

Article

Two-Stage Data-Driven Trip Purposes and Socio-Economic Attribute Inference Framework of Public Transit Riders

Chao Yang ^{1,2}, Wentao Dong ¹, Chengcheng Yu ², Yuliang Zhang ³ and Yingfei Tu ^{1,2,*} 

¹ The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, 4800 Cao'an Road, Shanghai 201804, China

² Urban Mobility Institute, Tongji University, 1239 Siping Road, Shanghai 200092, China

³ College of Civil Engineering and Architecture, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, China

* Correspondence: tuyingfei@tongji.edu.cn

Abstract

Data-driven research is becoming a new paradigm in transportation, but the natural lack of individual socio-economic attributes in transportation data makes research such as activity purpose inference and mobility pattern identification lack convincingness and verifiability. In this paper, a two-stage trip purpose and socio-economic attributes inference model is proposed based on travel resident survey and smart card data. In the first stage, the trip purpose of each trip is inferred by a combination of rule-based and XGBoost models. In the second stage, based on the trip purpose, a machine-learning model is built to infer the socio-economic attributes of individuals. A teacher–student model based on self-training is then applied to the models above to transfer them to smart card data. The impact of independent variables of the socio-economic attributes inference model is also investigated. The results show that the proposed framework achieves an overall accuracy of 92.7% for trip-purpose inference, and an average accuracy of 76.3% across the three socio-economic attributes (age, job status, and income). Travel time, arrival time, departure time, and purpose of the first two trips are the most important factors on age and job status, while the land price related to jobs and housing is significant when inferring individuals' incomes.

Keywords: transit; traffic big data; trip purpose inference; intelligent transportation; socio-economic attributes; teacher–student model

1. Introduction

Urban transit systems have been widely acknowledged to be essential to transportation sustainability in the way of alleviating the pressure of urban traffic congestion and environmental pollution. However, the current mismatch between the allocation of urban public transportation capacity and passenger demand has resulted in a waste of public resources, making it urgent to accurately grasp the characteristics of passenger demand [1–3]. To obtain a deeper insight into people's travel mode choices and the complexity of travel purposes, various information such as GPS data, smart card data, and social media data have been used to establish more precise disaggregated models, which have great value in understanding human mobility and travel behavior at both individual and aggregate levels [4].

Although a mass amount of data from different sources has been generated and collected from different sources, most of it lacks information related to the trip purpose or the subjects' socio-economic attributes, making these big data difficult to apply in travel



Academic Editor: Corrado Rindone

Received: 17 June 2026

Revised: 18 July 2026

Accepted: 21 July 2026

Published: 23 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

behavior research. Researchers have found that socio-economic attributes have proven to play important roles in human travel behavior [5]. Attributes such as age [6,7], gender [8], income [9], occupational status [10], and educational level [11] have great influence on people's spatiotemporal characteristics and selection of travel modes. Additionally, activity-based models (ABMs) have been carried out to capture the nuances of individual activity-travel behavior and homogeneity and heterogeneity between individuals [12,13]. ABMs focus on individuals' mechanism of travel generation and as a result place an emphasis on trip purposes [14]. On the other hand, manually collected datasets, such as household survey data, do include trip purpose and socio-economic attribute information, but these data always have a relatively small quantity of this information compared with those that are auto-collected. If the advantages of these two types of data are combined, we can create a large dataset with labeled trip purpose and socio-economic attributes, providing a better foundation to explore the relationship between travel demand and social and economic development.

In recent years, many approaches have been developed to deduce the socio-economic attributes of individuals in different kinds of data. Some researchers have proposed rule-based models [15], machine-learning models such as Bayesian conditional probability models [16], and deep-learning models [17,18] to infer the socio-economic attributes of users. However, this research has three shortcomings. First, these models were trained and validated on the same dataset and failed to consider the method of transferring the model from a small dataset to a large dataset. Second, the travel purpose is not explicitly included in the model, making it difficult to look into the relationships between socio-economic attributes and activities of individuals. Lastly, little research has considered the interpretability of inference models, thus hindering the understanding of the relationship between travel patterns and socio-economic attributes.

Considering these shortcomings in the existing studies, this paper proposes a two-stage method for inferring both travel purposes and socio-economic attributes of regular transit users. The first stage is to infer the purpose of each trip based on resident travel survey data and POI data by using a combined rule-based and machine-learning model, and a self-training teacher–student cycling strategy is built to transfer the model onto smart card data. Then, in the second stage, features including temporal and spatial travel characteristics, land price, population density, and trip purpose are used to build the socio-economic attributes inference model. The self-training strategy is also applied in this part. Finally, these models are compared with other existing models, and the effectiveness of the model is validated. Some interpretability analysis was also conducted. The contributions of this work lie in the following two aspects:

- (1) Proposing a novel two-stage joint inference framework that integrates trip purpose inference and socio-economic attribute inference, providing a holistic solution for enriching large-scale unlabeled smart card data with activity and demographic information;
- (2) Applying a teacher–student self-training strategy within the proposed framework to transfer knowledge from labeled survey data to unlabeled smart card data, and systematically investigating the interpretability of socio-economic attribute inference through SHAP value analysis, which provides behavioral insights into the relationship between travel patterns and individual attributes.

The rest of the paper is organized as follows. Section 2 presents a literature review on the inference of trip purpose and socio-economic attributes. Section 3 describes the datasets used in this study, mainly the smart card dataset and the resident travel survey dataset, and the overall framework of inference of trip purpose and socio-economic attributes. Section 4 evaluates the effectiveness of the model. Lastly, Section 6 concludes the paper with an outlook.

2. Literature Review

2.1. Trip Purpose Inference

At present, there have been many studies on the identification of travel purposes, and existing articles involve the identification of travel purposes from different data sources, including mobile GPS data [19,20], taxi trajectory data [21,22], and bus swiping-card data [23,24]. Although the sources of data vary, for identification most research methods use spatial geographic attributes near boarding and alighting, and the time of boarding and alighting.

In previous research, there were relatively few classifications of travel purpose. Chakirov and Erath [25] simply divided travel behavior into two categories: going to work and going home. Subsequently, Devillaine et al. [26] further refined travel behavior into three categories: going to work, going to school, and going home. With the development of data mining technology, the accuracy of classification has improved, and the number of travel purposes has been further divided. Xiao, Juan and Zhang [20] classified travel purposes into six categories based on mobile GPS data: picking up and dropping off people, social visits, shopping, dining, going to work (school), and returning home. Gong, Liu, Wu and Liu [22] proposed nine different travel purposes based on taxi trajectory data, including going home, working, transferring, eating, shopping, entertainment, going to school, accommodation, and seeking medical treatment.

In terms of modeling, Stopher et al. [27] proposed a travel purpose inference method based on custom rules, which can successfully identify over 60% of travel records. Alsger, Tavassoli, Mesbah, Ferreira and Hickman [24] adopted an inference method based on conditional probability, and the accuracy of the discrimination method using spatiotemporal conditional probability can reach 78%. Chen et al. [28] proposed a travel purpose recognition method based on the logit model, proposing four types of travel purposes with discrimination accuracies of 67% and 78% for both home and non-home groups, respectively. Xiao, Juan and Zhang [20] proposed a discrimination method based on artificial neural networks, which achieved a maximum accuracy of 97.22% on the test set. Sari Aslam et al. [29] established the ActiveNET model based on artificial neural networks and deep learning, using labeled data for prediction, with an overall accuracy rate of 94%.

2.2. Socio-Economic Attributes Inference

With the application of big data analysis in the field of transportation, many studies have begun to investigate individuals involved in transportation and classify them into different groups. Related research focused on two directions: individual activity patterns (spatiotemporal characteristics) and socio-economic attributes [30]. Extensive research has been conducted on individual activity patterns [31–34], while research on socio-economic attributes is also broad. Previous studies have shown that individuals' socio-economic attributes have strong impacts on their travel patterns [35,36]. Therefore, it is necessary to establish accurate and effective models for identifying socio-economic attributes.

In previous studies, researchers have proposed various models for inferring socio-economic attributes based on data related to transportation. Solomon et al. [37] used trajectory data to establish a Word2Vec model, extracted features from trajectory data, and predicted user attributes such as gender, age, and marital status. Results show that the proposed method achieves around 80% accuracy for various demographic prediction tasks. Kim, Kim and Sohn [18] proposed a novel method to estimate the trip purpose and socio-demographic attributes of smart card data by using a conditional generative adversarial network (CGAN), reaching an overall accuracy of 68.55%. Wu et al. [38] extracted spatiotemporal features from GPS trajectories and attached geographical context to trajectories to reflect people's activities. Supervised classification models are estab-

lished and results show the inference accuracies of demographics range from 66% to 80%. Zhang et al. [39] employed a multi-task convolutional neural network (CNN) on a smart card dataset provided by Transport for London. It shows that the highest accuracies of car ownership, age group, income, and gender inference achieve 80%, 76%, 69%, and 64%, respectively. Zhao et al. [40] draws upon the inverse discrete choice modeling (IDCM) approach to enrich people's socio-demographic attributes. The accuracies vary from 50% to 90%. Zhang et al. [41] proposed a method that first calculates the similarity of people's multidimensional daily activities and then applies dynamic time warping to measure the similarities of the multidimensional activity sequences. The results show that the proposed method outperforms existing methods in predicting four selected demographics: gender, age, education level, and work status, with an accuracy range between 91% and 94% for the national dataset and 88% to 91% for the GPS data.

Although so many different kinds of models have been proposed, few of them include the analysis of interpretability of models. Also, most of the existing studies used GPS data, which achieved good inference accuracy due to its fine granularity, ranging from 66% to 94%. However, smart card data is relatively sparse in time and space, containing little information. Therefore, the accuracy of socio-economic attribute estimation based on smart card data is relatively low, essentially not exceeding 80%.

In summary, the inference of trip purpose and socio-economic attributes is usually separated, and trip purpose is not considered when inferring socio-economic attributes. A joint inference model of travel purposes and socio-economic attributes needs to be developed to reveal the relationship between them and improve the accuracy of the model.

3. Materials and Methods

3.1. Data Description

In this study, the city of Shenzhen, China is chosen as a case study for empirical research. As of 2018, Shenzhen's public transportation contains 981 regular bus lines, 5331 pairs of bus stops, 21,000 km of regular bus operating line length, and an average daily passenger volume of 4.463 million. To establish an inference model for travel purposes and socio-economic attributes, resident travel survey data, smart card data, and Shenzhen's POI data are used in this research. The detailed descriptions of these three types of data are listed below.

Resident Travel Survey Data (RTSD). The RTSD consists of one-day resident travel information obtained from a face-to-face survey conducted in the fourth quarter of 2020. Seeing that COVID-19 had already been brought under control in China at that time, its impact on people's travel was negligible (It should be acknowledged that a temporal gap of approximately two years exists between the RTSD (Q4 2020) and SCD (October 2018). Although the COVID-19 pandemic was effectively controlled in China by late 2020 and routine travel patterns had largely recovered, potential systematic differences in travel behavior between the two periods may exist. To address this concern, we compared general bus travel characteristics in Shenzhen between 2018 and 2020 based on publicly available transit operation statistics. To further examine the comparability of the two datasets used in this study, we compared the trip-duration and departure-time distributions of the RTSD and SCD. Both datasets show that most bus trips are completed within one hour, and both exhibit clear morning and evening peak periods, mainly concentrated around 6:00–9:00 and 17:00–20:00. These similarities suggest that the main temporal structure of regular bus travel remained relatively stable. Nevertheless, the SCD contains a higher proportion of noon and late-evening trips, indicating that the RTSD may underrepresent some low-frequency and non-commuting bus users. This difference is explicitly considered when interpreting the model transfer results. Nevertheless, the potential impact of this temporal

mismatch on model transferability is recognized as a limitation and is further discussed in Section 6. It is also noteworthy that all of the surveys were carried out on workdays. The survey questions were divided into three categories: personal socio-economic attributes (e.g., gender, age, income, job type, car ownership, education, home location, work location) and travel-related attributes (e.g., departure time, arrival time, departure location, arrival location, activity type, travel mode). The origin data contain trips of different travel modes, but only trips by regular transit are used in order that we can focus on the travel of regular transit users. Finally, 20,802 trips of 5996 users of regular public transit are selected in this dataset. The socio-economic information of these users and the purposes of the trips they took are summarized in Table 1. The raw data contains 11 types of travel purposes, which are categorized into 4 purposes that are most relevant to the socio-economic attributes. Further, some of the socio-economic attributes have been merged to reduce the complexity of the model.

Smart Card Data (SCD). A one-month regular transit smart card dataset from Shenzhen, covering the period from 1 October to 31 October 2018, is used to validate the effectiveness of the proposed model. It includes card-swiping records from 12,979,574 users. Each trip contains the following fields: user ID, date, route, trip start and end time, boarding stop and alighting stop, trip start and end latitude and longitude. The data pre-processing eliminates some missing data and abnormal data, and transfer behaviors are also recognized and merged so that each record represents a trip instead of a ride. A survey revealed that more than 94% of the transfer activities took less than 60 min [42], which is taken in this study as a threshold to represent transfer activity between two rides.

POI Data. The POI data for Shenzhen in 2020 was obtained from Gaode Map API to add geospatial information to each trip.

Population Data and Land-Price Data. The population data and land-price data of Shenzhen in 2020 was obtained from WorldPop (<https://hub.worldpop.org/> (accessed on 15 July 2023)) and Lianjia (<https://sz.lianjia.com/> (accessed on 15 July 2023)), respectively.

Table 1. Descriptive statistics of socio-economic and trip purposes of RTSD.

Attributes	Category	Proportion (%)
Socio-Economic Attributes		
Gender	Male	48.0
	Female	52.0
Age	<20	17.3
	20–59	80.0
	≥60	3.7
Income	0	11.8
	0–10	60.6
	10–15	19.8
	≥15	7.8
Job Status	With job	69.8
	Student	15.6
	Retired and No job	14.6
Car ownership	Yes	32.7
	No	67.3
Trip Purpose		
Trip Purpose	Work	41.4
	Home	48.5
	Shopping and Entertainment (S&E)	9.6
	Medical	0.5

3.2. Overall Framework of the Model

In this part, a two-stage method for inferring both travel purposes and socio-economic attributes of regular transit users is proposed. The overall framework of the model is shown in Figure 1. Because the RTSD only includes one-day travel data for each resident, the residents with records of public transportation trips are more likely to be high-frequency users of public transportation. Moreover, the resident travel survey was carried out on weekdays, while the smart card data covers trips on both weekdays and weekends. These two differences between the RTSD and SCD make it difficult to establish a unified model on both datasets. Therefore, a semi-supervised self-learning-based method is proposed. In the first stage, both the RTSD and SCD are divided into two categories based on whether the homes and workplaces of the users can be recognized and two initial models are established on the RTSD. After that, by utilizing SCD for self-training, the trip purposes on the SCD can be inferred. In the second stage, the predicted trip purposes are then applied to establish the features of one-day trip chains, and again a self-learning-based model is utilized to infer socio-economic attributes on the SCD. Finally, considering multiple days of travel chain data, the socio-economic attributes of users are determined using the maximum probability estimation method.

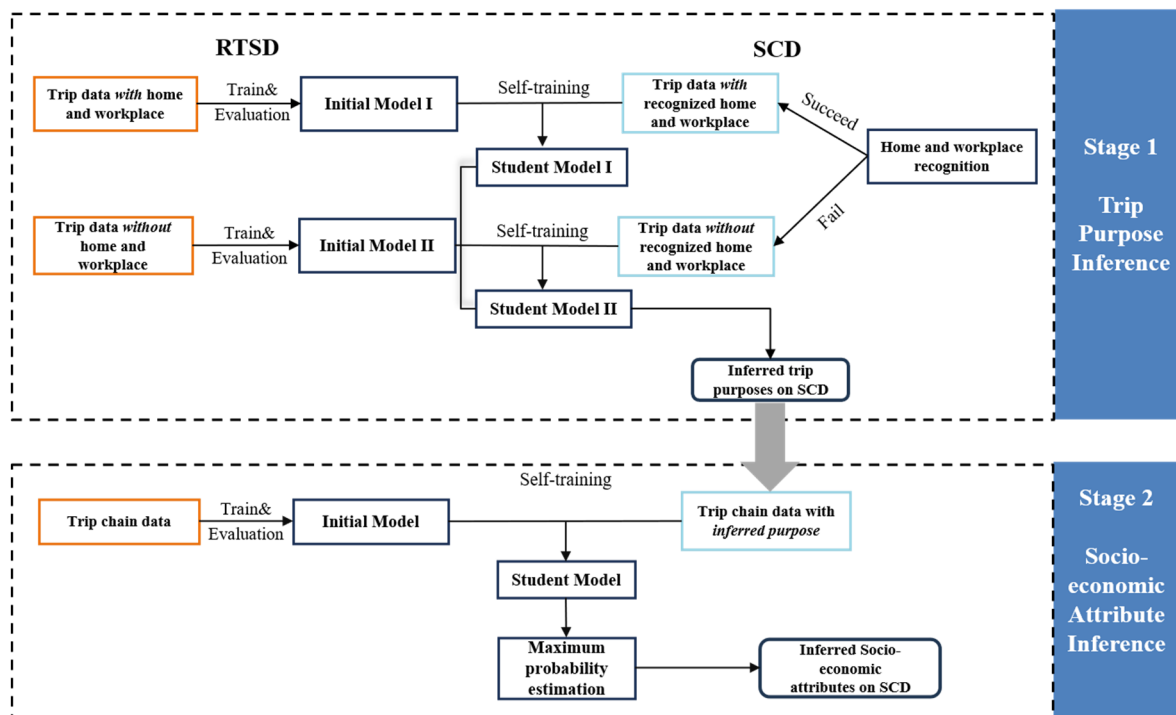


Figure 1. Overall model framework.

3.3. Initial Models for Inferring Trip Purposes

The anchor point theory was proposed by Golledge R [43], suggesting that the activity space of urban residents is framed by “anchor points” (frequently visited activity points such as homes and companies). The research pointed out that the frequency of trips between anchor points is much more than the frequency of other trips. As shown in the travel survey data of Shenzhen residents, 89.9% of trips are related to home or workplace. Therefore, the process of inferring trip purposes is shown in Figure 2. Whether the location of the traveler’s home and workplace can be identified is of crucial importance for deducing the purpose of travel. If a user’s home and workplace can be identified, we first identify their work and home purposes based on a rule-based model and then recognize the remaining purposes. Otherwise, all four purposes will be directly identified. Because the SCD lacks

information on activities between trips, home and workplace must be inferred solely from the regularity of passengers' boarding and alighting stops. The methods for identifying home and workplace are primarily based on the study by Ma et al. [44] and have been improved upon since then. The rules of identifying home station s_{Home} and workplace station s_{Work} are shown by Equations (1) and (2).

$$s_{Home} = \arg \max_{s_{first} \in \{s_1, s_2, \dots, s_n\}} f(s_{first}) \text{ if } f(s_{first}) > 0.6 \quad (1)$$

$$s_{Work} = \arg \max_{s_{last} \in \{s_1, s_2, \dots, s_n\}} f(s_{last}) \text{ if } f(s_{last}) > 0.6 \quad (2)$$

where s_{first} is the station where users first board buses in daily travels (departing before 3 p.m.), s_{last} is the station where users last board buses in daily travels (departing after 3 p.m.).

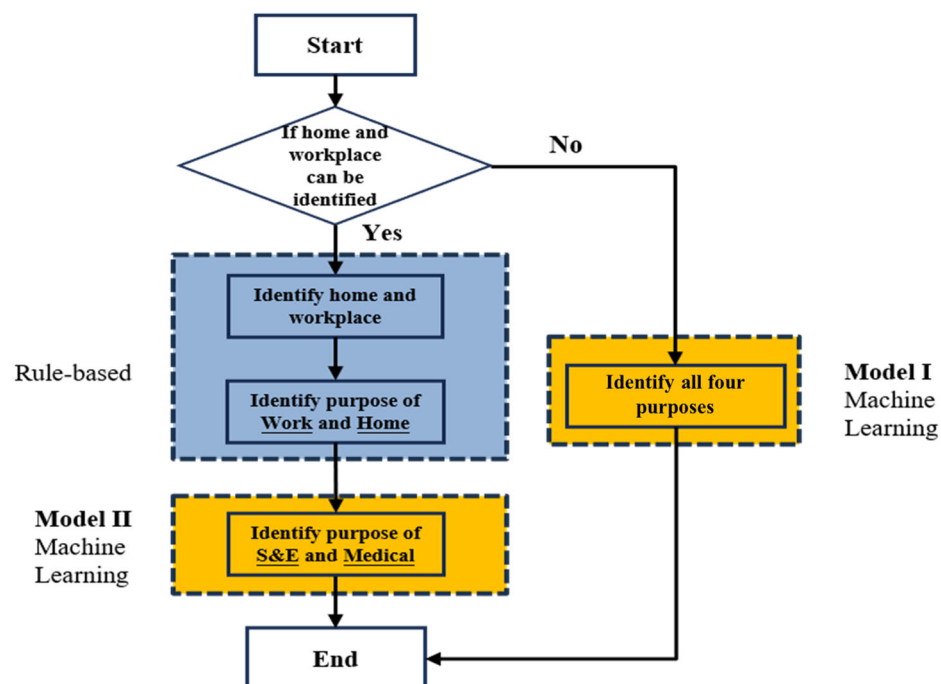


Figure 2. Diagram showing the travel purpose identification process.

A threshold of 0.6 in Equations (1) and (2) was used to ensure that the inferred home or workplace station appeared in the majority of a user's observed travel days. A lower threshold would increase the number of identified users but may introduce more false home/workplace locations, while a higher threshold would retain only highly regular commuters and reduce coverage. The 3 p.m. cutoff was used to distinguish morning-to-midday first departures and afternoon-to-evening last departures, which is consistent with the typical daily commuting structure observed in Shenzhen. In sensitivity checks using alternative frequency thresholds and time cutoffs, the identified home/workplace patterns were generally stable, and 0.6 with a 3 p.m. cutoff provided a reasonable balance between identification reliability and user coverage.

For POI processing, we used the Gaode Map POI data and selected POI categories that are behaviorally relevant to infer the purpose of the bus trip, including catering, education, leisure, shopping, hospital, company, and residence. Following common practice in public-transit station catchment analysis, the surrounding built environment for each boarding and alighting stop was measured within a 500 m buffer around the stop. For each stop, the number of POIs in each category was divided by the total number of selected POIs within the buffer, and the resulting category proportions were used as land-use features.

Proportional features were adopted instead of raw POI counts to reduce the influence of local POI density differences across urban areas.

Accordingly, if the alighting stop is the workplace, the purpose of the trip will be identified as work. Similarly, if the alighting stop is home, the purpose of the trip will be identified as work (corresponding to the rule-based model in Figure 2).

Then, concerning the trip purposes of the users whose home and workplace cannot be recognized (corresponding to Model I in Figure 2) or for the purpose of S&E and Medical for users who have identified home and workplace (corresponding to Model II in Figure 2), the XGBoost model is applied. XGBoost is an improved gradient-boosting decision algorithm created by Chen and Guestrin [45], which has high computational efficiency allowing for fast processing of large datasets, and its built-in regularization techniques help prevent overfitting and improve model generalization. The objective function of XGBoost is represented by Equation (3):

$$J(\theta) = L(\theta) + \Omega(\theta). \quad (3)$$

where L is the training loss function and Ω is the regularization function.

To solve the problem of an imbalanced distribution of sample categories, a weighted cross entropy function is used as the loss function in the model. The detailed loss function and regularization function is shown by Equations (4) and (5):

$$L = -\frac{1}{N} \sum_{i=1}^N \frac{1}{n_{y_i}} \log(\hat{p}_{i,y_i}) \quad (4)$$

$$\Omega = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (5)$$

where N represents the sample size, n_{y_i} represents the number of samples belonging to category y_i , $\hat{p}_{i,y_i}$ represents the predicting probability of sample i belonging to category y_i , T represents the depth of trees, and ω represents the vector of scores of leaves.

Finally, temporal and land-use features are included as the independent variables in the XGBoost model for inferring the trip purposes. The variables of boarding time and alighting time are transformed into cosine values to capture the periodic features of the data [46] and thus improve the model's generalization capability. The calculation method is shown by Equation (6):

$$\text{COS}_t = \cos\left(\frac{2\pi t}{T}\right) \quad (6)$$

where t represents the seconds of time from 0:00 on that day, T represents the whole seconds of a day. The independent variables of the trip purpose model and their descriptions are illustrated in Table 2.

Table 2. Independent variables of the trip purpose model.

Variable	Description
cos_boardtime	Cosine value of boarding time
cos_alighttime	Cosine value of alighting time
time_interval	Duration of the trip
board_catering	Proportion of catering POIs among all POIs at the boarding stop
board_education	Proportion of education POIs among all POIs at the boarding stop
board_leisure	Proportion of leisure POIs among all POIs at the boarding stop
board_shopping	Proportion of shopping POIs among all POIs at the boarding stop
board_hospital	Proportion of hospital POIs among all POIs at the boarding stop

Table 2. Cont.

Variable	Description
board_company	Proportion of company POIs among all POIs at the boarding stop
board_residence	Proportion of residence POIs among all POIs at the boarding stop
alight_catering	Proportion of catering POIs among all POIs at the alighting stop
alight_education	Proportion of education POIs among all POIs at the alighting stop
alight_leisure	Proportion of leisure POIs among all POIs at the alighting stop
alight_shopping	Proportion of shopping POIs among all POIs at the alighting stop
alight_hospital	Proportion of hospital POIs among all POIs at the alighting stop
alight_company	Proportion of company POIs among all POIs at the alighting stop
alight_residence	Proportion of residence POIs among all POIs at the alighting stop

3.4. Self-Training Model for Inferring Trip Purpose

Due to the previously mentioned significant differences between the RTSD and SCD, a semi-supervised self-training model inspired by Gao et al. [47] is utilized to infer users' socio-economic attributes based on the SCD. Self-training is particularly useful in scenarios where there are limited labeled examples but an abundance of unlabeled data available for training [48], and unlabeled data can help models learn by explicitly defining the structure of the data through a manifold [49]. In this way, large-scale unlabeled smart card data can be utilized and generalized within a smart card dataset.

The framework of the self-training method is illustrated in Figure 3. The model training process can be divided into two steps. The first step involves training an initial teacher model (ITM) on the RTSD using supervised learning methods. Then, the ITM is used to predict the activity purposes of the trips in the SCD. When the confidence of the predictions exceeds set thresholds, the predicted types are applied to the unlabeled data as pseudo-labels, and the confidence threshold is denoted as τ . Considering various learning status and learning difficulties within different classes, curriculum pseudo-labeling (CPL) [50] is used to determine the threshold of labeling. The core of CPL is to flexibly adjust thresholds for different classes at each time step, thus ensuring informative and reliable unlabeled data and their pseudo-labels. The calculation of confidence threshold T is illustrated by Equations (7)–(9):

$$\sigma_t(c) = \sum_{n=1}^N I(\max(p_{m,t}(y|u_n)) > \tau) \cdot I(\operatorname{argmax}(p_{m,t}(y|u_n)) = c) \quad (7)$$

$$\beta_t(c) = \frac{\sigma_t(c)}{\max_c \sigma_t} \quad (8)$$

$$T_t(c) = \beta_t(c) \cdot \tau \quad (9)$$

where $p_{m,t}(y|u_n)$ is the model's prediction for unlabeled data u_n at time step t , N is the total number of unlabeled data, $\sigma_t(c)$ reflects the number of high-quality sample of class c learned at time step t , and τ is the fixed threshold.

In our experiments, we observed that during the initial phase of training, the model tends to predict most unlabeled samples as belonging to a specific class, which is influenced by the parameter initialization. This behavior may cause a higher likelihood of confirmation bias, making the learning status estimates unreliable. Therefore, a warm-up process is introduced by rewriting the denominator in Equation (10) as:

$$\beta_t(c) = \frac{\sigma_t(c)}{\max\{\max_c \sigma_t(c), N - \sum_c \sigma_t(c)\}} \quad (10)$$

where the term $N - \sum_c \sigma_t(c)$ can be regarded as the number of unlabeled data that have not been used. This ensures the values of $\beta_t(c)$ increase gradually and prevents the model from categorizing all samples into the same type.

The second step involves the teacher–student switching model (TSSM), which is trained using a mix of labeled RTSD and pseudo-labeled SCD. Subsequently, the student model is used as new teacher model to generate predictions on the unlabeled SCD, which is used in the next training epoch. This alternating process is repeated until the model's performance on the new dataset reaches its best metrics. The ITM is trained on the RTSD and the TSSM is developed on the SCD. The detailed process is shown in Algorithm 1.

Algorithm 1. Self-training teacher–student model training process

```

1: Input:  $D_{SCD} = \{(x_m, y_m), m \in (1, 2, \dots, M)\}$ ,  $D_{RTSD} = \{u_n, n \in (1, 2, \dots, N)\}$  (M labeled data and N
unlabeled data.)
2: ITM Training
3: Create a supervised learning model as the initial teacher model  $M_t$ .
4: Initialize parameters  $\theta_t$  of initial teacher model  $M_t$ .
5: Train  $M_t$  by minimizing weighted cross-entropy loss (Equation (3)–(5)) on labeled dataset  $D_{RTSD}$ .
6: while not converged do
7:    $\theta_t \leftarrow \theta_t - \nabla_{\theta_t} \text{Loss}(D_{RTSD}(\theta_t))$ 
8: end while
9: TSSM Training
10: while not reach the maximum iteration do:
11:   for  $c = 1$  to  $C$  do
12:     Calculate  $\sigma(c)$  using Equation (7).
13:     if  $\max \sigma(c) < \sum_{n=1}^N I(\hat{u}_n = -1)$  then
14:       Calculate  $\beta(c)$  using Equation (10).
15:     else
16:       Calculate  $\beta(c)$  using Equation (8).
17:     end if
18:     Calculate  $T(c)$  using Equation (9).
19:   end for
20:   Predict activity type on the unlabeled dataset  $D_{SCD}$  by  $M_t$ . If the confidence level of predicted activity
type  $> T(c)$ , the predicted activity type is adopted as ‘pseudo-label’.
21:   Get  $D_{mixed}$  by combining  $D_{pseudo}$  with  $D_{RTSD}$ 
22:   Create an equal model as a student model  $M_s$ .
23:   Initialize parameters  $\theta_s$  of student model  $M_s$ .
24:   Train  $M_s$  by minimizing weighted cross-entropy loss on  $D_{mixed}$ 
25:   while not converged do
26:      $\theta_s \leftarrow \theta_s - \nabla_{\theta_s} \text{Loss}(D_{mixed}(\theta_s))$ 
27:   end while
28:   if metrics perform better than the last iteration then
29:     close
30:   else
31:     Set the student model  $M_s$  as the new teacher model  $M_t$ 
32:     goto while
33:   end if
34: end while

```

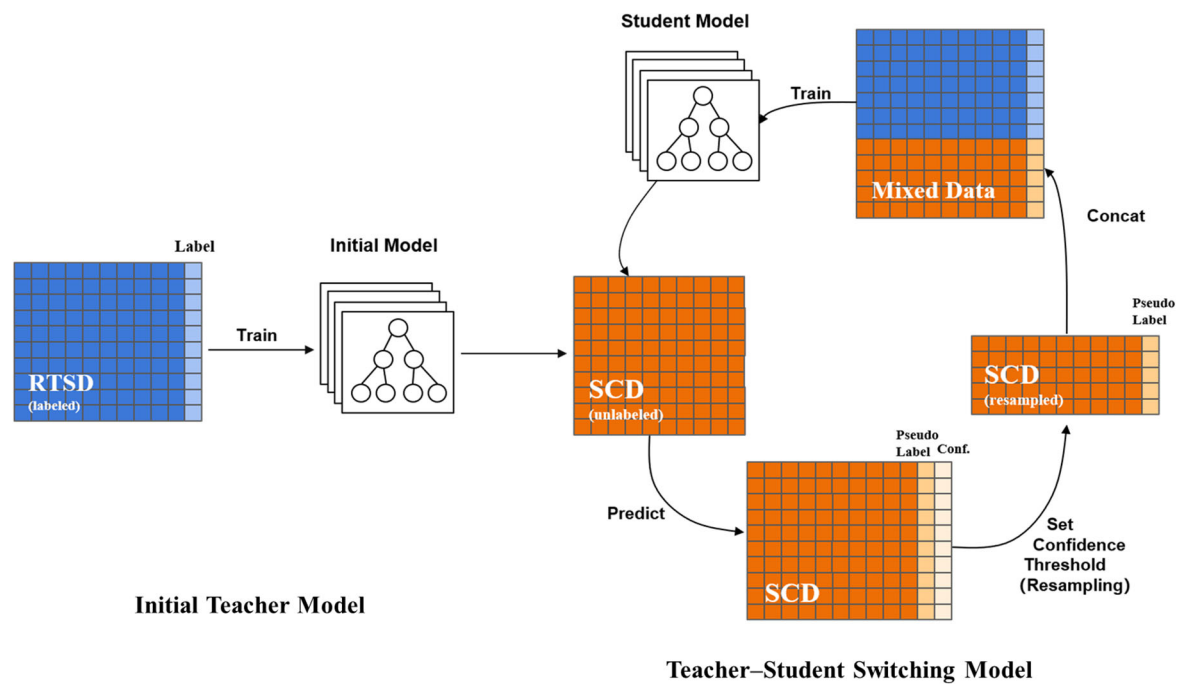


Figure 3. Framework of the self-training method.

3.5. Models for Inferring Socio-Economic Attributes

After finishing inferring the purpose of each trip, the model for inferring socio-economic attributes of regular transit users is proposed in this part. Much research involves the concept of social segregation, indicating that people who have the same socio-economic attributes tend to have similar temporal and spatial attributes of travel and live in similar communities, especially in big cities [51,52]. Because resident survey data only contains one-day travel information of a single user, we can only obtain the relationship between daily travel and the user's socio-economic attributes. Therefore, a daily travel chain is established to infer the user's socio-economic attributes. As more than 99% of transit users make five or fewer trips in one day in the RTSD, the spatial-temporal information about the users' first five trips in one day are selected in the feature engineering (if a user traveled fewer than five times in a day, the variables of the unfinished trips is set as -1). The variables about spatial information of the user's home and workplace, such as population density and land price, are also included. Shenzhen city is divided spatially into $1\text{ km} \times 1\text{ km}$ grids and numbered. The departure and arrival position of each trip is represented by the grid number and coded as cosine values, which is shown as Equations (11) and (12):

$$\text{COS}_x = \cos\left(\frac{2\pi x}{X}\right) \quad (11)$$

$$\text{COS}_y = \cos\left(\frac{2\pi y}{Y}\right) \quad (12)$$

where x, y represents the number of the grid on the x -axis and y -axis, respectively. X and Y represent the total number of grid columns and rows on x -axis and y -axis, respectively. The variables of socio-economic attributes model are shown in Table 3.

After the feature engineering, a socio-economic attribute inference method similar to travel purpose inference is built. All the RTSD is initially placed into the XGBoost model for training, and then pseudo-labels are generated on the SCD based on the self-training strategy (mentioned in the previous section). Considering confidence threshold control,

the mixed data is used for the next round of model training to obtain the socio-economic attribute inference model based on one-day travel chains.

Although socio-economic attributes are inferred based on one-day trip chains, we want to obtain the actual socio-economic attributes of regular transit users. Thus, a rule is proposed that the user's actual socio-economic attributes are the ones that appear most frequently among all the attributes inferred during the long term, as shown in Figure 4. Assuming that the multi-day inferred travel purposes of user i are $C_i = \{c_i^1, c_i^2, \dots, c_i^n\}$, the actual socio-economic attribute is defined by Equation (13):

$$c_i^{real} = \underset{c \in C}{\operatorname{argmax}} f(c_i) \quad (13)$$

where $f(c_i)$ represents the occurrence number of c_i in set C . In this way, the actual socio-economic attributes of users can be uniquely determined.

Table 3. Variables in the socio-economic attributes model.

Category	Variable	Description
Independent Variables		
Spatial Information of Home (4)	Home_X	Home's X cosine value of the grid
	Home_Y	Home's Y cosine value of the grid
	Home_Pop	Population density of home location grid
	Home_LP	Land Price of home location grid
Spatial Information of Workplace (4)	Work_X	Workplace's X cosine value of the grid
	Work_Y	Workplace's Y cosine value of the grid
	Work_Pop	Population density of workplace location grid
	Work_LP	Land Price of workplace location grid
Information of Trips (11 × 5)	Dep_Time1	Departure time of the 1st trip of the day (cosine value)
	Arr_Time1	Arrival time of the 1st trip of the day (cosine value)
	Travel_Time1	Travel time of the 1st trip of the day
	Dep_X1	X cosine value of the grid of 1st trip's departure location
	Dep_Y1	Y cosine value of the grid of 1st trip's departure location
	Arr_X1	X cosine value of the grid of 1st trip's arrival location
	Arr_Y1	Y cosine value of the grid of 1st trip's arrival location
	Purpose1_1	If the trip purpose of 1st trip is Work
	Purpose1_2	If the trip purpose of 1st trip is Home
	Purpose1_3	If the trip purpose of 1st trip is S&E
	Purpose1_4	If the trip purpose of 1st trip is Medical
		
	Dep_Time5	Departure time of the 5th trip of the day (cosine value)
	Arr_Time5	Arrival time of the 5th trip of the day (cosine value)
	Travel_Time5	Travel time of the 5th trip of the day
	Dep_X5	X cosine value of the grid of 5th trip's departure location
	Dep_Y5	Y cosine value of the grid of 5th trip's departure location
	Arr_X5	X cosine value of the grid of 5th trip's arrival location
	Arr_Y5	Y cosine value of the grid of 5th trip's arrival location
	Purpose5_1	If the trip purpose of 5th trip is Work
	Purpose5_2	If the trip purpose of 5th trip is Home
	Purpose5_3	If the trip purpose of 5th trip is S&E
	Purpose5_4	If the trip purpose of 5th trip is Medical
Dependent Variables		
Socio-Economic Attributes (3)	Age	Age of the individual
	Job Status	Job status of the individual
	Income	Income of the individual

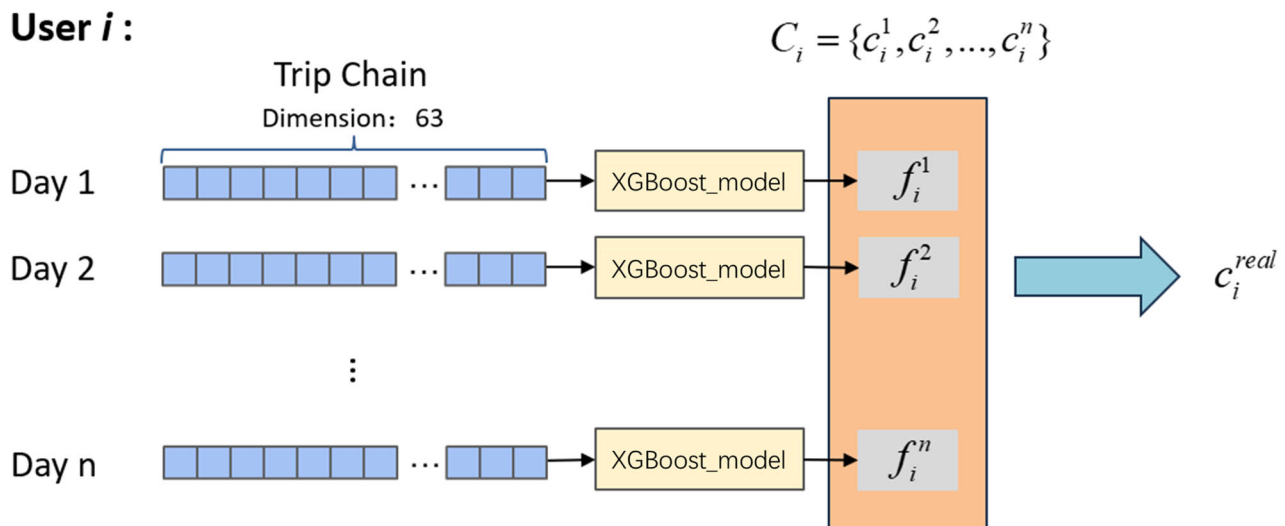


Figure 4. Socio-economic attribute inference based on multi-day travel data.

3.6. Evaluation Metrics

During the training process, all train–test splits were performed at the user level rather than the trip level. Specifically, users in the RTSD were randomly divided into 80% training users and 20% testing users, and all trips belonging to the same user were kept in the same subset. This user-level separation was adopted to avoid information leakage caused by placing different trips of the same individual into both the training and testing sets. The same user-level splitting principle was also applied in the held-out RTSD pseudo-label simulation experiment. The selection of XGBoost hyperparameters was determined by grid search on the training users only. For initial training and the teacher–student switching process, we designed different evaluation metrics to evaluate the performance of the models. In the process of initial training, accuracy, precision, and recall are considered. The metrics are calculated by Equations (14)–(16):

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \quad (14)$$

$$Precision = \frac{tp}{tp + fp} \quad (15)$$

$$Recall = \frac{tp}{tp + fn} \quad (16)$$

where tp is the number of true positive, fp is the number of false positive, fn is the number of false negative, and tn is the number of true negative. The overall metrics is calculated by Equation (17):

$$Weighted\ metric = \sum_{i=1}^N \left(\frac{metric_i * n_i}{\sum_{i=1}^N n_i} \right) \quad (17)$$

where N represents the number of classes, n_i represents the sample numbers of class i , and $metric_i$ represents the accuracy, precision, or recall of class i .

Because the SCD does not contain observed trip-purpose or socio-economic attribute labels, the validation of the teacher–student self-training strategy cannot rely on conventional classification metrics directly computed on the SCD. To avoid over-interpreting unsupervised clustering indicators as classification accuracy, we designed a three-part validation strategy. First, we conducted a source-domain pseudo-label simulation experiment

on the RTSD, where ground-truth labels are available. Specifically, the RTSD was split at the user level into a labeled subset and a held-out subset. The labels of the held-out subset were temporarily masked, and this subset was treated as unlabeled data during the self-training process. The initial teacher model (ITM), trained only on the labeled subset, first generated pseudo-labels for the masked subset. The teacher–student switching model (TSSM) was then trained using the labeled subset and the selected pseudo-labeled samples under the same curriculum pseudo-labeling procedure as that applied to the SCD. After training, the hidden labels of the held-out subset were recovered only for evaluation, and accuracy, macro-F1, weighted-F1, and class-wise recall were calculated. This experiment provides a controlled validation of whether the self-training strategy can improve pseudo-label quality when ground-truth labels are available. Second, for the actual unlabeled SCD, the silhouette coefficient and Davies–Bouldin index were used only as indirect structural consistency indicators rather than direct measures of classification accuracy. For trip-purpose inference, the underlying assumption is that trips assigned to the same inferred purpose should share similar temporal and spatial features. For socio-economic attribute inference, the use of these indicators is motivated by social segregation theory and the travel-behavior homogeneity hypothesis, which suggest that individuals with similar socio-economic attributes tend to share comparable activity-travel, residential, and employment-location characteristics. Therefore, improvements in these two indicators can only suggest more internally coherent pseudo-label assignments in the original feature space, but they cannot by themselves prove the correctness of the inferred labels. Third, to examine the robustness of the pseudo-labeling process, we conducted a confidence-threshold sensitivity analysis. The base confidence threshold τ was varied within {0.70, 0.75, 0.80, 0.85, 0.90, 0.95}. For each threshold, we reported the pseudo-label coverage ratio, pseudo-label accuracy, and macro-F1 on the held-out RTSD simulation experiment, as well as the silhouette coefficient and Davies–Bouldin index on the SCD. This sensitivity analysis helps assess the trade-off between pseudo-label reliability and the number of unlabeled samples incorporated into training. The final threshold was selected based on the joint consideration of pseudo-label quality, sample coverage, and model stability. The calculations of the two metrics are illustrated by Equations (18) and (19):

$$S(i) = \frac{1}{n} \sum_{i=1}^n \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (18)$$

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma(i) + \sigma(j)}{d(i, j)} \right) \quad (19)$$

where a represents the average distance from other samples in the same cluster, b represents the average distance from the cluster where the nearest sample is located, n represents the number of samples, $\sigma(i)$ represents the average distance from samples within cluster i to the centroid of the cluster, $d(i, j)$ represents the distance between the centroids of cluster i and cluster j , and k represents the number of samples. The range of values for the silhouette coefficient is between -1 and 1 , with the values closer to 1 indicating better models. A smaller Davies–Bouldin index indicates that the model’s classification effect is better.

4. Results

4.1. Results for Inferring Trip Purposes

The confusion matrices of the two machine-learning models are shown in Figure 5. Figure 5a,b show the confusion matrices of Model I and Model II, respectively. The average accuracies of Model I and Model II are 86.7% and 78.3%. In Model I, the purpose of Work and Home are well recognized, achieving an accuracy of around 90%, while the accuracies

of the S&E and Medical purpose are relatively low, at only 57.0% and 29.4%. The average accuracy of Model I is 78.3%. In Model II, the purpose of Medical is also not very well predicted, with an accuracy of only 47.3%. It seems that both Model I and Model II tend to classify the purpose of Medical as a purpose of S&E. Thus, the low accuracies for S&E and Medical may be due to the fact that the two purposes are usually dispersed in time and space with no clear boundaries, which is not conducive to recognition by the model. Although the recognition accuracy of the Medical category is relatively low, we did not merge it with S&E in the main experiment for two reasons. First, medical-related travel has a distinct behavioral and policy meaning from shopping and entertainment travel, especially for identifying special mobility needs of elderly or health-related transit users. Merging Medical into S&E would improve numerical balance but would also remove this semantic distinction. Second, Medical accounts for only 0.5% of the RTSD trips, and therefore its influence on the downstream socio-economic attribute inference is limited at the aggregate level. In the socio-economic inference model, trip purposes are used together with temporal, spatial, job–housing, population-density, and land-price features. As shown in the SHAP analysis, the most important predictors are mainly the temporal features of the first two trips, job–housing spatial variables, and major trip-purpose indicators such as Home and S&E, while Medical-purpose variables do not dominate the prediction of age, job status, or income. Therefore, the low accuracy of the Medical category should be interpreted as a limitation of fine-grained, rare-purpose inference, but it does not substantially undermine the overall socio-economic profiling framework. Considering that rule-based models are generally accurate for the purpose of returning home and going to work, the accuracy of identifying the purpose of home and work is over 95%, and the overall accuracy of our proposed trip-purpose inference model in this study is 92.7%.

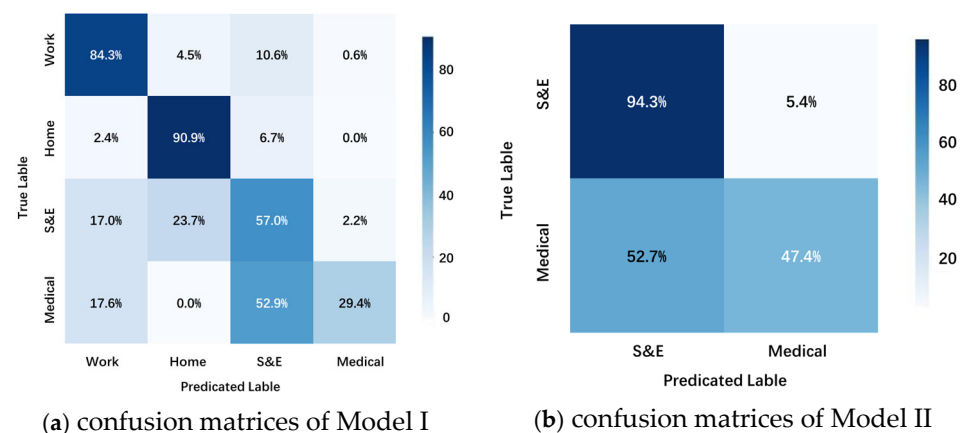


Figure 5. Confusion matrices of the two models on the testing set.

We give further attention to the metrics between our proposed model and other machine learning models. The performance of the three models is shown in Table 4. As can be seen from the table, apart from our proposed model, XGBoost outperforms the other machine-learning models, which provides us with the reason for choosing it as the base model in Model I and Model II. Moreover, our proposed model outperforms other models in all evaluation metrics with a 14.4% increased accuracy. This result demonstrates the effectiveness of identifying users' homes and workplaces first. By pre-identifying users' home and workplace locations, our proposed model can effectively distinguish primary activities, such as returning home and going to work, from other secondary activities.

Table 4. Performance comparison between proposed model and other models (unit: %).

Model	Accuracy	Precision	Recall
Proposed model	92.7	91.2	92.7
XGBoost	78.3	77.4	78.3
SVM	68.4	65.6	68.4
KNN	71.2	69.7	71.2
RF	72.8	67.2	72.8
LightGBM	74.9	73.3	74.9

Note: The “proposed model” in the table refers to the metrics of the joint rule-based model and machine learning model.

Before applying the self-training strategy to the actual unlabeled SCD, we further conducted a source-domain pseudo-label simulation experiment using the RTSD, where ground-truth trip-purpose labels are available. It is important to emphasize that the held-out subset was completely unseen during the ITM training phase; the user-level split ensured that no trip records from these users were included in the labeled training set. This setting allows the simulation experiment to directly evaluate the generalization capability of the self-training strategy to new users. Specifically, the RTSD was split at the user level into a labeled subset and a held-out subset. The labels of the held-out subset were masked during self-training and were recovered only for post hoc evaluation. This setting allowed us to directly compare pseudo-labels generated by the ITM and the TSSM with the ground-truth labels. The results show that the TSSM consistently improved pseudo-label quality for trip-purpose inference. For Model I, pseudo-label accuracy increased from 86.7% to 89.1%, macro-F1 increased from 73.6% to 77.8%, and weighted-F1 increased from 86.2% to 88.7%. For Model II, pseudo-label accuracy increased from 78.3% to 81.2%, macro-F1 increased from 59.6% to 64.9%, and weighted-F1 increased from 76.9% to 80.3%. These results provide direct evidence, under a controlled setting with available ground-truth labels, that the teacher–student self-training strategy improves the quality of pseudo-labels rather than merely changing the clustering structure of the unlabeled data.

To quantify the contribution of each component in the proposed teacher–student framework, we conducted an ablation analysis using the held-out RTSD pseudo-label simulation experiment. Four variants were compared: (1) ITM, which uses only the supervised model trained on labeled RTSD; (2) ITM + ST, which adds standard self-training with a fixed confidence threshold; (3) ITM + ST + CPL, which further incorporates curriculum pseudo-labeling to adaptively adjust class-specific confidence thresholds; and (4) the full TSSM, which combines self-training, curriculum pseudo-labeling, and teacher–student switching. The results show a gradual improvement as more components are introduced. For trip-purpose inference, the average pseudo-label accuracy of Model I and Model II increased from 82.5% under ITM to 83.8% after adding standard self-training, then to 84.6% after incorporating curriculum pseudo-labeling, and finally to 85.2% under the full TSSM. For socio-economic attribute inference, the average accuracy of age, job status, and income increased from 76.3% under ITM to 77.3%, 77.9%, and 78.5%, respectively. These results indicate that the performance improvement does not come from a single component alone. Standard self-training provides the basic benefit of incorporating unlabeled samples, curriculum pseudo-labeling improves the reliability of pseudo-label selection under class imbalance and different learning difficulties, and teacher–student switching further stabilizes the iterative transfer process.

In the SCD, the home and workplace of users are recognized first. The success rate of recognizing both home and workplace is 14.5%. It should be noted that this relatively low identification rate implies that the rule-based model with high accuracy is only applicable to a subset of users with regular commuting patterns. Consequently, the reported overall

accuracy of 92.7% may partially reflect the performance on this more predictable subgroup. However, since the XGBoost-based model covers all remaining trips and the self-training strategy is applied to the entire dataset, the framework is not limited to this 14.5% of users. The potential selection bias introduced by the rule-based identification is acknowledged as a limitation and further discussed in Section 6. Both Model I and Model II are then trained through the self-training strategy. The results are shown as Table 5. For the actual unlabeled SCD, Table 5 reports the silhouette coefficient and Davies–Bouldin index of ITM and TSSM. The TSSM achieves higher silhouette coefficients and lower Davies–Bouldin indexes than the ITM for both Model I and Model II, suggesting that the self-training strategy produces more structurally coherent pseudo-label assignments in the SCD feature space. However, because the SCD does not contain ground-truth trip-purpose labels, these clustering metrics are interpreted only as auxiliary evidence of improved label consistency, rather than as direct proof of classification accuracy. We further conducted a sensitivity analysis on the base confidence threshold τ . As shown in Table 6, increasing τ generally improved pseudo-label reliability on the held-out RTSD subset but reduced the pseudo-label coverage ratio. When τ was too low, more unlabeled samples were incorporated into training, but the risk of noisy pseudo-labels increased. When τ was too high, pseudo-label accuracy improved only marginally, while the number of usable pseudo-labeled samples decreased substantially.

Next, we predicted the purpose of each trip to determine whether the distributions of purposes recorded in those datasets are similar. Given the RTSD are all on weekdays, the SCD is divided into a weekday dataset and a weekend dataset to look into the differences. The results are shown in Figure 6. The distribution of trip purposes in SCD-weekday exhibits greater similarity to that of the RTSD than that in SCD-weekend. The proportion of S&E purposes on weekends is higher than proportions on the other two datasets, while the proportion of Work purpose is lower. This is because people tend to have more entertainment on non-working days. The proportion of S&E in SCD-weekday is also much higher compared to the RTSD. This may be due to the fact that the respondents in the RTSD are mostly high-frequency users, while SCD-weekday covers a large number of low-frequency users whose travel purposes are mostly non-commuting. As a whole, although the distribution of travel purposes has subtle differences between the two datasets, the overall distribution remained consistent.

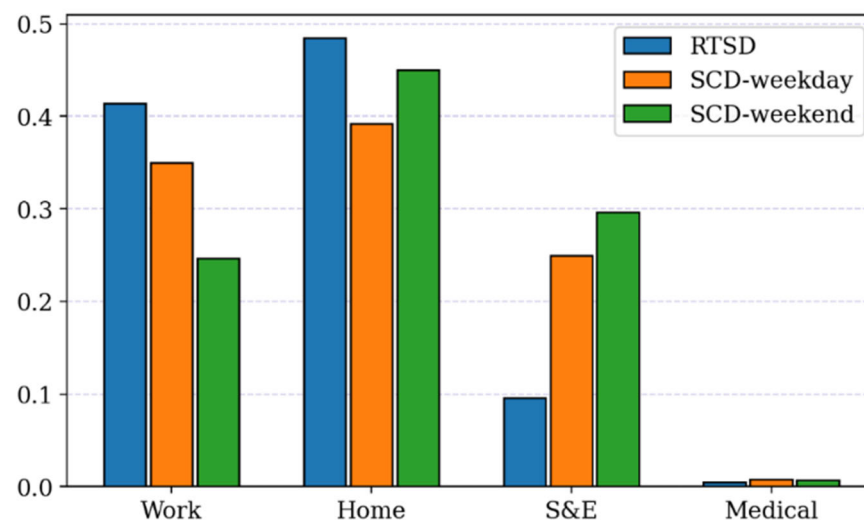
To examine possible error propagation from the first-stage trip-purpose inference to the second-stage socio-economic attribute inference, we conducted an additional source-domain comparison on the RTSD. Two versions of the socio-economic attribute inference model were trained and evaluated: one using ground-truth trip purposes from the RTSD, and the other using inferred trip purposes generated by the first-stage model. The comparison shows that replacing ground-truth trip purposes with inferred trip purposes led to only moderate performance degradation. Specifically, the accuracy decreased from 86.1% to 85.5% for age inference, from 84.4% to 83.8% for job-status inference, and from 60.5% to 59.7% for income inference. This suggests that the proposed framework is relatively robust to first-stage trip-purpose errors, mainly because socio-economic inference also uses temporal, spatial, job-housing, population density, and land-price features rather than relying solely on trip-purpose variables. Nevertheless, error propagation remains possible, especially for rare or ambiguous trip purposes.

Table 5. Comparison between trip purpose ITM and TSSM on SCD.

	Model	Silhouette_Coef	DB_Index
Model I	ITM	0.156	2.938
	TSSM	0.173	1.672
Model II	ITM	0.049	18.186
	TSSM	0.054	7.839

Table 6. Confidence-threshold sensitivity analysis for trip-purpose inference.

T	Coverage (%)	Accuracy on Held-Out RTSD (%)	Macro-F1 (%)	Silhouette Coef on SCD	DB Index on SCD
0.70	87.8	82.9	68.9	0.098	7.254
0.75	81.6	84.4	71.6	0.105	5.690
0.80	73.5	85.7	73.4	0.111	4.995
0.85	64.2	86.4	74.1	0.114	4.756
0.90	51.9	86.8	74.8	0.111	5.085
0.95	35.6	87.1	75.0	0.106	6.150

**Figure 6.** Trip-purpose distribution on different datasets.

4.2. Results for Inferring Socio-Economic Attributes

The confusion matrices of initial models of socio-economic attributes inference are shown in Figure 7. Figure 7a–c show the confusion matrices of age, job status, and income models, respectively. Among the three models, the job status model performs the best, with all three category accuracies higher than 70%. For the age model, people aged under 20 and 20–59 are well predicted, with accuracies at 71.6% and 89.3%, respectively. However, people aged above 60 are not clearly identified, which may be due to the small sample size and behavioral differences within the elderly categories. For the income model, people who do not earn money and those who earned 0–100 thousand yuan per month are fairly well recognized, with accuracies at the levels of 62.1% and 66.5%, respectively. The other two category accuracies are both lower than 50% and the average accuracy of income is merely 59.5%, indicating that the relationship between income and behavior of using public transit is relatively weak. Note that the relatively lower performance of income inference can be explained from both behavioral and data perspectives. First, compared with age and job status, income has a more indirect relationship with daily bus travel behavior. Age and job status are often reflected in regular schedules, commuting intensity, and the

presence or absence of workplace-related trips, whereas income may affect travel behavior through more complex channels such as residential choice, car ownership, job location, and multimodal substitution. Second, regular bus users are overrepresented by low- and middle-income groups, which reduces income variance within the bus-user sample and makes fine-grained income classification more difficult. Third, the four-category income setting is relatively granular for smart card-based inference. The income brackets used in this study, namely 0, 0–100,000 RMB, 100,000–150,000 RMB, and $\geq 150,000$ RMB, follow the original income categories in the RTSD survey and were retained to preserve consistency with the survey design. Although income inference is less accurate than age and job-status inference, it does not invalidate the overall application value of the proposed framework. The proposed model achieves much higher accuracy for age and job status, which are key attributes for transit user profiling. In addition, income is mainly used for aggregate-level passenger profiling and route-level analysis rather than individual-level decision-making. Therefore, income predictions should be interpreted cautiously and used as supplemental information. A binary or coarser income classification, such as low-income versus middle-to-high-income users, may be more robust and is worth further investigation).

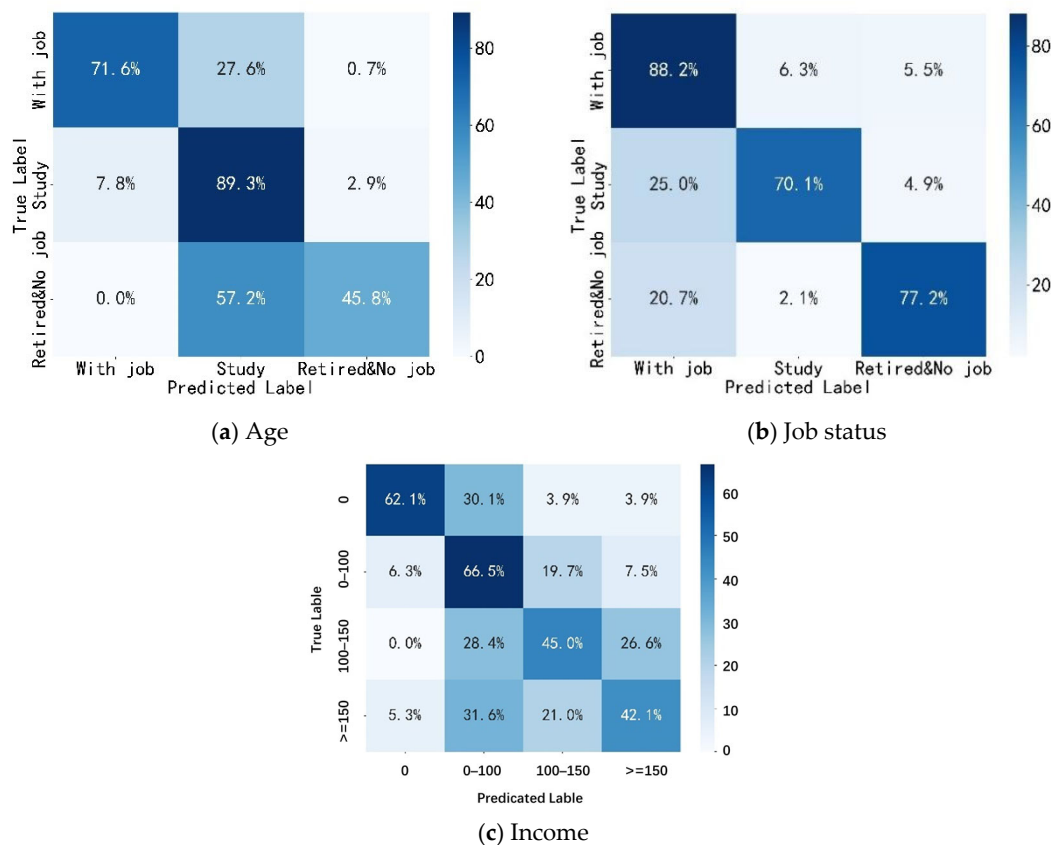


Figure 7. Confusion matrices of socio-economic inference models for the testing set.

The performances of the initial teacher XGBoost models are compared with other models, including Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Random Forest (RF), and LightGBM, across various attributes. Table 7 presents the results, which indicate that XGBoost consistently outperforms the other models in most cases, achieving overall accuracies of 85.5% and 83.8% for predicting age and job status, respectively, though its performance in inferring income was slightly lower at 59.5%. Despite lower accuracy and recall for income prediction compared to RF and LightGBM, XGBoost exhibited notably higher precision. The model achieved an average accuracy of 76.3% across all three attributes,

demonstrating strong performance in inferring socio-economic characteristics from public transit trip data alone.

Table 7. Performance comparison of socio-economic inference models (unit: %).

Attributes	Model	Accuracy	Precision	Recall
Age	SVM	81.4	76.4	81.4
	KNN	84.3	81.8	84.3
	XGBoost	85.5	88.2	85.5
	RF	85.3	85.0	85.3
	LightGBM	85.2	84.3	85.2
Job Status	SVM	79.1	76.2	79.1
	KNN	81.7	80.5	81.7
	XGBoost	83.8	85.3	83.8
	RF	83.5	84.1	83.5
	LightGBM	83.4	83.1	83.4
Income	SVM	59.5	51.1	59.5
	KNN	56.5	50.7	56.5
	XGBoost	59.7	71.8	59.7
	RF	63.1	56.5	63.1
	LightGBM	60.1	55.5	60.1

In addition to the conventional machine-learning baselines, we further compared the proposed TSSM with two semi-supervised pseudo-labeling baselines in the held-out RTSD pseudo-label simulation experiment. The first baseline used standard self-training with a fixed confidence threshold, in which high-confidence predictions generated by the ITM were directly selected as pseudo-labels for the next round of training. The second baseline used curriculum pseudo-labeling to adaptively adjust class-specific confidence thresholds, but did not include the teacher–student switching mechanism. The results show that the supervised ITM achieved an average trip-purpose pseudo-label accuracy of 82.5% and an average socio-economic attribute accuracy of 76.3%. The fixed-threshold pseudo-labeling baseline improved these values to 83.8% and 77.3%, respectively, while the curriculum pseudo-labeling baseline further improved them to 84.6% and 77.9%. The proposed full TSSM achieved the best performance, with 85.2% for trip-purpose inference and 78.5% for socio-economic attribute inference. These results suggest that semi-supervised pseudo-label learning is beneficial for model transfer, and that the proposed teacher–student switching mechanism provides additional improvement beyond fixed-threshold or curriculum-only pseudo-labeling.

After verifying the effectiveness of the initial model, the performance of TSSM is also investigated. The model is tested on the RTSD and the results are shown in Table 8. For all three types of socio-economic attribute inference models, their self-trained models achieved higher silhouette coefficients and lower Davies–Boulding indexes. The results indicate that the self-training strategy exhibits good transferability and effectively improves the initial model’s performance when applied to the SCD.

The comparison of socio-economic attribute distributions on the two datasets is shown in Figure 8. Although the distribution of socio-economic attributes is roughly the same, there are still some subtle differences. The age distribution is generally consistent between the two datasets, with slightly fewer individuals under 20 years old in the SCD and a somewhat higher proportion of those over 60. The number of retired individuals is similar across both datasets, but there is a lower proportion of employed people in the SCD, while students are overestimated in the SCD. Concerning income, there is a significantly larger proportion of individuals with incomes over 150,000, while those earning less

than 150,000 are underestimated. This could suggest that the SCD includes more low-frequency public transit users, who tend to have relatively higher incomes and use public transportation only occasionally.

Table 8. Comparison between socio-economic ITM and TSSM on SCD.

Attributes	Model	Silhouette_Coef	DB_Index
Age	ITM	−0.079	11.514
	TSSM	−0.001	7.644
Job status	ITM	−0.039	9.972
	TSSM	−0.012	8.124
Income	ITM	−0.020	20.074
	TSSM	−0.007	5.922

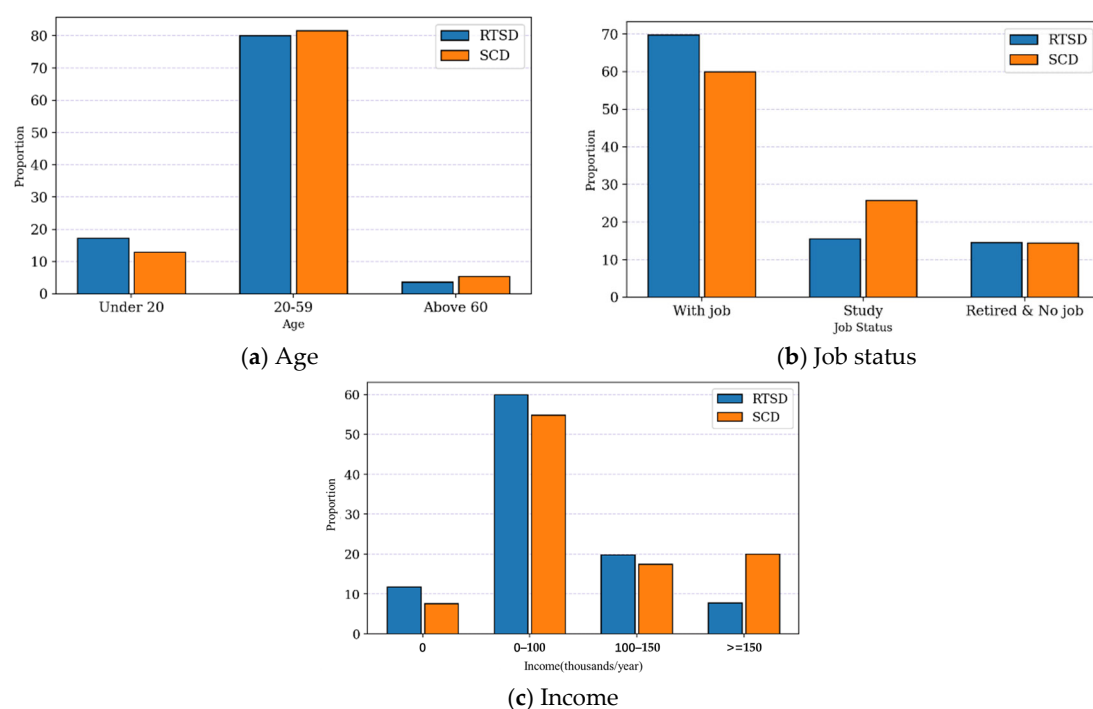


Figure 8. Socio-economic attribute distribution on different datasets.

5. Discussion

To investigate the contributions of spatiotemporal, job–housing, and trip-purpose features to socio-economic attributes, Shapley Additive Explanations (SHAP) [53] is employed to evaluate the average contribution of each independent variable. The larger the absolute SHAP value of an independent variable, the greater its impact is on the model’s dependent variable. The SHAP values of the three models are shown in Figure 8; the spatiotemporal, job–housing, and trip-purpose features marked as different colors and their actual mean absolute SHAP value are noted under the variable name.

For the age inference model, as shown in Figure 9a, the arrival and departure times of the first two trips in one day make the largest contribution to the model predictions for users aged under 20 and users aged between 20 and 59. In addition, whether the first inferred trip purpose is Home or S&E also contributes to the prediction of users aged under 20. This pattern may reflect differences in daily activity schedules among younger transit users. Specifically, when the first observed bus trip is inferred as a Home-purpose trip, it suggests that the user may have completed an earlier activity or trip using another travel

mode before the first observed bus ride. In the SCD, 9.2% of users have a first observed trip inferred as Home, and most of these trips occur in the afternoon. A plausible explanation is that some students may not use regular bus services for their morning trip but may take the bus when returning home in the afternoon. Similarly, some college students under 20 may have more flexible daily schedules, making S&E-purpose first trips more likely to appear in their observed trip chains. For users aged above 60, job-housing-related variables have relatively larger SHAP contributions. In particular, missing or weak workplace-related information is associated with a higher predicted probability of being classified into the older-age group, which is consistent with the fact that retired users may not have regular workplace-related travel patterns. The spatial clustering pattern of inferred residential locations also provides additional predictive information for this age group.

For the job-status model, as shown in Figure 9b, the SHAP results show a pattern similar to that of the age model. The prediction of employed users and students is mainly associated with temporal features of the first two trips, while the prediction of retired and non-employed users is more closely related to workplace-related variables and whether the first inferred trip purpose is S&E. Compared with the age model, trip-purpose features appear more frequently among the top-ranked variables in the job-status model. This suggests that inferred activity purposes provide useful predictive information for distinguishing employment-related daily schedules, especially between regular commuting patterns and more flexible non-work activity patterns.

For the income model, as shown in Figure 9c, job-housing-related spatial variables, especially land price and residential-location features, have relatively larger SHAP contributions than in the age and job-status models. This result suggests that residential and workplace spatial contexts are informative for income prediction in the fitted model. For users with non-zero income, the land price around the inferred home location is associated with the model's income classification. For users with monthly income above 10,000 yuan, the Y-coordinate of the home grid also appears among the top-ranked variables, suggesting that the relative residential location within the city contributes to the model prediction. However, these findings should not be interpreted as causal effects of land price or residential location on income. Rather, they indicate that the model captures statistical associations between urban spatial context and inferred income categories. Overall, the SHAP analysis provides interpretable evidence on which behavioral and spatial features the model relies on when predicting socio-economic attributes, but causal interpretation would require additional research designs and external validation.

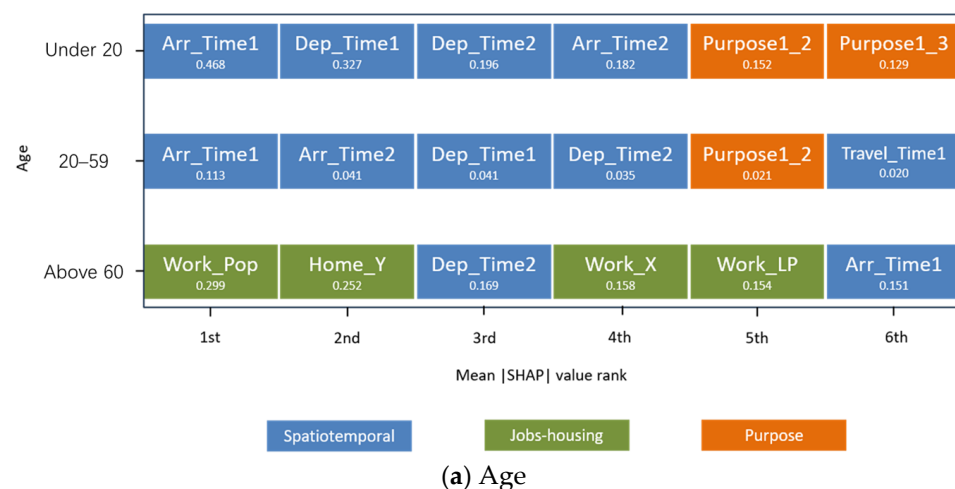


Figure 9. Cont.

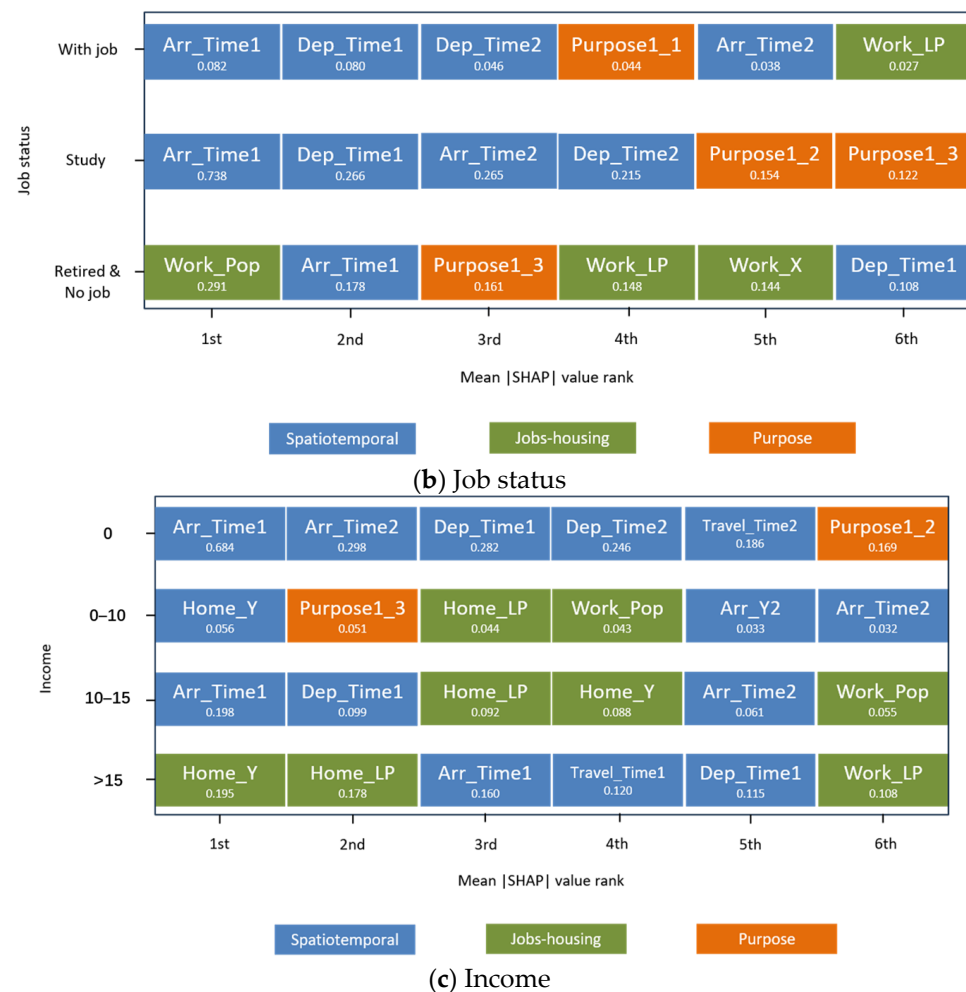


Figure 9. SHAP value of the socio-economic inference model.

6. Conclusions

The primary objective of this study is to develop a two-stage method for inferring both trip purposes and socio-economic attributes. A combined rule-based and machine-learning model is established to infer trip purposes and a machine-learning model is built to infer socio-economic attributes such as age, job status, and income. A self-training teacher–student switching model is also established to transfer the model from resident travel survey data to smart card data. The interpretability of the socio-economic model is also investigated. From the results, some major conclusions can be drawn. First, models for inferring trip purposes have an overall accuracy of 92.7%. The rule-based model can improve recognition performance for purposes of Work and Home, achieving an accuracy of over 95%. The proportion of trips for shopping and entertainment purposes in smart card data is higher than that in resident travel survey data. Second, the models for inferring socio-economic attributes achieve an average accuracy of 76.3% across age, job status, and income, with individual accuracies of 85.5%, 83.8%, and 59.7%, respectively. The self-training framework can enhance the model's transferability across different datasets. Finally, travel time, arrival time, and departure time of the first two trips in a day of an individual play the most important roles in recognizing all three socio-economic models. The results indicate that the first trip purpose being Home is an important feature for identifying students and individuals under 20, whereas jobs–housing location land prices contribute significantly to the inference of individuals' incomes.

Despite the contributions of this study, several limitations must be acknowledged. First, due to the lack of subway smart card data, only regular bus trips are considered. This may neglect the transfer behavior between subway and bus, and affect the accuracy in describing travel behaviors. Second, this method only experimented on data from Shenzhen, which may limit the generalizability of the model to other urban environments. To demonstrate the generality of the model, experiments can be conducted on datasets from additional cities in the future. Another limitation concerns the validation of inferred labels in the actual SCD. Furthermore, our comparison with semi-supervised learning methods is limited to variants of our own framework (i.e., fixed-threshold self-training and curriculum pseudo-labeling). We acknowledge that comparisons with more recent state-of-the-art semi-supervised or domain adaptation methods are not included, which would provide a broader benchmark for the proposed teacher–student model. Although we added a held-out RTSD pseudo-label simulation experiment and confidence-threshold sensitivity analyses to directly examine the effectiveness of the self-training strategy under a setting with available ground-truth labels, the actual SCD still lacks observed trip-purpose and socio-economic attribute labels. Therefore, the evaluation on the SCD remains indirect and relies on structural consistency indicators and distributional plausibility checks. Future research should procure a small, manually validated smart card sample or link smart card records with privacy-preserving surveys or census information to further validate the inferred labels at the individual or aggregate level. In addition, future work should validate the proposed framework on transit smart card data from other cities with diverse urban forms and transit systems, to further demonstrate its generalizability. Such cross-city validations would help assess the robustness of the two-stage inference approach under different spatial and demographic contexts.

Author Contributions: Conceptualization, C.Y. (Chao Yang) and Y.T.; methodology, W.D.; software, C.Y. (Chengcheng Yu); validation, C.Y. (Chao Yang) and Y.T.; formal analysis, Y.Z.; investigation, W.D.; resources, C.Y. (Chao Yang) and Y.T.; data curation, C.Y. (Chao Yang) and Y.T.; writing—original draft preparation, C.Y. (Chao Yang); writing—review and editing, Y.T.; visualization, Y.Z.; supervision, Y.T.; project administration, C.Y. (Chao Yang); and funding acquisition, C.Y. (Chao Yang). All authors have read and agreed to the published version of the manuscript.

Funding: The authors appreciate the funding support from NSFC grants (Grant No. 52572346).

Institutional Review Board Statement: Ethical review and approval were waived for this study by the Institutional Committee as per Article 32 of the “Measures for Ethical Review of Life Sciences and Medical Research Involving Humans” (National Health Commission Document No. 4).

Informed Consent Statement: Informed consent for participation was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, C.; Yu, C.; Dong, W.; Yuan, Q. Substitutes or complements? Examining effects of urban rail transit on bus ridership using longitudinal city-level data. *Transp. Res. Part A Policy Pract.* **2023**, *174*, 103728. [[CrossRef](#)]
2. Yu, C.; Dong, W.; Liu, Y.; Yang, C.; Yuan, Q. Rethinking bus ridership dynamics: Examining nonlinear effects of determinants on bus ridership changes using city-level panel data from 2010 to 2019. *Transp. Policy* **2024**, *151*, 85–100. [[CrossRef](#)]
3. Fan, Q.; Yu, C.; Zuo, J. The Nonlinear Causal Effect Estimation of the Built Environment on Urban Rail Transit Station Flow Under Emergency. *Sustainability* **2025**, *17*, 5829. [[CrossRef](#)]
4. Barbosa, H.; Barthélemy, M.; Ghoshal, G.; James, C.R.; Lenormand, M.; Louail, T.; Menezes, R.; Ramasco, J.J.; Simini, F.; Tomasini, M. Human mobility: Models and applications. *Phys. Rep.* **2018**, *734*, 1–74. [[CrossRef](#)]

5. Lenormand, M.; Louail, T.; Cantú-Ros, O.G.; Picornell, M.; Herranz, R.; Arias, J.M.; Barthelemy, M.; Miguel, M.S.; Ramasco, J.J. Influence of sociodemographic characteristics on human mobility. *Sci. Rep.* **2015**, *5*, 10075. [[CrossRef](#)] [[PubMed](#)]
6. Wong, R.C.P.; Szeto, W.Y.; Yang, L.; Li, Y.C.; Wong, S.C. Public transport policy measures for improving elderly mobility. *Transp. Policy* **2018**, *63*, 73–79. [[CrossRef](#)]
7. Shin, E.J. Exploring the causal impact of transit fare exemptions on older adults' travel behavior: Evidence from the Seoul metropolitan area. *Transp. Res. Part A Policy Pract.* **2021**, *149*, 319–338. [[CrossRef](#)]
8. Simićević, J.; Milosavljević, N.; Djoric, V. Gender differences in travel behaviour and willingness to adopt sustainable behaviour. *Transp. Plan. Technol.* **2016**, *39*, 527–537. [[CrossRef](#)]
9. Di Ciommo, F.; Comendador, J.; López-Lambas, M.E.; Cherchi, E.; Ortúzar, J.d.D. Exploring the role of social capital influence variables on travel behaviour. *Transp. Res. Part A Policy Pract.* **2014**, *68*, 46–55. [[CrossRef](#)]
10. Silm, S.; Tominga, A.; Saidla, K.; Poom, A.; Tammara, T. Socio-economic and residential differences in urban modality styles based on a long-term smartphone experiment. *J. Transp. Geogr.* **2024**, *115*, 103810. [[CrossRef](#)]
11. He, S.Y.; Thøgersen, J.; Cheung, Y.H.Y.; Yu, A.H.Y. Ageing in a transit-oriented city: Satisfaction with transport, social inclusion and wellbeing. *Transp. Policy* **2020**, *97*, 85–94. [[CrossRef](#)]
12. Rasouli, S.; Timmermans, H. Activity-based models of travel demand: Promises, progress and prospects. *Int. J. Urban Sci.* **2014**, *18*, 31–60. [[CrossRef](#)]
13. Allahviranloo, M.; Regue, R.; Recker, W. Modeling the activity profiles of a population. *Transp. B Transp. Dyn.* **2017**, *5*, 426–449. [[CrossRef](#)]
14. Yin, L.; Lin, N.; Zhao, Z. Mining Daily Activity Chains from Large-Scale Mobile Phone Location Data. *Cities* **2021**, *109*, 103013. [[CrossRef](#)] [[PubMed](#)]
15. Alencar, L.A.; Alvares, L.O.; Renso, C.; Raffaeta, A.; Bogorny, V. A Rule-based Method for Discovering Trajectory Profiles. In Proceedings of the 27th International Conference on Software Engineering and Knowledge Engineering, Pittsburgh, PA, USA, 6–8 July 2015; pp. 244–249.
16. Bi, B.; Shokouhi, M.; Kosinski, M.; Graepel, T. Inferring the demographics of search users: Social data meets search queries. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 131–140.
17. Tkaličič, M.; Ferwerda, B. Eudaimonic Modeling of Moviegoers. In Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, New York, NY, USA, 8–11 July 2018; pp. 163–167.
18. Kim, E.-J.; Kim, D.-K.; Sohn, K. Imputing qualitative attributes for trip chains extracted from smart card data using a conditional generative adversarial network. *Transp. Res. Part C Emerg. Technol.* **2022**, *137*, 103616. [[CrossRef](#)]
19. Cui, Y.; Meng, C.; He, Q.; Gao, J. Forecasting current and next trip purpose with social media data and Google Places. *Transp. Res. Part C Emerg. Technol.* **2018**, *97*, 159–174. [[CrossRef](#)]
20. Xiao, G.; Juan, Z.; Zhang, C. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transp. Res. Part C Emerg. Technol.* **2016**, *71*, 447–463. [[CrossRef](#)]
21. Wang, P.; Liu, G.; Fu, Y.; Zhou, Y.; Li, J. Spotting Trip Purposes from Taxi Trajectories: A General Probabilistic Model. *ACM Trans. Intell. Syst. Technol.* **2017**, *9*, 1–26. [[CrossRef](#)]
22. Gong, L.; Liu, X.; Wu, L.; Liu, Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2015**, *43*, 103–114. [[CrossRef](#)]
23. Kusakabe, T.; Asakura, Y. Behavioural data mining of transit smart card data: A data fusion approach. *Transp. Res. Part C Emerg. Technol.* **2014**, *46*, 179–191. [[CrossRef](#)]
24. Alsger, A.; Tavassoli, A.; Mesbah, M.; Ferreira, L.; Hickman, M. Public transport trip purpose inference using smart card fare data. *Transp. Res. Part C Emerg. Technol.* **2018**, *87*, 123–137. [[CrossRef](#)]
25. Chakirov, A.; Erath, A. Activity identification and primary location modelling based on smart card payment data for public transport. In Proceedings of the 13th International Conference on Travel Behaviour Research (IATBR 2012), Toronto, ON, Canada, 15–20 July 2012. [[CrossRef](#)]
26. Devillaine, F.; Munizaga, M.; Trépanier, M. Detection of Activities of Public Transport Users by Analyzing Smart Card Data. *Transp. Res. Rec. J. Transp. Res. Board* **2012**, *2276*, 48–55. [[CrossRef](#)]
27. Stopher, P.; FitzGerald, C.; Zhang, J. Search for a global positioning system device to measure person travel. *Transp. Res. Part C Emerg. Technol.* **2008**, *16*, 350–369. [[CrossRef](#)]
28. Chen, C.; Gong, H.; Lawson, C.; Bialostozky, E. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transp. Res. Part A Policy Pract.* **2010**, *44*, 830–840. [[CrossRef](#)]
29. Sari Aslam, N.; Ibrahim, M.R.; Cheng, T.; Chen, H.; Zhang, Y. ActivityNET: Neural networks to predict public transport trip purposes from individual smart card data and POIs. *Geo-Spat. Inf. Sci.* **2021**, *24*, 711–721. [[CrossRef](#)]
30. Han, Y.; Li, W.; Wei, S.; Zhang, T. Research on Passenger's Travel Mode Choice Behavior Waiting at Bus Station Based on SEM-Logit Integration Model. *Sustainability* **2018**, *10*, 1996. [[CrossRef](#)]

31. Ma, X.; Wu, Y.-J.; Wang, Y.; Chen, F.; Liu, J. Mining smart card data for transit riders' travel patterns. *Transp. Res. Part C Emerg. Technol.* **2013**, *36*, 1–12. [\[CrossRef\]](#)
32. Kim, K.; Oh, K.; Lee, Y.K.; Kim, S.; Jung, J.-Y. An analysis on movement patterns between zones using smart card data in subway networks. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1781–1801. [\[CrossRef\]](#)
33. Cho, W.; Chung, J.-H.; Kim, J. Need-based approach for modeling multiday activity participation patterns and identifying the impact of activity/travel conditions. *Transp. Res. Part A Policy Pract.* **2023**, *172*, 103665. [\[CrossRef\]](#)
34. Yu, C.; Lin, H.; Dong, W.; Fang, S.; Yuan, Q.; Yang, C. TripChain2RecDeepSurv: A novel framework to predict transit users' lifecycle behavior status transitions for user management. *Transp. Res. Part C Emerg. Technol.* **2024**, *167*, 104818. [\[CrossRef\]](#)
35. Pas, E.I. A Flexible and Integrated Methodology for Analytical Classification of Daily Travel-Activity Behavior. *Transp. Sci.* **1983**, *17*, 405–429. [\[CrossRef\]](#)
36. Hasan, A.; Hasan, U.; Aljassmi, H.; Whyte, A. Transit Behaviour and Sociodemographic Interrelation: Enhancing Urban Public-Transport Solutions. *Eng* **2023**, *4*, 1144–1155. [\[CrossRef\]](#)
37. Solomon, A.; Bar, A.; Yanai, C.; Shapira, B.; Rokach, L. Predict Demographic Information Using Word2vec on Spatial Trajectories. In Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, Singapore, Singapore, 8–11 July 2018; pp. 331–339.
38. Wu, L.; Yang, L.; Huang, Z.; Wang, Y.; Chai, Y.; Peng, X.; Liu, Y. Inferring demographics from human trajectories and geographical context. *Comput. Environ. Urban Syst.* **2019**, *77*, 101368. [\[CrossRef\]](#)
39. Zhang, Y.; Sari Aslam, N.; Lai, J.; Cheng, T. You are how you travel: A multi-task learning framework for Geodemographic inference using transit smart card data. *Comput. Environ. Urban Syst.* **2020**, *83*, 101517. [\[CrossRef\]](#)
40. Zhao, Y.; Pawlak, J.; Sivakumar, A. Theory for socio-demographic enrichment performance using the inverse discrete choice modelling approach. *Transp. Res. Part B Methodol.* **2022**, *155*, 101–134. [\[CrossRef\]](#)
41. Zhang, B.; Rasouli, S.; Feng, T. Social demographics imputation based on similarity in multi-dimensional activity-travel pattern: A two-step approach. *Travel Behav. Soc.* **2024**, *37*, 100843. [\[CrossRef\]](#)
42. Jang, W. Travel Time and Transfer Analysis Using Transit Smart Card Data. *Transp. Res. Rec.* **2010**, *2144*, 142–149. [\[CrossRef\]](#)
43. Golledge, R.G. Learning about urban environments. In *Timing Space and Spacing Time: Making Sense of Time*; Arnold: London, UK, 1978; Volume 1, pp. 76–98.
44. Ma, X.; Liu, C.; Wen, H.; Wang, Y.; Wu, Y.-J. Understanding commuting patterns using transit smart card data. *J. Transp. Geogr.* **2017**, *58*, 135–145. [\[CrossRef\]](#)
45. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
46. Reitboeck, H.; Brody, T.P. A transformation with invariance under cyclic permutation for applications in pattern recognition. *Inf. Control* **1969**, *15*, 130–154. [\[CrossRef\]](#)
47. Gao, L.; Huang, H.; Ye, J.; Wang, D. Activity type detection of mobile phone data based on self-training: Application of the teacher–student cycling model. *Transp. Res. Part C Emerg. Technol.* **2024**, *161*, 104550. [\[CrossRef\]](#)
48. Amini, M.-R.; Feofanov, V.; Pauletto, L.; Devijver, E.; Maximov, Y. Self-Training: A Survey. *arXiv* **2022**, arXiv:2202.12040.
49. Niyogi, P. Manifold regularization and semi-supervised learning: Some theoretical analyses. *J. Mach. Learn. Res.* **2013**, *14*, 1229–1250.
50. Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; Shinozaki, T. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 6–14 December 2024; p. 1407.
51. Nilforoshan, H.; Looi, W.; Pierson, E.; Villanueva, B.; Fishman, N.; Chen, Y.; Sholar, J.; Redbird, B.; Grusky, D.; Leskovec, J. Human mobility networks reveal increased segregation in large cities. *Nature* **2023**, *624*, 586–592. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Massey, D.S.; Denton, N.A. The Dimensions of Residential Segregation. *Soc. Forces* **1988**, *67*, 281–315. [\[CrossRef\]](#)
53. Lundberg, S.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *27*, 4765–4774.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.