Sugiyama, K., Hatano, K. and Yoshikawa, M. (2004), "Adaptive web search based on user profile constructed without any effort from use's", *Proc. 13th Int. Conf. World Wide*, pp. 675-684.

Tang, J., Yao, L., Zhang, D. and Zhang, J. (2010), "A combination approach to web user profiling", *ACM Transactions on Knowledge Discovery from Data*, Vol. 5 No. 1, p. 2.

Thompson, S., Eva, N. and Shea, E. (2017), "Watching the movie: using personas as a library marketing tool", *Reference and User Services Quarterly*, Vol. 57 No. 1, pp. 17-19.

Van Oort, N. and Cats, O. (2015), "Improving public transport decision making, planning and operations by using big data: cases from Sweden and The Netherlands", *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, IEEE, pp. 19-24.

Verma, T., Sirenko, M., Kornecki, I., Cunningham, S. and Araújo, N.A. (2021), "Extracting spatiotemporal commuting patterns from public transit data", *Journal of Urban Mobility*, Vol. 1, p. 100004.

Wang, Z., Ma, D., Sun, D. and Zhang, J. (2021), "Identification and analysis of urban functional area in Hangzhou based on OSM and POI data", *Plos One*, Vol. 16 No. 5, p. e0251988.

Xia, D., Jiang, S., Yang, N., Hu, Y., Li, Y., Li, H. and Wang, L. (2021), "Discovering spatiotemporal characteristics of passenger travel with mobile trajectory big data", *Physica A: Statistical Mechanics and Its Applications*, Vol. 578, p. 126056.

Yu, Z., Zhou, X., Hao, Y. and Gu, J. (2006), "TV program recommendation for multiple viewers based on user profile merging", *User Modeling and User-Adapted Interaction*, Vol. 16 No. 1, pp. 63-82.

Zannat, K.E. and Choudhury, C.F. (2019), "Emerging big data sources for public station planning: a systematic review on current state of art and future research directions", *Journal of the Indian Institute of Science*, Vol. 99 No. 4, pp. 601-619.

Zhao, J., Qu, Q., Zhang, F., Xu, C. and Liu, S. (2017), "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.

Zhiyuan, H., Liang, Z., Ruihua, X. and Feng, Z. (2017), "Application of big data visualization in passenger flow analysis of shanghai metro network", *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, IEEE, pp. 184-188.

Corresponding author
Ailing Huang can be contacted at: alhuang@bjtu.edu.cn