Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 2012579, 16 pages
https://doi.org/10.1155/2022/2012579

Research Article

Individual Travel Knowledge Graph-Based Public Transport Commuter Identification: A Mixed Data Learning Approach

Song Hu,[1] Jiancheng Weng,[1] Quan Liang,[2] Wei Zhou,[3] and Peizhao Wang[4]

[1]*Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, Beijing, China*
[2]*Department of Road Teaching and Research, Transport Management Institute Ministry of Transport of the China, Beijing, China*
[3]*Ministry of Transport of the People's Republic of China, Beijing, China*
[4]*Centre for Advanced Spatial Analysis, University College London, London, UK*

Correspondence should be addressed to Quan Liang; lquan0730@163.com

Received 21 July 2021; Revised 12 January 2022; Accepted 29 January 2022; Published 24 February 2022

Academic Editor: Tao Liu

Commuters are the stable travel group for the public transportation (PT) service system. Accurately identifying the PT commuters is conducive to promoting PT service quality and development of urban sustainable transportation. This paper extracts individual PT travel chain information and constructs individual travel knowledge graphs of PT passengers based on the association matching algorithm and the theory of multilayer planning. A mixed dataset is formed by associating individual travel chains with travel survey data. Seven travel characteristic indicators regarding travel performance and spatiotemporal travel characteristics are extracted. The identification model of PT commuters is developed based on a three-layer backpropagation neural network (BPNN). The optimal model structure of neuron node number, transfer function, and learning rate are discussed quantitatively according to the minimization of model errors. The evaluation indexes of overall accuracy and *kappa* coefficient of the constructed model are 94.5% and 87.9% separately. The results indicate that the model identification accuracy is acceptable, and the proposed characteristic indicators and systematic modelling procedure are effective. Then, the model performance is compared with the other five machine learning models further. The results confirm that the proposed model has a better identification accuracy and viability, and the model performance will improve with the increase of the sample size.

1. Introduction

With the continuous penetration of the concept of sustainable transport and green traveling, especially, the Chinese government put forward the goal of "carbon peak" and "carbon neutral" in 2021, and public transportation (PT) has become an increasingly important transportation option for the residents. According to official statistics, the total number of PT trips was 12.6 million accounting for 55.9% of the total number of motorized trips downtown in 2019 as compared to 48.1% in 2015, Beijing [1]. PT has occupied the largest share in the urban transportation market in the Chinese context. With a better understanding of the travel patterns of transit riders, transit authorities will be able to evaluate their current services to reveal how best to adjust their marketing strategies to attract higher PT usage [2]. Nevertheless, it is pointed out that there are prominent differences in the travel characteristics between PT commuting passengers and others [3]. Therefore, it is of great significance to effectively grasp the travel demands and mobility characteristics of the PT commuters, which is conducive to improving urban sustainable transport service. For this purpose, realizing accurate identification of the PT passenger category is the premise of revealing the travel demand and characteristics of heterogeneous passengers.

Most of the previous studies have attempted to apply smart card transaction data and travel survey data for analyzing mobility characteristics of PT commuters and detecting their behaviour differences. Some studies identified PT commuters by analyzing the commuting travel characteristics including travel mode, travel spatiotemporal regularity, and travel-route selection diversity [4, 5]. Jun and

Deng took the threshold values of travel frequency and departure time standard deviation as the classification standard; then, PT passengers from IC card data were divided into three categories: commuting, ordinary, and random [6]. Ma et al. proposed that commuters' travel regularity and spatiotemporal repeatability can be measured from the aspects of residence, workplace, and departure time; then, PT commuters were identified by leveraging spatial clustering and multicriteria decision analysis approaches [3]. Zou et al. proposed a rule-based recognition method that is utilized to identify the commuters from the perspective of spatiotemporal features, personal property, and travel behaviour [7]. However, although the aforementioned studies enable us effectively realize the passenger classification with significant commuting characteristics, they are inapplicable to accurately identify commuters with unapparent commuting behavioural characteristics. Thus, the selection of multidimensional and spatiotemporal travel behavior indicators is the crucial link to realize the identification of these atypical commuters. Besides, some relevant studies collected the travel data including travel purposes, travel mode, origins, and destinations of trips through the resident travel survey [8, 9], but the behaviour classification of passengers in the whole sample could not be realized due to the high cost and limited samples of travel survey. Moreover, many previous studies only used a single data source such as a travel survey or smart card transaction and lacked the integrated utilization of multisource travel data to extract more multidimensional travel behavior characteristics.

The intelligent PT system has been effectively improved with the emerging technology development of Internet of Things, big data, and cloud computing. In addition, the rapid evolution of artificial intelligence and machine learning technology also provides methodological support for data-driven PT passenger classification. Some previous studies identified the PT commuters based on the intelligent algorithm including association rules algorithm [10], convolutional neural networks (CNNs) [8], Naïve Bayes probabilistic model [11, 12], support vector machine and decision tree [12], and statistical analysis model [13]. Zhang et al. identified the commuters among numerous bus passengers by using the IC data with the cluster analysis [14]. Allahviranloo and Recker used Markov chain models to study the sequential choice of activities; then, the sequential multinomial logit (MNL) models and multiclass support vector machines (K-SVMs) were adopted to identify the activity pattern of in-home, work, maintenance, personal, pick up/drop off, and stop [15]. Rafiq and McNally analyzed transit-based activity-travel patterns by classifying users via latent class analysis, and data from the household travel survey were collected to classify the transit users [16]. Manley et al. also used the density-based spatial clustering of applications with noise (DBSCAN) algorithm to identify the travel spatiotemporal regularity of individuals, and the spatial and temporal regularity difference of each cluster was derived through a continuous long-term observation period [17]. The DBSCAN model was improved and applied to classify the passengers under a much lower calculation complexity [18, 19]. Sun and Yang established the Bayesian probabilistic relations from travel survey data; then, a Naive Bayesian method was constructed to identify PT commuters [11]. Moreover, they proposed a Naive Bayesian classifier model to identify PT commuters. The results showed that the model can identify the objectives using smart card data without requiring travel regularity assumptions of PT commuters [20]. Bösehans and Walker utilized the centroid clustering algorithm and k-means procedure to cluster the staff and students; then, the main travel mode of staff commuters and student commuters was identified and analyzed [21]. Weng and Lv selected the characteristic indexes of the average number and gap time of smart card transactions and departure time stability of weekdays from IC card transaction data; then, a commuter identification model was constructed by using the gradient boosting decision tree (GBDT) algorithm [22]. The above studies showed the methods of using intelligent models could identify the variation of traveler identity attributes and detect the categories of PT passengers. However, most of the previous studies on the analysis of passengers' commuting characteristics oversimplified the definition of PT commuters [6], and the characteristic variables of identification models were incomplete. Many studies characterized the PT commuters considering partial travel pattern characteristics, such as the simple frequency count [4], spatial travel patterns, [6] and travel time characteristics [8]; more comprehensive indicators including the spatiotemporal travel modes and travel choice characteristics should be adopted. Additionally, the structure design and parameter adjustment were not discussed in the modelling process quantitatively. Therefore, the applicability and extensibility of these methods need to be improved further.

Figure 1 shows the keyword structure relationship visualization of the aforementioned related literature. The prominent keywords of PT passenger identification are travel pattern, behaviour, information, neural network, and prediction model. It can be acquired that the literature structure relationship among neural networks, knowledge graphs, and travel patterns is relatively weak. Therefore, exploring the types of residents' travel patterns combined with the neural network and knowledge graph is beneficial to enrich the research achievements.

The artificial neural network (ANN) algorithm has the advantages of self-learning, self-organization, favorable fault tolerance, and the ability of highly nonlinear mapping from the input to the output which can figure out the classification problems with better performance. What is more, more than 80% of ANNs employ the error backpropagation (BP) algorithm or its improved algorithm to construct their structures [23]. Therefore, we adopt the BP neural network (BPNN) framework to develop a mixed data learning model that is employed to identify PT commuters accurately.

This paper is aimed at proposing a systematic process approach to identify the PT commuters based on the PT travel chain data and multimode travel graph. The proposed method based on a three-layer BPNN model contributes to illustrating the relationships between the estimated passenger categories and multidimensional travel behavior

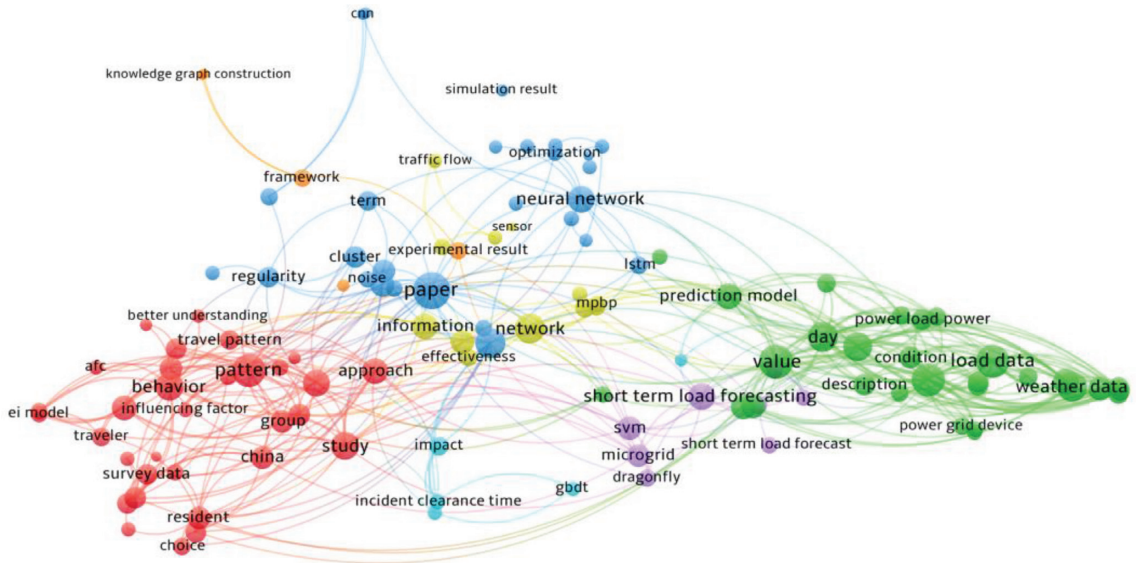

Figure 1: Keyword structure relationship of the related literature.

characteristics. Wherein, a quantitative analysis method is used to determine the model parameters and structure, which improves the scientific and systematic construction of the proposed PT commuter identification model. The result is expected to lay a solid foundation for multidimensional analysis of passenger's travel demands and enhance understanding of the composition of urban travel groups and their behaviour performance during monitoring of the smart card data. Besides, the identified behavior characteristics of commuter groups are conducive to traffic managers to improve PT services and their sharing rate.

This paper is structured as follows. The data foundation is introduced first, followed by the extraction of individual passenger travel chains. The identification method of PT passengers is explained in the following order: (1) construction of travel knowledge graph of passengers, (2) extraction of travel characteristic indicators of PT passengers, and (3) structural design and parameter adjustment of the BPNN model, after which the constructed model is applied to detect the categories of PT passengers, and the model results are verified effectively. The paper concludes by summarizing the research findings and suggesting directions for future research.

2. Extraction Methodology of Travel Chain Data

This section proposes a method for extracting individual travel chains that reflect the whole travel process of passengers through the collection, correlation, and matching for multisource PT data, and attempt to lay a foundation for the construction of the PT commuter identification model.

2.1. Multisource PT Data Acquisition and Processing. The multisource PT trip data used in this paper including the smart card (automated fare collection card, AFC card; integrated circuit card, IC card) transaction data, PT network data, and global positioning system (GPS) data of bus are collected at the entire city scale, according to Beijing Transportation Operation Coordination Center (TOCC) and Transit Metropolis Platform. To improve the quality and availability of obtained raw bus data, the GPS data, PT network data were utilized to calibrate the information on the boarding and alighting stations and time; also, the missing data of stations were restored by adopting similar handling methods in literature [24, 25]. Considering the detailed process of data handling was not the focus in this section, the related contents can be learned from the aforementioned literature. Moreover, the location of the AFC system in metro stations is fixed, so it is not necessary to check the smart card information of the metro system using the data obtained from the automatic vehicle location (AVL) system.

To effectively extract and analyze the travel information of PT passengers, some valuable fields related to the mobility of passengers can be obtained from the raw data of smart card transaction data, PT network data, and GPS data of the bus. Table 1 shows the selected valid fields of these data.

2.2. Extraction of Individual PT Travel Chains. To clearly understand the individual PT travel behaviour and mine more useful information from smart card transaction data, an extraction process of the individual PT travel chains will be implemented in this section. Each smart card transaction record is defined as a travel stage that reflects information about a segment of a passenger's journey. A travel chain means a continuous journey of passengers over time. Hence, a travel chain could contain multiple travel stages. Figure 2 shows the two-dimensional structure of the individual PT travel chain that includes two transfers and three travel stages in the spatiotemporal dimensions. The horizontal axis represents the travel time and the duration of the trip, and the vertical coordination indicates the spatial mobility from the origin (O) and destination (D); the slopes of the slanted lines can intuitively reflect the speed of mobility for each travel mode. Additionally, the definitions of several

Table 1: Valid fields of multisource PT data.

	Bus	Metro
Smart card transaction data	User card code Boarding/alighting line number Boarding/alighting station number Boarding/alighting time	User card code Inbound/outbound line number Entry/exit station code Inbound/outbound time
GPS data	Line number Data return time Latitude and longitude of return point	—
Network and station data	Arc start/end number Arc length Longitude and latitude of stations Station spacing	Entry/exit station code Inbound/outbound name Longitude and latitude of stations Station spacing

Note: the card codes of the smart card data are not always identical to the individuals. For example, a smart card can be shared among family members and friends, or a traveler can hold several cards. However, such usage may not be the majority, especially when registered monthly passes belong to the smart cards [9]. With the rapid development of mobile payment, the PT systems apply to the quick response (QR) code payment besides the traditional smart card payment in Beijing. However, the code rules of QR codes are not consistent with those of the smart cards, and the service operators do not provide the number of QR codes in the transaction application software. Therefore, it is infeasible to associate the individual travel data and QR code transaction data that account for about 20% of all transaction data in Beijing. Thus, this study focuses on the smart card transaction data to effectively introduce and match the corresponding individual travel survey data. What is more, smart card data are ticket-dependent methods, and they typically underestimate the travel demand owing to possible fare evaders in many worldwide transit systems [26]. However, no ticketing system can avoid fare evasion, and the percentage of possible fare evaders is relatively low, so this limitation is ignored in the study. The PT operating companies allowed the use of the smart card data only for research purposes; the individual information had been anonymized prior to the analysis to protect the privacy of cardholders throughout this study.

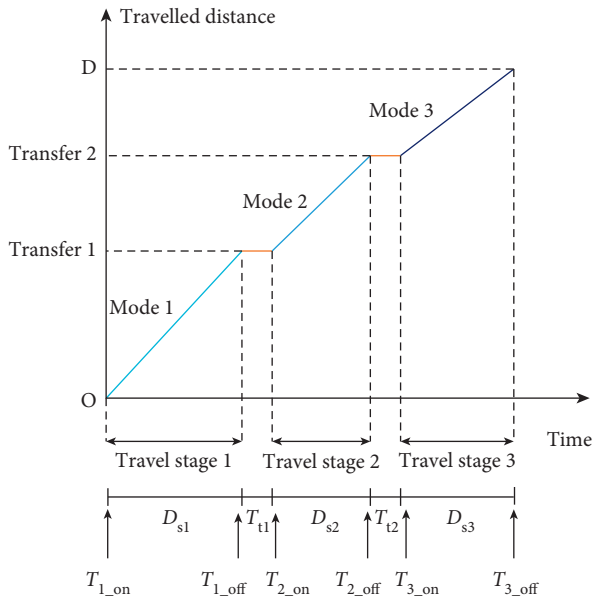

Figure 2: Two-dimensional structure of an individual PT travel chain.

parameters in Figure 2 are described as follows: T_{i_on} means the boarding time at the beginning of the i^{th} travel stage, T_{i_off} presents the alighting time at the end of the i^{th} travel stage, D_{si} demonstrates the duration at the i^{th} travel stage, and T_{ti} illustrates the transferring time between the i^{th} travel stage and the $(i+1)^{th}$ travel stage. We note that D_{si} and T_{ti} are available when the card code of a cardholder is provided, since the data of these variables are defined from two consecutive transaction records. In addition, "Mode i" indicates different modes of PT; "Transfer i" is the process of traffic mode conversion from the travel phase i to the travel phase $(i+1)$ and "Travelled distance" means the distance between OD.

The method of extracting individual travel chains based on multisource PT data includes two steps: multisource PT data integration and the association and matching of passengers' travel information [27]. Figure 3 describes the whole extraction process of the individual travel chains for PT passengers.

The first step focuses on integrating the spatiotemporal mobility data and presenting the travel stages of PT passengers. Besides selecting the corresponding attributes per algorithm, some general preprocessing operations were applied to the data. The smart card transaction data were merged into a dataset and then were grouped by the card code and sorted by timestamp. Thus, the individual smart card transaction data with key fields can be organized preliminarily.

The next step consists of four processing substeps: the judgment of transferring time threshold, travel chain structure acquisition, O/D inference of travel stage, and travel feature information matching need to be executed to extract the individual travel chains. We note that three transferring time thresholds need to be discussed resulting from three kinds of mode transferring relations including bus to bus, bus to metro, and metro to bus. In addition, the passengers can transfer to another metro line within the station, and there are no transaction records for tracking the transferring time of metro trips. Therefore, the transferring time threshold of the metro to the metro is not included in this study. The smart card transaction records of the bus only provide the alighting time, and the smart card transaction records of the metro contain both the boarding and alighting time, so the transferring time gaps of the bus to bus and metro to bus contain the riding time on PT. Thus, the three kinds of transferring time thresholds have great differences.

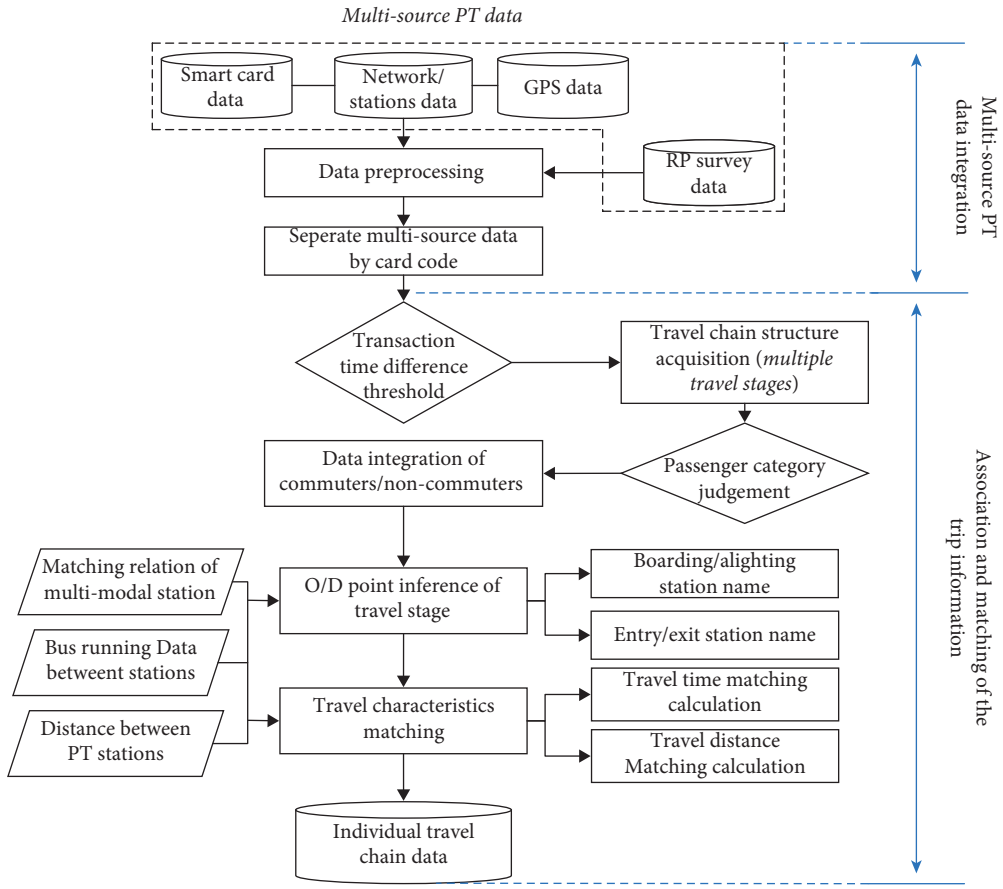

Figure 3: Extraction process of individual travel chains for PT passengers.

We use the probability distribution statistical method to extract the values of 95% of cumulative frequency as the transferring time thresholds, which are 112 min, 20 min, and 104 min, respectively.

In the PT travel chain dataset, any travel information of PT can be included more than what appears in smart card transaction data. Therefore, more mixed travel information can be effectively obtained from the travel chain data, such as OD points, travel distance, transferring time, the number of transfers, and travel model. Table 2 shows some important travel information of PT passengers obtained from the individual travel chain dataset.

Especially, the field of card type in IC card transaction data provides the elementary category information of PT passengers. It is not difficult for us to intuitively recognize the identity including students, adults, and seniors of PT passengers by the field of card type. Thus, the numbers of PT travel chains of the passengers with different categories can be obtained severally, and the day-to-day changes in passenger numbers of different passengers can be observed. Figure 4 shows the changes and statistical results of PT travel chains of different passengers from 1st to 7th June in 2019, Beijing. The travel chain data covering four consecutive days from 3rd to 7th June were workdays, and the number of travel chains is about 8 million every day. The days of 1st and 2nd June were weekends, and 7th June was Dragon Boat Festival which is a Chinese traditional festival, so the number of travel chains in each of these days was slightly lower than that of the workday, and the number was about 6 million. From the relative perspective, the scale of student passengers' travel chains accounts for 4.2% to 5% of the total number of daily travel chains, which was the smallest group. The senior passengers who travel for leisure and recreation by PT account for nearly a third of all trips on weekends or festivals. Additionally, it is not surprising that the group of adult passengers makes up the largest proportion of trips reaching 62% to 65% of the total PT passenger flow.

However, the above analysis is just a coarse-grained category identification of PT passengers, and it is infeasible to infer the main daily travel purposes of the adult passenger. Namely, the passengers' behavioural categories including commuting and noncommuting activities cannot be identified merely according to the card types. Therefore, the following part focuses on the model construction and category analysis of the adult passengers selected from the whole sample.

2.3. Construction of Individual Travel Behaviour Graph. To observe and extract the individual travel characteristic variables for identifying the PT commuters more intuitively and effectively, we introduced the knowledge graph system to establish the individual travel behaviour graph in this study. Knowledge graph, as a visual expression way of

Table 2: Travel chain data sample of a PT passenger.

Card code Card type	24XXXX73	24 XXXX73 1 (adult card)	...	24 XXXX73
Travel mode	Metro	Metro	...	Bus-metro
Boarding/inbound time	2017/5/1 8:28	2017/5/1 17:53	...	2017/5/31 17:04
Alighting/outbound time	2017/5/1 8:55	2017/5/1 16:29	...	2017/5/31 17:39
Boarding/inbound line number	4	1	...	114
Alighting/outbound line number	1	4	...	4
Travel distance (m)	8115	8115	...	8620
On station	Beijing south station	Muxidi station	...	Baiyunqiao west
Off station	Muxidi station	Beijing south station	...	Beijing south station
On station longitude (°)	116.3779	116.3369	...	116.3395
On station latitude (°)	39.8641	39.9075	...	39.8973
Off station longitude (°)	116.3369	116.3779	...	116.3779
Off station latitude (°)	39.9075	39.8641	...	39.8641

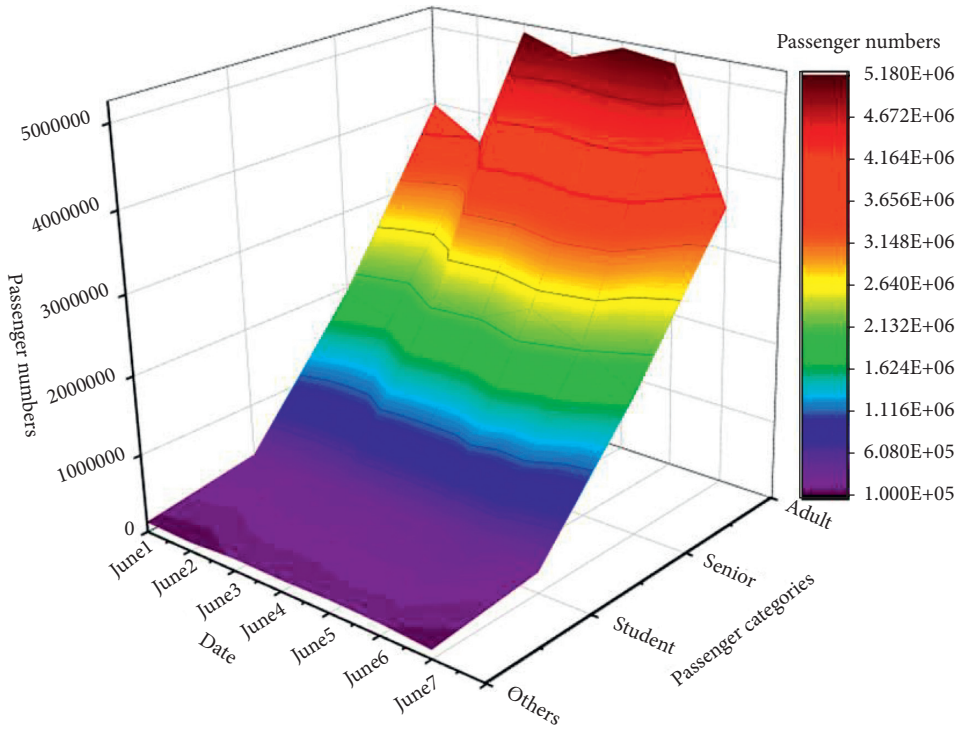

Figure 4: Day-to-day changes in the number of every category passengers.

characteristic information, owns the advantages of describing the concepts and mutual relationships among objects in the form of symbols, and the network structure which can realize the intuitive expression of characteristic indicators is formed by the connection of relations [28]. Therefore, we can realize the individual travel behaviour expression based on the knowledge graph theory from the fragmented and incomplete individual data. The significant effect of the individual travel behaviour graph is to transform the low dimensional numerical data into a high dimensional visual structure.

Based on the individual PT travel chain dataset, the spatiotemporal physical relation network of PT passengers' travel behaviour information including spatial positions, time distributions, and trip routes can be constructed. The construction steps of the individual travel behaviour graph are shown as follows [29]:

(1) The first step is to cluster the individual travel spatial locations. The hierarchical system cluster model is applied to cluster the longitude and latitude data of passengers' travel OD points from the travel chain dataset. Thus, these OD points are divided into different groups from the spatial dimension.

(2) Then, the individual travel time of PT passengers is further classified based on the results of travel space position clustering. Firstly, the travel time range of 05:00 to 23:00 which is the PT operation time in Beijing needs to be split into 2-hour intervals; thus, the travel time was divided into 9 intervals.

Thereafter, the departure time is classified in each OD group, and a dataset is established to store the cluster results.

(3) Next, the actual travel paths of the PT-passenger trips are clustered. The travel paths are represented by the actual travel distance and travel direction from the individual travel chain dataset. The results of the travel path clustering represent the travel modes of individual passengers. What is more, each travel time cluster was further classified in each travel path cluster.

(4) The last step is to construct an individual travel knowledge graph. Based on the multilayer planning theory, the foregoing clusters of spatial location, travel time, and travel paths were adopted to respectively construct the first, second, and third layers of the individual travel knowledge graph. The statistical probabilities of different travel behaviour modes in each layer of the travel knowledge graph are calculated to present the travel choice behaviour.

The individual travel behaviour graph represents intuitively the behavioural attributes in the spatiotemporal dimensions. The individual travel behaviour graphs are constructed to better understand PT passengers' travel characteristics including trips' spatiotemporal characteristics and travel stability, which is conducive to accurately and hierarchically extracting the input indicators of the PT commuter identification model. Figure 5 depicts the individual travel knowledge graph of a PT passenger that is selected randomly from the travel chain dataset in May 2017, Beijing. The spatial and temporal characteristics of individual travel behaviour in several continuous weekdays were intuitively expressed.

3. PT Commuter Identification Modelling

Neural network algorithms are among the most widely applied supervised learning methods in the field of machine learning and artificial intelligence. The method is well known in computer science, and there have been some successful applications of the method in traffic flow prediction [30], traffic model selection [31], and traffic congestion detection [32]. The BPNN method as a multilayer feedforward network is among the most widely applied supervised classification methods. The method is well known in computer science, and there are many successful applications of this method in the field of transportation [33–35]. However, the application of this methodology in studying aspects of travel behaviour and passenger category has been extremely limited, especially when it comes to identifying PT passenger categories in the Chinese context. What is more, Guo, et al. proved that a three-layer BPNN can satisfy most of the problems according to the universal approximation theory [36].

Therefore, a BPNN model with a three-layer structure was constructed as the identification model for the categories of PT passengers. The overall architecture of BPNN is composed of the input layer, hidden layer, and output layer. The description of the calculation flow of the BPNN model is as follows:

(1) Input layer to hidden layer:

$$\alpha_h = \sum_{i=1}^{d} v_{ih} * x_i, \tag{1}$$

where α_h represents the input value of the h^{th} neuron, d is the number of input variables, x_i indicates the input variables of the model, and v_{ih} is the weight to connect x_i in the input layer to the neuron α_h in the hidden layer.

(2) Activation function processing in the hidden layer:

$$b_h = f(\alpha_h - \gamma_h), \tag{2}$$

where b_h is the output value of the h^{th} neuron in the hidden layer, $f(x)$ illustrates the activation function, and γ_h presents the threshold of the h^{th} neuron.

(3) Hidden layer to output layer:

$$y_k = \sum_{h=1}^{q} w_{hk} * b_h, \tag{3}$$

where y_k is the model output value, w_{hk} illustrates the weight to connect the neuron α_h to the output variables in the output layer, k indicates the number of output indicators of the PT commuter prediction model, and q means the number of neurons in the hidden layer.

In addition, the model errors between the model results and the expected results are adopted to improve the model parameters. The model errors E are calculated using the least square method as follows:

$$E = \frac{1}{2}\sum_{k=1}^{K} (y_{k'} - y_k)^2, \tag{4}$$

where $y_{k'}$ indicates the prediction results, y_k represents the training results. The error is taken as the control target to capture the best functions and parameters of the BPNN model.

The following part describes the construction of a three-layer BPNN model from two aspects: structure design (feature variables of the input layer and passenger category of the output layer) and parameter adjustment (neuron node number of the hidden layer, transfer function, and learning rate). The details of the model are discussed quantitatively.

3.1. Structure Design

3.1.1. Input Layer Design. The input layer of the BPNN model contains the typical characteristic variables of passengers' PT travel behaviour. To extract individual travel behaviour characteristics, several methods of entity extraction, relationship extraction, and attribute extraction in the domain of knowledge graph are used. Table 3 shows the selected seven characteristic indicators and their profile. By

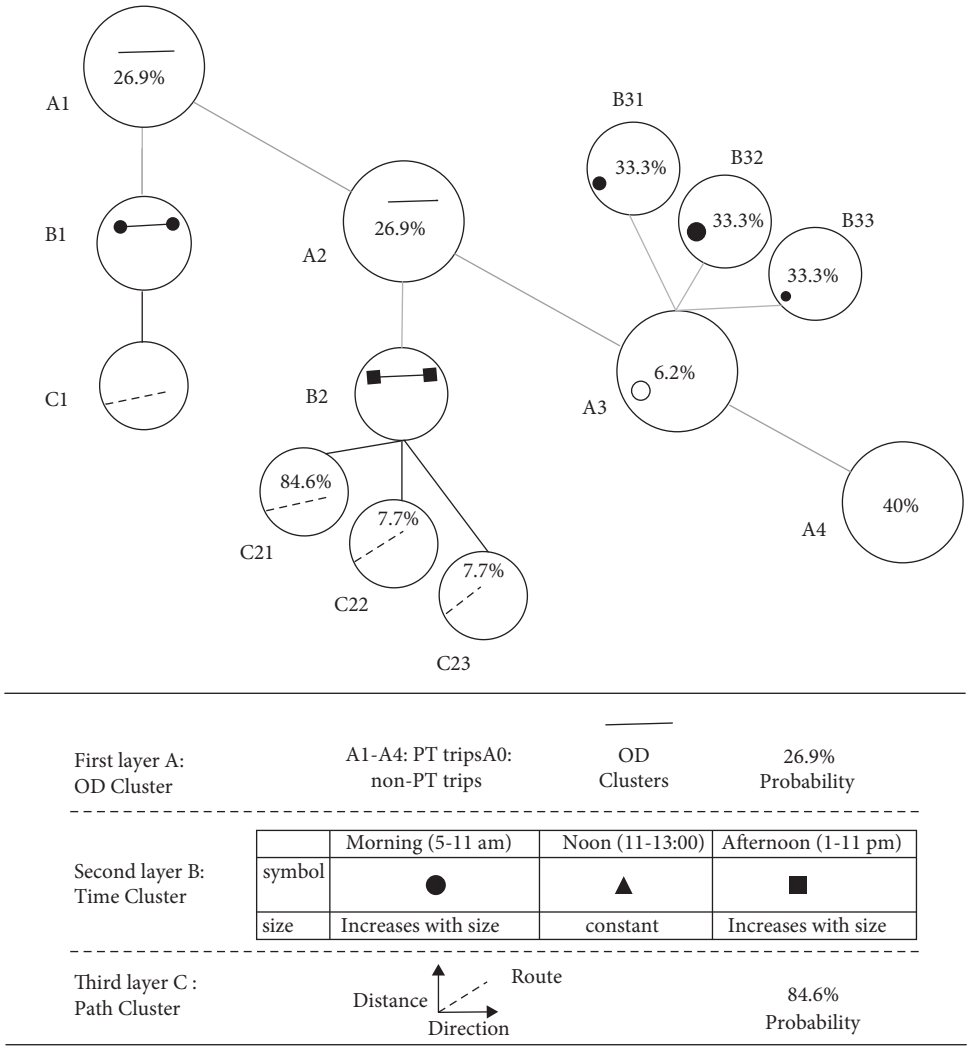

Figure 5: Travel knowledge graph of a PT passenger.

Table 3: Profile of characteristic indicators of PT commuter travel behaviour.

Indicators	Profile
ATD	The average number of days a passenger traveled by PT per month
ANT	The average number of trips by PT per month
ODCN	The cluster number of trips' OD points per month
RC	RC = 1, if a PT trip has a corresponding return trip by PT on the same day, 0 otherwise
DTC	The most concentrated departure period of the day for traveling by PT per month, where DTC = 1, 2, . . ., 9
TPF	TPF = 1, if the travel path is fixed for the same OD pairs in different departure time, 0 otherwise
TSE	The spatial distribution equilibrium characteristics of travel frequency of PT passengers to different activity OD points per month

observing the structure and features of individual travel knowledge graphs from multiple perspectives, four travel behaviour characteristic indexes including average travel days (ATDs), the average number of trips (ANTs), OD cluster number (ODCN), and PT roundtrip coefficient (RC) were extracted from the spatial dimension. While from the temporal dimension, the indicator of departure time concentricity (DTC) was developed from the second layer of the travel knowledge graph. Likewise, we selected the indicator of travel path fixity (TPF) from the third layer of the individual travel behaviour graph. Besides, one more comprehensive indicator of travel space equilibrium (TSE) was proposed to measure the frequency with which passengers travel to different activity OD points, through travel behaviour analysis based on individual travel knowledge graphs. These seven indicators can be used to define and

identify the PT commuter from different travel behavior perspectives. In general, the bigger the values of ATD, ANT, RC, and TPF, the smaller the values of ODCN and TSE, and the more concentrated the DTC; then, the respondents were more likely to be the commuters.

The first six indicators in Table 3 can be directly acquired based on individual travel knowledge graphs. Meanwhile, TSE is denoted by (5) and (6) which introduce the conception of information entropy. Therefore, we defined the TSE combined with the paradigm of information entropy function as follows:

$$\mathrm{TSE} = -\sum_{i=1}^{m}\left(\frac{1}{N}\sum_{n=1}^{N}\alpha_i\right) * \log_2\left(\frac{1}{N}\sum_{n=1}^{N}\alpha_i\right), \tag{5}$$

$$\alpha_i = \begin{cases} 0, & \text{the passenger didn't go to activity point } i \text{ on } N^{\text{th}} \text{ day}, \\ 1, & \text{the passenger went to activity point } i \text{ on } N^{\text{th}} \text{ day}, \end{cases} \tag{6}$$

where m denotes the total number of different activity OD points, $i = 1, 2, 3, \ldots, m$, N indicates the total number of travel days in a month, $N = 31$, and α_i presents the decision variable.

3.1.2. Output Layer Design. The output layer of this model is designed to predict the PT passenger categories including the commuter and noncommuter. To acquire accurately the ground truth of the passenger category to train the model proposed in this paper, the RP survey was designed and conducted to obtain the individual attributes and travel behaviour information of PT passengers from 10^{th} to 27^{th} May in 2017, Beijing. From the temporal perspective, the survey period covered the morning peak hours (7 : 00–9:00), evening peak hours (17 : 00–19 : 00), and off-peak peak hours. From the spatial perspective, this survey activity involves five subway stations and three bus stations in the downtown area of Beijing, and the land-use attributes cover residential, commercial, and leisure areas. Thus, 453 valid questionnaires were collected on purpose.

The survey comprises two parts. The first part is the travel records for activities which ask for detailed information about respondents' mobility information in the past one week, including the travel days, travel purpose, departure and arrival time, travel mode, and the number of trips. The second part is the sociodemographic characteristics including the main travel purpose (commuting and noncommuting), age, gender, occupation, monthly income, vehicle ownership, and educational status. The key information of travel purpose, which is utilized to mark the passenger categories, is significant to the results and the accuracy of the proposed model. Therefore, we emphasized the importance of this question to the interviewees and asked them to complete the question according to the actual situation during the field survey. In general, the commuters are the population whose daily travel purpose are commuting; they may work full time or several days per week. Besides, the card codes of respondents' smart cards were also collected in the form of anonymity through this survey. Thus, the corresponding travel chain dataset can be matched, and continuous one-month travel chain data of respondents were extracted from the whole travel chain dataset.

Thereafter, the survey data were correlated and matched with the travel chain database through the field of card code, and the respondents whose card type belonged to the adult card were further selected. Thus, the multidimensional survey data containing both individual travel chain data and travel survey data of 147 commuters and 42 noncommuters were achieved. Then, Cronbach's alpha test and Kaiser–Meyer–Olkin (KMO) test were used to measure the reliability and validity of the collected survey data. The results show that Cronbach's alpha coefficients and KMO coefficients of the collected survey data are all above 0.836 and 0.851, respectively, which implies that the survey data are effective and representative. Table 4 presents the basic information statistics of the respondents.

From the overall perspective, a large proportion of PT passengers are between 26 and 35 years old (about 39%), followed by the group of 21–25 years old which accounts for almost a quarter of the whole samples. Surprisingly, the respondents own a relatively high education level, and about 80% of respondents have a bachelor's degree or above in Beijing. As expected, the overall income level of the PT travel group is slightly lower because the monthly salary of three-quarters of the respondents is lower than the average monthly salary (8,476 RMB) of residents in Beijing, 2017. Besides, nearly half of the PT passengers only own one car, and around 40% of respondents do not have private cars. The low private car ownership saliently results from the strict policy restrictions on the license plate which was introduced in 2011, Beijing.

From the relative perspective, it is interesting to note that the ratio of women in the commuter group is slightly higher, while that is lower in the noncommuter group. Besides, the expected results were found that the proportion of passengers over 50 years old in the noncommuter group is higher than that in the commuter group, due to the retirees in that group. Additionally, the proportion of households with more than two cars is slightly higher in the noncommuting group, which may result from the actual condition that the elderly have more probability to own cars, and a majority of the elderly tend to live with their children who belong to the major car ownership groups in China.

Table 4: Basic information statistics.

Attribute		Share (%)	Attribute		Share (%)
Passenger category	1 (commuter)	61.82	Passenger category	0 (noncommuter)	38.18
Gender	Men	48.73	Gender	Men	53.26
	Women	51.27		Women	46.74
Age	18–20	2.96	Age	18–20	2.48
	21–25	25.37		21–25	26.72
	26–35	39.75		26–35	38.57
	36–45	17.76		36–45	17.36
	46–50	9.73		46–50	7.44
	≥50	4.43		≥50	7.43
Education	High school or below	4.86	Education	High school or below	4.14
	High school	11.63		High school	14.92
	Undergraduate	74.21		Undergraduate	72.10
	Graduate or above	9.30		Graduate or above	8.84
Monthly income (RMB)	≤1500	17.30	Monthly income (RMB)	≤1,500	18.28
	1,501–3,000	7.81		1,501–3,000	9.42
	3,001–5,000	16.46		3,001–5,000	17.17
	5,001–8,000	33.12		5,001–8,000	29.36
	8,001–15,000	20.68		8,001–15,000	20.78
	≥15,000	4.63		≥15,000	4.99
Vehicle ownership	0	40.80	Vehicle ownership	0	38.67
	1	52.43		1	52.21
	2	6.13		2	7.73
	≥3	0.64		≥3	1.39
Number of weekly travel days	0	0	Number of weekly travel days	0	0
	1	2.33		1	22.73
	2	0		2	27.27
	3	9.30		3	13.64
	4	6.98		4	9.09
	≥5	81.39		≥5	27.27
Number of weekly commuting trips	0	2.50	Number of weekly commuting trips	0	59.09
	1–3	2.50		1–3	22.73
	4–6	7.50		4–6	18.18
	7–9	7.50		7–9	0
	10–12	62.50		10–12	0
	13–15	5.00		13–15	0
	≥16	12.50		≥16	0
Number of weekly leisure trips	0	22.50	Number of weekly leisure trips	0	4.55
	1–3	50.00		1–3	45.45
	4–6	25.00		4–6	40.91
	7–9	0		7–9	0
	10–12	0		10–12	9.09
	13–15	2.50		13–15	0
	≥16	0		≥16	0

3.2. Parameter Adjustment. In this section, the neuron node number, transfer functions between adjacent layers, and the learning rate are discussed, which is conducive to saliently improving the efficiency and accuracy of the identification model proposed in this paper.

3.2.1. Selection of Neuron Node Number in the Hidden Layer. The number of neuron nodes in the hidden layer of the BPNN network follows the following functional relationship (7) with the number of input variables and output variables [23]:

$$n = \sqrt{n_{\text{in}} + n_{\text{out}}} + \alpha, \tag{7}$$

where n is the number of neuron nodes in the hidden layer, n_{in} indicates the number of input variables, n_{out} means the number of output variables, and α is a constant between 0 and 10.

Since the model has seven input variables and 1 output variable, we can obtain the number of neuron nodes $n \subseteq [3, 13]$ in the hidden layer according to (7). Considering the difference of prediction results with the change of the model structure, the BPNN model is executed 10 epochs, while the number of neuron nodes (n) in the hidden layer

increases linearly, $n = 3, 4, \ldots, 13$. Thus, a model classification accuracy can be obtained after each model runs. Therefore, the average classification accuracy of the proposed model with different neuron nodes numbers can be achieved, respectively.

Figure 6 gives some insight into the relationship between the average classification accuracy and the number of neuron nodes in the hidden layer. It can be acquired that the average classification accuracy of the BPNN model with four neural nodes is highest when other model parameters remain unchanged. Therefore, four neuron nodes as the optimal parameter selection for the proposed model were structured in the hidden layer.

3.2.2. Transfer Function Selection. Transfer functions as the local computing function map the output of neurons to the input of neurons in the adjacent network layers. The transfer functions determine the weights and thresholds of the whole neural network and have an important influence on the prediction results. Some transfer functions such as hyperbolic tangent function *Tansig* and linear function *Purelin* which are denoted by (8) and (9) have been used in the three-layer neural network in the field of image processing [37], water quality treatment [38], and environmental engineering [39]. And these corresponding models have achieved prominent prediction effects. Therefore, all the training functions were used to train BPNN 10 times, respectively; then, one of them would be selected as the optimal model function based on the prediction accuracy and training time.

$$tansig(N) = \frac{2}{1 + \exp(-2N)} - 1, \tag{8}$$

$$purelin(N) = N, \tag{9}$$

where N is the input vector of the characteristic indicators of PT commuter travel behaviour.

The indicators of prediction accuracy and convergence rate of the model with different training functions were selected to evaluate model efficiency, respectively. Analogously, the runtime was adopted to investigate the model performance with respect to different thresholds [40]. Figure 7 shows the results of prediction accuracy and training time for diverse training functions. From the relative perspective, it can be acquired from the figure that the training function *Trainrp* which is developed based on the elastic gradient descent method develops the BPNN model to achieve the best prediction accuracy. Besides, the training time of *trainrp* is only 18.6% lower than that of the function *traingd* and 41.7% faster than that of the function *traincgf*. What is more, the advantage of the training functions will be further highlighted if a larger scale of data is calculated. Therefore, we adopted *trainrp* as the model training function considering the comprehensive performance of the transfer functions.

3.2.3. Learning Rate Selection. Another key hyperparameter of the BPNN model is the learning rate for gradient descent. This parameter scales the magnitude of our weight updates to

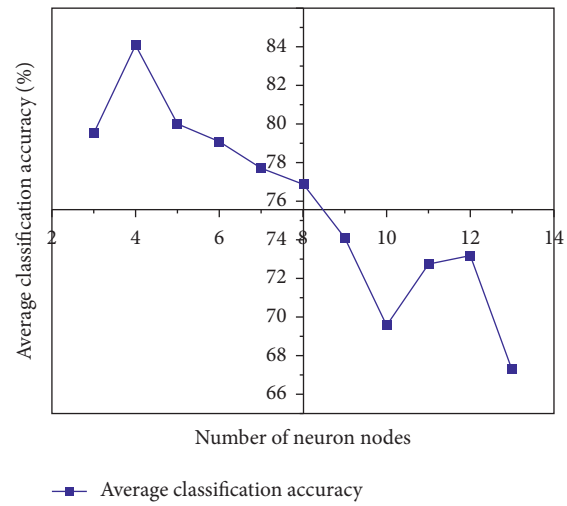

Figure 6: Relationship between average classification accuracy and neuron node number.

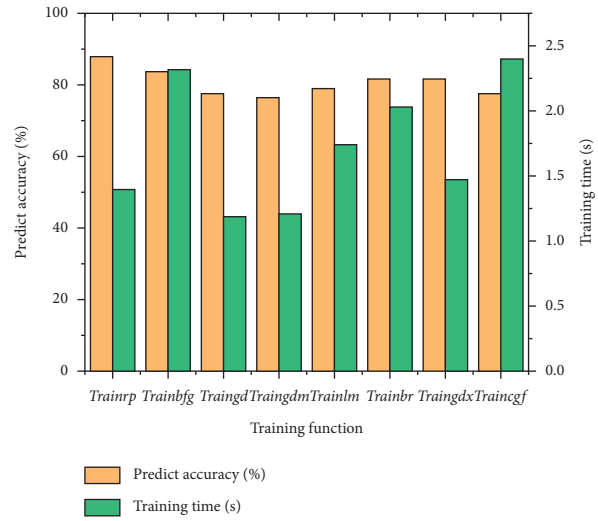

Figure 7: Prediction accuracy and training time of the model.

minimize the network's loss function, affects the stability and training time of the model, and determines the weight change in each cyclical training. If the values of the learning rate are too small, model training would progress very slowly due to very tiny updates to the weights in the network. Inversely, if the value of the learning rate is set too large, that could cause undesirable divergent behaviour in loss function and lead to the instability of the neural network. Substantial studies suggest that the learning rate of 0.01 has made the neural network model achieve salient prediction performance [40–42]. Therefore, we adopted the value of 0.01 as the learning rate of the BPNN model considering these previous achievements.

Through the aforementioned parameter adjustment and optimization, a stable PT commuter identification model proposed in the paper is finally constructed through the foregoing discussion on the parameters and structure of the model. Figure 8 shows the conceptual structure of the PT commuter identification model based on the three-layer neural network model.

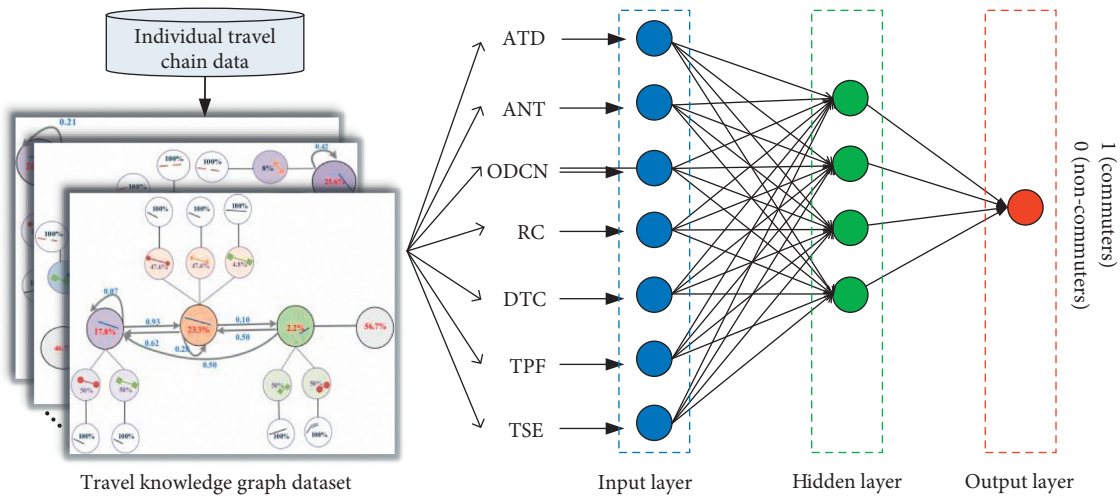

Figure 8: Conceptual structure of PT commuter identification model.

4. Empirical Results

In this section, we describe empirical analysis using the method proposed in Section 2, and the identification model was built and tested by using the travel chain data of respondents harvested from Beijing. To train and complete the proposed BPNN model, we randomly selected the datasets of 145 respondents as the training datasets and 44 respondents' datasets as verification datasets. Thus, there forms a $145 * 7$ matrix from the training dataset and a $44 * 7$ matrix from the verification dataset.

Figure 9 illustrates the feature dataset processing flowchart for PT commuter identification by the BPNN model. Firstly, the training data containing the characteristic indicator data derived from individual travel chain datasets and the category attribute information of passengers extracted from the survey data are input into the BPNN model. Then, the identification model is trained through the iterative adjustment of weight and parameters based on the error backpropagation and self-learning mechanism. Thereafter, the verification dataset is fed into the trained model developed in the previous step. Thus, the input data are computed and transmitted at each layer of the BPNN model, and the passenger categories could be predicted and estimated by the model.

The model performance depends on whether the heterogeneity in the attributes accurately indicates the difference in the passenger categories. Therefore, to evaluate the predicted classification accuracy and validity of the PT commuter identification model and data fusion approach proposed in this paper, we adopted the evaluation indicators of overall accuracy (OA) and *kappa* coefficient (*Kappa*) which have been successfully applied in the previous studies [43–45]. Although the model validation method is relatively simple, it is very effective and clear, and also easy to compare with other model results. Then, these two indicators were applied in evaluating the PT commuter identification model proposed in this paper, and the OA and *Kappa* are estimated by equations (10) and (11):

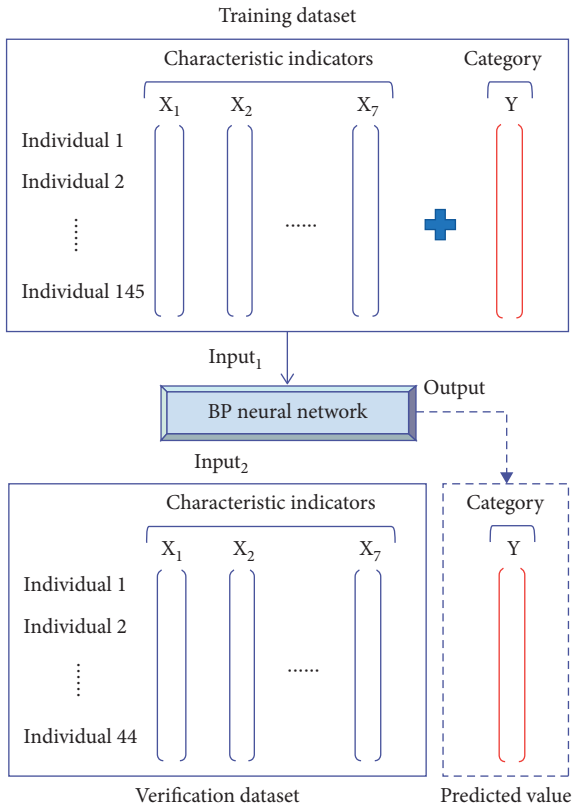

Figure 9: Feature set processing framework for PT commuter identification.

$$\mathrm{OA} = \frac{\sum a_{ii}}{N}, \tag{10}$$

$$Kappa = \frac{N \times \sum a_{ii} - \sum\left(T_{*j} \times T_{i*}\right)}{N^2 - \sum\left(T_{*j} \times T_{i*}\right)}, \tag{11}$$

where OA indicates the ratio of the number of correctly classified passengers to the total number of passengers,

Kappa represents the reduced error percentage of predicted classification results compared with the random classification, a_{ij} is the diagonal elements of the confusion matrix, N is the overall sample size, T_{*j} is the sum of the j^{th} column values of the confusion matrix, and T_{i*} is the sum of the i^{th} row of the confusion matrix.

We note that the two evaluation indicators are calculated based on the confusion matrix, which is commonly used to compare the errors between ground truth and predicted values in the field of artificial intelligence, especially supervised learning. Therefore, the confusion matrix regarding the passenger category was constructed. Then, seven characteristic indicators in Table 3 derived from the verification dataset were input into the identification model to estimate the categories of the samples. Table 5 shows the estimation results of the PT passenger category in the confusion matrix.

Thus, the evaluation indicators of OA and *Kappa* can be calculated based on the above confusion matrix and equations (10) and (11). The calculated results of these two values are 95.4% and 87.9%, respectively. The classification accuracy of the model can be considered almost identical to the ground truth when the value of *Kappa* is between 0.81 and 1.00 [46]. The good model accuracy also means that the model-overfitting problem is not prominent. The majority of commuters have high values in the indicators of ATD, ANT, and RC, and they travel by PT at least 3 days per week. However, the commuters who are not correctly identified travel by PT only once or twice a week and have few PT trips because they shift to PT for commuting only when their trips by car are limited by the motor vehicle restriction policy, and adverse weather or major events occurred. These PT passengers have the commuting purpose while they do not show the spatiotemporal characteristics of typical commuting travel. Fortunately, these passengers are not the focus of traffic regulators and policymakers because their trips do not have much impact on PT network planning and traffic demand forecast. Regarding the group of noncommuters, it is also worth noting that though all noncommuting passengers are identified correctly in this experiment, the noncommuters with similar commuting travel characteristics are likely to be identified as commuters. Particularly, the aforementioned two groups of passengers were not expected to be identified and were in small proportion; thus, such identification errors can be ignored.

In addition, the accuracy and performance of the trained BPNN model in this paper are compared with those of the previous studies further [9, 22]. The method of comparative analysis is a common technique to highlight the advantages of models more or less, though different methods have their characteristics under certain conditions. For example, the existing literature verified the better performance of the proposed model in automatically identifying a pilot's brain workload compared with its seven peers including the Gaussian mixture model, infinite student's t-mixture model, and DBSCAN model [47]. The involved models are gradient boosting decision tree (GBDT), Bayes, decision tree (DT), random forest (RF), and Naïve Bayes probabilistic model (NBPM), respectively. Figure 10 presents the compared results of these models. The results show that the BPNN model constructed in this paper has high prediction accuracy and is appropriate to identify the categories of PT passengers. And the indicators described in Section 3.1.1 are effective to distinguish the commuters and noncommuters in terms of these characteristics. In addition, the differences between the results of alternative methods in Figure 10 are not so prominent, which was caused partially by the limited data. With the increase of sample scale, the accuracy and superiority of the proposed model would improve further. Furthermore, the mixed data learning approach based on the BPNN model achieves the transit market segments of PT passengers including the commuter, noncommuter, student, senior, and staff who contribute to the changes in the transit demand. In addition, the results suggest that the proposed PT passenger category identification method can help the transport operators to analyze the travel behaviour differences during the monitoring of travelers. That enables them to analyze the relationship between the originally observed attributes of the integrated trip data and the estimated attributes that are originally unobserved [9].

Table 5: Confusion matrix of estimated PT passenger category.

Ground truth	Estimated value		
	Commuter	Noncommuter	Total
Commuter	32	2	34
Noncommuter	0	10	10
Total	32	12	44

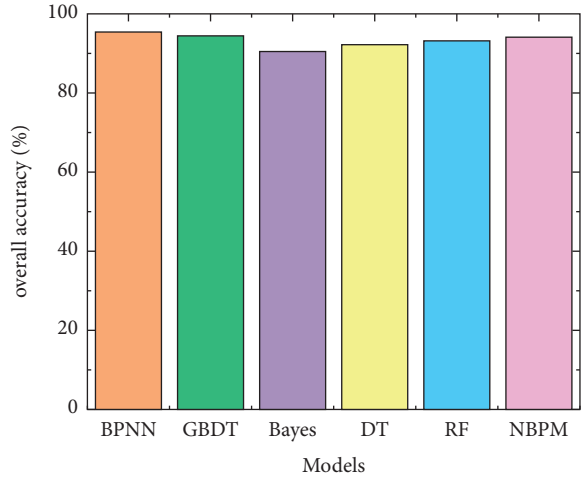

Figure 10: The comparison results of model performance.

5. Conclusions and Discussion

In this paper, the individual travel chains reflecting the whole travel process of PT passengers were exacted through the correlation and matching method based on the multisource PT data collected in Beijing, China. Then, the field of card type in smart card transaction data was applied to analyze the day-to-day changes in passenger flow of PT passengers. Thereafter, the multimode travel knowledge graphs were constructed for extracting the characteristic

indicators including ATD, ANT, ODCN, RC, DTC, TPF, and TSE hierarchically from multiple perspectives. Besides, the multimode travel knowledge graphs also contribute to understanding the travel habits and travel features of individuals. The control variable method and the comparative analysis method were applied to fit the optimal parameters of the BPNN model to identify the PT commuters more accurately.

The evaluation indicators of OA and *Kappa* are adopted to verify the model identification accuracy, as shown in equations (10) and (11), demonstrating that the proposed method correctly estimated 95.4% of the passenger categories while the incorrect estimations were caused by residents' noncommuting trips with similar commuting travel characteristics and residents' commuting trips with similar noncommuting travel characteristics. For example, some noncommuters regularly take PT to go shopping or exercise in parks every day, while some commuters rarely use PT to travel only under some special conditions. The model results indicate that the BPNN model proposed in this paper can effectively realize the identification of PT commuters. Also, the relatively good model predictive power shows that the parameter selection process is effective and scientific. Considering that BPNN is a deep learning algorithm, the accuracy of the model would increase with the increase of training samples. In addition to the parameter selection and calculation principle [48], the recognition accuracy of the estimation model is also related to the selected characteristic indicators and the data features and size of the selected sample due to the random selection effect of samples. The results indicate the different features in each travel purpose [9]. This reflects the necessity of individual travel chains and multimode travel knowledge graphs which conduce to capturing and extracting the travel characteristics. Moreover, similar to other inference models, researchers should consider the overfitting properties normally caused by using a large set of features, and cross validation is necessary to prevent overfitting [15]. The empirical data mining analysis in Section 3 showed that the proposed method is capable of helping us to find the behavioural features and illustrate the share of travel purposes and the relationship between the travel characteristics and the passenger categories observed in the smart card data. In addition, the multilayer BP neural network owns good adjustability and adaptability for different types of data, which means that the proposed methodology can also be used to explore the travel demands of passengers using shared travel, taxi, and intercity transportation.

This study aims to propose a systematic modelling procedure to set up the BPNN model with optimal parameters to identify the PT commuters based on mixed data, which contributes to refining passenger travel demands and helping transport operators to grasp and capture behavioural features of different travel groups observed in the smart card data. Some studies, such as the study proposed by Guo et al. [49], also have adopted similar ideas. Besides, comprehensive indicators of the spatiotemporal travel modes and travel choice characteristics were extracted to depict PT-commuter travel behavior. The relationships between the travel characteristic indicators and the passenger categories were captured through the proposed method at the individual level. Additionally, exploring the categories of PT passengers is especially conducive to the traffic management department to carry out targeted research on PT services and increase the attraction of the PT system. And the findings of this paper have been useful to augment the passenger characterization and to better cater to individual transit passengers.

There are also some limitations in this paper. The heterogeneity and categories of PT passengers were identified in this paper, but the causes and mechanisms of passengers' behavioural differences have not been revealed. In addition, the sample size is not enough due to the low matching rate of smart card and survey data to mine deeply the self-learning ability of the BPNN model, which limits the further optimization of the accuracy of the model. Thus, the travel behavior analysis of PT passengers based on traffic big data is the next work. Though the results have some boundedness due to the data sample size, this research provides a feasible method and process for identifying the PT commuters. In addition, the computational complexity that is associated with the model structure and parameters is an issue worth discussing, especially in a big data environment. The more complex the model structure and parameters are, the higher the calculation complexity is, which is reflected in the longer calculation time. However, the sample size is limited due to the low matching rate of smart card and survey data, so the computational complexity is not a prominent problem in this manuscript, while this issue will be concerned further when the travel demands of PT passengers were identified based on large-scale data in the next work.

In the future, the travelers' dependence on PT will be studied based on the achieved research basis in this paper. Then, various travel service modes for different types of passengers will be developed, such as bus rapid transit (BRT), customized bus, demand response bus, and minibus, to provide support for the refined traffic demand scheduling of operating management departments. Besides, the travel choice behaviour and the behavioural influence mechanism of PT commuters with different travel dependence on PT could be studied further.

Data Availability

The data generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant no. 52072011 and the Major Program of the National Natural Science Foundation

of China under Grant no. U1811463. The authors would like to show great appreciation for the support.

References

[1] J. Guo, X. Li, H. Wen, T. Gu, and Z. Wang, "2020 Beijing Transport Development Annual Report," Transportation Development Research Institute, Beijing, China, 2020.

[2] X. Ma, Y.-J. Wu, Y. Wang, 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.

[3] 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.

[4] Y. Yao, X. Jiang, and Z. Li, "Spatiotemporal characteristics of green travel: a classification study on a public bicycle system," *Journal of Cleaner Production*, vol. 238, Article ID 117892, 2019.

[5] S. B. Osoba, "Travel characteristics and commuting pattern of lagos Metropolis residents: an assessment," *The Indonesian Journal of Geography*, vol. 47, no. 1, pp. 40–51, 2015.

[6] L. Jun and H. Deng, "Research on passenger travel classification based on bus IC card data," *Journal of Chongqing Jianzhu University*, vol. 35, no. 6, pp. 109–114, 2016.

[7] Q. Zou, P. Zhao, X. Yao, and B. Wang, "Commuters identification for urban rail transit using automatic fare collection data," *Journal of Beijing Jiaotong University*, vol. 42, no. 3, pp. 44–52, 2018.

[8] Y. Cui, H. Qing, and K. Alireza, "Travel behavior classification: an approach with social network and deep learning," *Journal of the Transportation Research Board*, vol. 2672, no. 47, pp. 68–80, 2018.

[9] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: a data fusion approach," *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 179–191, 2014.

[10] S. Hu, Q. Liang, H. Qian, J. Weng, W. Zhou, and P. Lin, "Frequent-pattern growth algorithm based association rule mining method of public transport travel stability," *International Journal of Sustainable Transportation*, vol. 15, no. 11, pp. 1–14, 2020.

[11] S. Sun and D. Yang, "Identification of transit commuters based on naïve bayesian classifier," *Journal of Traffic and Transportation*, vol. 15, no. 6, pp. 216–221, 2015.

[12] J. Chen, K. Qi, and S. Zhu, "Traffic travel pattern recognition based on sparse Global Positioning System trajectory data," *International Journal of Distributed Sensor Networks*, vol. 16, no. 10, Article ID 15501477209, 2020.

[13] R. Rastogi and K. V. Krishna Rao, "Segmentation analysis of commuters accessing transit: Mumbai study," *Journal of Transportation Engineering*, vol. 135, no. 8, pp. 506–515, 2009.

[14] D. Zhang, X. Zhang, and J. Wang, "Commuter travel identification based on bus IC data," *Procedia-Social and Behavioral Sciences*, vol. 96, no. 6, pp. 1547–1555, 2013.

[15] M. Allahviranloo and W. Recker, "Daily activity pattern recognition by using support vector machines with multiple classes," *Transportation Research Part B: Methodological*, vol. 58, no. dec, pp. 16–43, 2013.

[16] R. Rafiq and M. G. McNally, "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, 2021.

[17] E. Manley, C. Zhong, and M. Batty, "Spatiotemporal variation in travel regularity through transit user profiling," *Transportation*, vol. 45, no. 3, pp. 703–732, 2018.

[18] L.-M. Kieu, A. Bhaskar, and E. Chung, "A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 193–207, 2015.

[19] L. M. Kieu, 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.

[20] S. Sun and D. Yang, "Identifying public transit commuters based on both the smartcard data and survey data: a case study in xiamen, China," *Journal of Advanced Transportation*, vol. 2018, pp. 1–10, 2018.

[21] G. Bösehans and I. Walker, "Do supra-modal traveller types exist? A travel behaviour market segmentation using Goal framing theory," *Transportation*, vol. 47, no. 1, pp. 243–273, 2018.

[22] X. Weng and P. Lv, "Subway IC card commuter crowd identification based on GBDT algorithm," *Journal of Chongqing Jiaotong University*, vol. 38, no. 5, pp. 8–12, 2019.

[23] W. Wang, W. Zhang, H. Guo, H. Bubb, and K. Ikeuchi, "A safety-based approaching behavioural model with various driving characteristics," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1202–1214, 2011.

[24] I. Laña, I. Olabarrieta, M. Vélez, and J. Del Ser, "On the imputation of missing data for road traffic forecasting: new insights and novel techniques," *Transportation Research Part C: Emerging Technologies*, vol. 90, pp. 18–33, 2018.

[25] B. Barabino, M. Di Francesco, and S. Mozzoni, "Time reliability measures in bus transport services from the accurate use of automatic vehicle location raw data," *Quality and Reliability Engineering International*, vol. 33, no. 5, pp. 969–978, 2016.

[26] B. Barabino, C. Lai, and A. Olivo, "Fare evasion in public transport systems: a review of the literature," *Public Transport*, vol. 12, no. 1, pp. 27–88, 2020.

[27] J. Weng, C. Wang, Y. Wang, Z. Chen, and S. Peng, "Extraction method of public transit trip chains based on the individual riders' data," *Journal of Transportation Systems Engineering and Information*, vol. 17, pp. 67–73, 2017.

[28] G. Qi, "Knowledge graph construction and reasoning," in *Proceedings of the EEE International Conference on Progress in Informatics and Computing*, pp. 17-18, Shangai, China, December 2016.

[29] Q. Liang, J. Weng, W. Zhou, S. B. Santamaria, J. Ma, and J. Rong, "Individual travel behavior modeling of public transport passenger based on graph construction," *Journal of Advanced Transportation*, vol. 2018, pp. 1–13, 2018.

[30] K. Y. Chan, T. S. Dillon, J. Singh, and E. Chang, "Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and l-marquardt algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 644–654, 2012.

[31] V. L. Bernardin, S. Trevino, G. Slater, and J. Gliebe, "Simultaneous travel model estimation from survey data and traffic counts," *Journal of the Transportation Research Board*, vol. 2, no. 2494, pp. 69–76, 2015.

[32] X. Ke, L. Shi, W. Guo, and D. Chen, "Multi-dimensional traffic congestion detection based on fusion of visual features and convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2157–2170, 2019.

[33] W. W. Liu and C. S. Shi, "Analysis of university science research capability elements and evaluation based on BP neural

network," *Journal of Service Science and Management*, vol. 1, no. 3, pp. 266–271, 2008.

[34] L. J. Li and W. Huang, "A short-term power load forecasting method based on BP neural network," *Applied Mechanics and Materials*, vol. 494-495, no. 495, pp. 1647–1650, 2014.

[35] J. Li, D. Zhao, B. Ge, K. Yang, and Y. Chen, "A link prediction method for heterogeneous networks based on BP neural network," *Phys. A Stat. Mech. its Appl*, vol. 491, pp. 1–17, 2017.

[36] L. Guo, J. Gao, J. Yang, and J. Kang, "Criticality evaluation of petrochemical equipment based on fuzzy comprehensive evaluation and a BP neural network," *Journal of Loss Prevention in the Process Industries*, vol. 22, no. 4, pp. 469–476, 2009.

[37] Z. Y. Bing and D. Y. Xing, "Design of BP network image predictor based on MATLAB," *Computer Simulation*, vol. 24, no. 3, pp. 223–226, 2007.

[38] A. R. Soleymani, V. Moradi, and J. Saien, "Artificial neural network modeling of a pilot plant jet-mixing UV/hydrogen peroxide wastewater treatment system," *Chemical Engineering Communications*, vol. 206, pp. 1–13, 2018.

[39] F. Nabizadeh Chianeh, J. Basiri Parsa, and H. Rezaei Vahidian, "Artificial neural network modeling for removal of azo dye from aqueous solutions by Ti anode coated with multiwall carbon nanotubes," *Environmental Progress & Sustainable Energy*, vol. 36, no. 6, pp. 1778–1784, 2017.

[40] Z. H. Zhang, F. Min, G. S. Chen, S. P. Shen, Z. C. Wen, and X. B. Zhou, "Tri-partition state alphabet-based sequential pattern for multivariate time series," *Cognitive Computation*, pp. 1–19, 2021.

[41] L. I. Ruimin, L. U. Huapu, and Q. Shi, "ANN-based prediction of turning rate of traffic flows at intersection," *Journal of Southwest Jiao Tong University*, vol. 7, no. 1, pp. 1–34, 2007.

[42] C. I. Muntean, F. M. Nardini, F. Silvestri, and R. Baraglia, "On learning prediction models for tourists paths," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 1, pp. 1–34, 2015.

[43] S. Maher and P. Biswajeet, "Severity prediction of traffic accidents with recurrent neural networks," *Applied Sciences*, vol. 7, no. 6, pp. 476–493, 2017.

[44] E. J. Adams, M. Goad, S. Sahlqvist, F. C. Bull, A. R. Cooper, and D. Ogilvie, "Reliability and validity of the transport and physical activity questionnaire (TPAQ) for assessing physical activity behaviour," *PLoS One*, vol. 9, no. 9, Article ID e107039, 2014.

[45] G. Sun, C. Webster, and A. Chiaradia, "Objective assessment of station approach routes: development and reliability of an audit for walking environments around metro stations in China," *Journal of Transport & Health*, vol. 4, pp. 191–207, 2017.

[46] H. Q. Du and W. W. Fan, "The application of self-organizing neural network to remote sensing image classification based on matlab," *Journal of Northeast Forestry University*, vol. 32, no. 4, pp. 51–53, 2003.

[47] E. Q. Wu, M. C. Zhou, D. Hu, L. Zhu, and H. Ren, "Self-paced dynamic infinite mixture model for fatigue evaluation of pilots' brains," *IEEE Transactions on Cybernetics*, no. 99, pp. 1–16, 2020.

[48] A. Boubezoul and S. Paris, "Application of global optimization methods to model and feature selection," *Pattern Recognition*, vol. 45, no. 10, pp. 3676–3686, 2012.

[49] M. Guo, P. Wang, and L. Zhao, "Research on recognition method of transportation modes based on deep learning," *Journal of Harbin Institute of Technology*, vol. 51, no. 11, pp. 1–7, 2019.