Electronic Research Archive

ERA, 31(11): 6844–6865.
DOI: 10.3934/era.2023346
Received: 12 September 2023
Revised: 17 October 2023
Accepted: 17 October 2023
Published: 23 October 2023

http://www.aimspress.com/journal/ERA

Research article

Interpretable machine learning models for detecting fine-grained transport modes by multi-source data

Yuhang Liu, Jun Chen, Yuchen Wang and Wei Wang*

School of Transportation, Southeast University, Nanjing 210096, China

* **Correspondence:** Email: wwang@seu.edu.cn.

Abstract: Analysis of transport mode choice is crucial in transportation planning and optimization. Traditionally, the transport mode of individuals is detected by discrete choice models (DCMs), which rely on data regarding individual and household attributes. Using these attribute data raises privacy concerns and limits the applicability of the model. Meanwhile, the detection results of DCMs may be biased, despite providing insight into the impact of variables. The machine learning models are more effective for mode detection, but most models need more interpretability. In this study, an interpretable machine learning model is developed to detect the transport modes of individuals. The mobility features of individuals, which introduce the velocity and acceleration of the center of mass (COM) are innovatively considered in the detection model. These mobility features are combined with multi-source data, including land use mix, GDP, population and online map service data as detection features. Using the travel survey data from Nanjing, China in 2015, the effects of different machine learning models on fine-grained detection performance are investigated. The results indicate that the deep forest model presents the best detection performance and achieves an accuracy of 0.82 in the test dataset, demonstrating the effectiveness of the proposed detection model. Furthermore, t-distributed stochastic neighbor embedding (t-SNE) and ablation experiments are conducted to overcome the non-interpretability issue of the machine learning models. The results show that the mobility features of individuals are the most critical features for improving detection performance. This study is essential for improving the structure of transport modes and maintaining low-carbon and sustainable development in urban traffic systems.

Keywords: transport mode choice; machine learning modeling; multi-source data analysis; model interpretability

1. Introduction

Rapid urbanization generates growing amounts of travel flows, urging the requirement for efficient transport planning policies. During 1980's, in Shanghai, China, the urbanization processes were in their nascent stages, with an approximate daily total of 20 million trips. By 2019, this daily total had surged to approximately 57.31 million trips. Similarly, in Beijing, since the inception of the first travel survey in 1986, the daily volume of trips in the central urban area has exhibited an average annual increase of 4%. By 2019, the daily volume of trips had reached a significant 39.57 million. The substantial growth has resulted in severe traffic congestion, with an average congestion duration of three hours in 2019 [1,2]. These trends underscore the urgent need for the development and implementation of efficient transport planning policies. Accurate comprehension of an individual's transport mode choice enables the development of customized guidance policies for different demographic groups. These policies can significantly contribute to the optimization of transportation system infrastructure, the promotion of increased public transportation utilization and the relief of traffic congestion. Consequently, the development of precise models for transport mode detection assumes paramount importance [3].

1.1. Background

In fact, the choice of transport modes is influenced by many types of factors, including individual, household and built environment factors. Individual factors include age, gender, income, occupation, attitudes and trip purpose [4–9]. Household factors include household structure, household characteristics, number of household members and car ownership, etc. [10–12]. Furthermore, several attributes of the built environment have been recognized to influence the choice of transport modes, such as building density, land use mix and distance to the transport facilities [13–16]. The primary method of collecting the data is via travel survey. Collecting individual mobility patterns and rich sociodemographic information are critical advantages of travel surveys [17–20]. However, respondents may provide incorrect information out of a desire to protect their privacy, such as by providing lower-income information, etc. This reason results in insufficient data collection accuracy and impacts the validity of the subsequent study [21–22]. Additionally, ensuring the security of respondents' privacy can be challenging. Balancing the detection performance of transport mode with the imperative of ensuring respondents' privacy remains a significant challenge in this field of research.

Most detection models for transport mode are predicted by econometric discrete choice models (DCMs) based on the random utility framework, such as multinomial logit model, nested logit model and mixed logit model [23–25]. The main advantage of these models is that they have good interpretability because of explicit mathematical formulas. The utility functions in the model are defined manually before fitting the model. This step allows for the integration of established behavioral theory into the model [26,27]. However, manual definitions may be violated under real complex problems, leading to biased predictions [28,29]. In contrast to traditional DCMs, machine learning models applied to model the choice of transport modes are promising methods to address current limitations [30–34]. Instead of making the necessary prior definitions, machine learning models represent complex relationships between variables through data-driven learning. Although machine learning models have better predictive accuracy, they focus on prediction and lack model interpretability [35]. Novel techniques and experiments such as t-SNE, SHapley Additive exPlanations

(SHAP) and ablation experiments are proposed to overcome the non-interpretability issue of the machine learning models [36–38]. At the same time, most of the current detection models detect transport modes, with their categories including car, walk, bike, and public transport. Some studies refine "bike" into bike and e-bike. For public transport, studies consider them as a broad category. For example, they do not distinguish between the bus and the subway [30,39,40]. The coarse-grained classification of public transport results in the inability to target groups with specific transport modes for interventions in transportation planning. Therefore, achieving accurate individual transport mode detection for improved travel demand management still presents many areas for research.

1.2. Literature review

1.2.1. Features for transport mode detection

The detection of transport mode plays a crucial role in managing travel demand. The choice of transport mode is influenced by a variety of factors including age [4], gender [5], income [6], occupation [7], attitudes [8], trip purpose [9], household structure [10], household characteristics [11] number of household members [12], car ownership [13] and built environment [14]. For example, Kim et al. found that the built environment has a greater impact on young and elderly individuals compared to middle-aged individuals in the choice of public transportation [4]. Subjective determinants like travel satisfaction can also influence transport mode detection. A study found that pro-environmental attitudes have a positive impact on the utility of subway and walking as modes of travel, and this preference increases with the duration of walking trips [41]. Another study found that cars promote subjective well-being by facilitating leisure activities, but they also diminish a sense of belonging and achievement through shopping activities [42]. The acquisition of this information can be time-consuming and challenging [26,43]. Respondents may provide inaccurate information due to privacy concerns, which affect the efficacy of subsequent research. Therefore, there is potential to propose a method to both protect individual privacy and accurately detect the transport mode.

Research on individual spatial mobility features has provided the potential for detecting modes. Gonzalez et al. first introduced physical concepts such as the COM and radius of gyration to study individuals' mobility features [44]. They found that individuals' trajectories exhibit high spatiotemporal regularities, with each individual displaying time-invariant features such as travel distance and a significant probability of returning to specific frequent locations. Another study revealed that if a person's future mobility behaviour is influenced by their historical movements, it can be considered a memory-dependent human mobility model [45]. Later, Hong et al. observed that individuals from both datasets maintain a consistent number of combinations of travel modes and activity locations over time [46]. However, the currently proposed mobility features alone may not be sufficient for detecting travel modes. Velocity and acceleration play crucial roles in transport mode detection. To improve the accuracy of transport mode detection using trip information, we have introduced velocity and acceleration of the COM in the mobility features.

1.2.2. Models for transport mode detection

Traditionally, DCMs have been employed to detect transport modes of individuals. DCMs based on the random utility maximization theory commonly include forms such as the multinomial logit

model, nested logit model and mixed logit model [34]. These models have a mathematical structure, making them highly interpretable. Nevertheless, DCMs require many specific features for assessing their impact on transport mode detection, demanding the collection of extensive and detailed data [26]. The acquisition of this data is time-consuming and involves significant implementation costs. Additionally, DCMs are based on model assumptions, such as the independence of irrelevant alternatives (IIA) for the multinomial logit model. If these assumptions are violated, it will lead to biases in parameter estimation and model detection.

Machine learning models which are driven by data are emerging as feasible alternatives for detecting individual transport modes. Various machine learning methods like support vector machines (SVM) [47], k-nearest neighbors (KNN) [27], decision trees (DT) [48] and artificial neural networks (ANN) [49] have been employed in transport mode detection without stringent assumptions. As machine learning models develop, individual models have exhibited limitations in predicting various aspects. To enhance detection accuracy and robustness, ensemble models such as random forests (RF) [30], XGBoost [50] and CatBoost [51] have been developed. For example, a study compared the performance of seven different machine learning models in detecting transport modes and successfully detected four modes: walk, bike, public transport and car [27]. Omrani used four machine learning models to detect individual transport modes in Luxembourg, aggregating walking and biking into one category and recognizing three transport modes: car, public transport and soft mode [52]. In another study, Zhao et al. compared four different types of machine learning models, namely Naive Bayes, tree-based models, SVM and neural networks (NN) for predicting individual choices among four transport modes: car, walk, bike and public transport [26].

Recently, a novel machine learning model called Deep Forest (DF) has been developed and has achieved success in various domains, including image recognition, natural language processing and anomaly detection [53]. The DF model demonstrates the ability to handle high-dimensional data, capture intricate patterns, and make accurate predictions with limited data. The advantages of the DF model make it a potential method to enhance the accuracy and reliability of transport mode detection models. However, most of the methods proposed in literature lack interpretability, which limits their applicability in policy-making. Additionally, these models can detect only a limited number of transport modes, typically three or four. They can detect public transport modes but are unable to distinguish between buses and subways within the category of public transport.

1.3. Objectives of this research

The review of these studies reveals that some aspects need to be improved in the current research: (i) The traditional DCMs employ individual and household attribute data to detect transport modes. These attribute data involve individual privacy and affect the applicability of the model. (ii) The DCMs produce biased detections, though they reveal the contribution of the variables. The machine learning models are more effective for mode detection, but most models need more interpretability. (iii) Current studies mainly treat public transport as a broad category, making it ill-suited for implementing fine transportation planning.

Therefore, an interpretable machine learning model is proposed for the detection of transport modes. We ignore individual and household attribute data and consider individual mobility features to train the detection models. A wide range of multi-source data that can be utilized to detect transport modes is easily accessible with the rapid development of big data, such as land use mix, GDP,

population and online map service data. We combine individual mobility features and multi-source data as detection features for the models. The different machine learning models using detection features were applied to detect fine-grained transport modes, which were divided into seven categories. Specifically, the categories of public transport are subdivided into bus, metro and public transport (including transfers). We calculate four performance metrics to verify the performance of the proposed models for the transport modes. Furthermore, t-SNE and ablation experiments are applied to improve the interpretability of the machine learning models.

The main contributions of this study are as follows. First, the individual and household factors are ignored in the transport mode detection models in order to protect individual privacy and improve the applicability of the model. Thus, the study innovatively considers individual mobility features and combines them with multi-source data, including land use mix, GDP, population and online map service data as detection features. Third, the performance of the deep forest machine learning models has been investigated in a comparative study with traditional classifiers for transport mode detection. We detect the transport modes at the fine-grained level, which are divided into seven categories: car, walk, bike, e-bike, bus, metro and public transport (including transfers). Finally, t-SNE and ablation experiments are employed to enhance the interpretability of the machine learning models.

The remainder of the paper is organized as follows: Section 2 describes the proposed method for transport mode detection, including the extraction of individual mobility features, processing of multi-source data and interpretable machine learning detection models. Sections 3 and 4 discuss the experimental results of different machine learning models using Nanjing as an example. Finally, in Section 5, we summarize the study and propose future research directions.

2. Methodology

The proposed model for detecting transport modes compromises three steps. First, individual mobility features are extracted, making use of the spatial movement. Next, land use mix and online map service data are extracted as multi-source data to enhance the performance of transport mode detection. Finally, these extracted features and multi-source data serve as inputs for the proposed detection model. The performance of models is assessed using performance metrics such as accuracy, recall, precision and F1-measure to verify its effectiveness. The following sections discuss in detail the steps and substeps of the proposed method.

2.1. Overview

Historical movements influence mobility of individuals. Information about the choice of transport modes is implied in individual mobility features [45,54]. The COM of trips is the center of individual spatial mobility, which is a vital feature in describing individual mobility. Radius of gyration is used to measure the extent of individual travel activity. These are the essential features of individual mobility [44,55]. Velocity and acceleration are critical features in transport mode detection. These two features are introduced to detect the transport mode more accurately from the trip information. We innovatively introduce the velocity and acceleration of the COM in the mobility features.

Many multi-source datasets used to detect transport modes are readily available with the rapid development of big data [56,57]. First, the share of different transport modes varies in areas with different land uses. Land use mix is an important indicator of the choice of transport mode [58,59].

Second, individual income levels and regional development levels also influence the choice of transport modes [60]. Owing to the anonymity of the data, it is impossible to identify the socioeconomic and demographic backgrounds of individuals. The proposed model considers the analysis in different administrative regions, using regional GDP and population as detection features. Finally, online map service data can be used to obtain the travel path, distance and time of different transport modes, which can further improve the performance of the detection models [61]. Hence, this study uses the land use mix, GDP, population, and online map service data as the features to detect the transport modes of individuals.

We combine individual mobility features and multi-source data, including land use mix, GDP, population and online map service data as detection features. The machine learning models are applied to detect transport modes, which are divided into seven categories: car, walk, bike, e-bike, bus, metro and public transport (including transfers). To enhance the interpretability of the machine learning models, t-SNE and ablation experiments are conducted in the analysis.

2.2. Individual mobility features extraction

Center of mass. COM $\vec{c}_m$ for individual trips is determined by considering an individual's stay locations and the frequency of their stays at various spatial positions. This metric can be viewed as the central location within an individual's spatial mobility range. To calculate the COM for individual trips, we map the stay location of an individual onto a two-dimensional (2D) coordinate system and use the following formula:

$$\vec{c}_m = \frac{\sum_i n_i \vec{r}_i}{\sum_{i \in L} n_i} \tag{1}$$

where i is the stay location of the individual, n_i is the number of stays at location I, L is the set of stay locations and $\vec{r}_i$ is the 2D vector coordinate of stay location i.

Velocity and Acceleration of the COM. The velocity and acceleration of the COM are applied to calculate the velocity and acceleration of the COM for the trips of an individual. The velocity of the COM indicates the speed at which individuals move through space, while the acceleration of the COM reflects changes in an individual's mobility, such as instances of acceleration or deceleration during a journey. Stay location i to stay location $i + 1$ constitutes one trip i, where each trip includes a transport mode. The distance from stay location i to $i + 1$ is l_i, and the trip duration is t_i. The velocity $\vec{v}_i$ and acceleration $\vec{a}_i$ of trip i are calculated for each individual trip.

$$l_i = \sqrt{\left(\vec{r}_{i+1} - \vec{r}_i\right)^2} \tag{2}$$

$$\vec{v}_i = \frac{l_i}{t_i}, \quad \vec{a}_i = \frac{l_i}{t_i^2} \tag{3}$$

Except for the velocity of the COM mentioned below, the calculation procedure for the acceleration of the COM is the same. The calculation procedure is illustrated in Figure 1. Specifically, the velocity of trip i is decomposed into velocity in the horizontal $\vec{v}_{ix}$ and velocity in the vertical

direction $\vec{v}_{iy}$.

$$\vec{v}_{iy} = \frac{\left|y_{r_{i+1}} - y_{r_i}\right|}{l_i}\vec{v}_i = \frac{\left|y_{r_{i+1}} - y_{r_i}\right|}{t_i}, \quad \vec{v}_{ix} = \frac{\left|x_{r_{i+1}} - x_{r_i}\right|}{l_i}\vec{v}_i = \frac{\left|x_{r_{i+1}} - x_{r_i}\right|}{t_i} \tag{4}$$

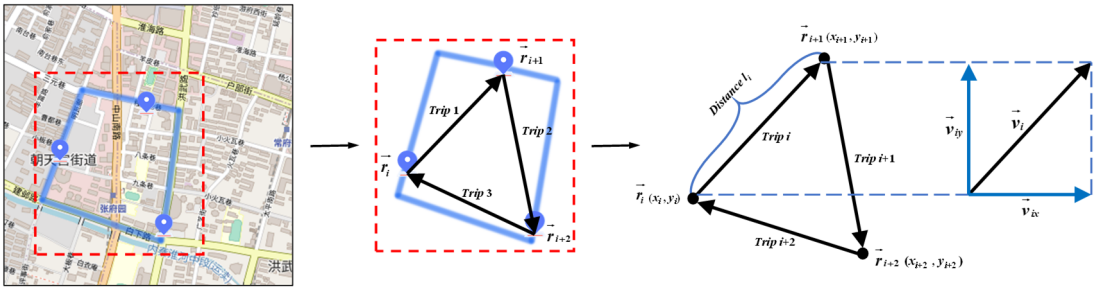
where x_{r_i} is the latitude of stay location I and y_{r_i} is the longitude of stay location i.

Then the velocity of COM v_c is calculated using the frequency of each stay location.

$$\vec{v}_{cx} = \frac{\sum_{i=1}^{N-1} n_i \vec{v}_{ix}}{\sum_{i \in L} n_i}, \quad \vec{v}_{cy} = \frac{\sum_{i=1}^{N-1} n_i \vec{v}_{iy}}{\sum_{i \in L} n_i} \tag{5}$$

$$v_c = \sqrt{\vec{v}_{cx}^{\,2} + \vec{v}_{cy}^{\,2}} \tag{6}$$

where $\vec{v}_{cx}$ is the horizontal velocity of the COM and $\vec{v}_{cy}$ is the vertical velocity of the COM.



(a) Trip of individual (b) Mathematization of individual trip (c) Calculation of the velocity of the COM

Figure 1. Procedure for calculating the velocity for the center of mass (COM).

Radius of gyration. Radius of gyration r_g represents the weighted average of distances from an individual's stay locations to the COM. This metric is employed to measure the spatial extent covered by individual mobility. COM is used as the axis of rotation, and the frequency of the individual at the stay location is used as the mass. r_g for the individual trip is calculated as follows:

$$r_g = \sqrt{\frac{1}{N}\sum_{i \in L} n_i \left(\vec{r}_i - \vec{c}_m\right)^2} \tag{7}$$

2.3. Multi-source data process

Land Use Mix. Land use mix is an important indicator that reflects the choice of the transport mode. In this study, the land use mix were used as detection features. Land use data were obtained from the Chinese essential urban land use categories (EULUC) dataset [62]. We draw on the code for planning standards of development land, and classify land use into six categories: residential,

administration and public services, commercial and business facilities, industrial, street and transportation and green space. The processing method involves the following steps, as illustrated in Figure 2.

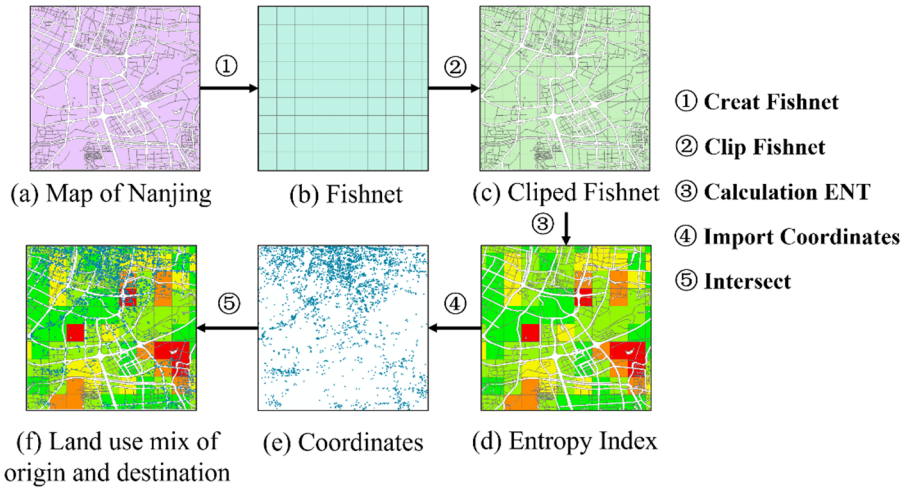


Figure 2. Calculation process of land use mix.

First, the size of the land use map was chosen as the basis for creating the fishnet. The width and height of the cell size were set to 1 km in the fishnet. We clip the fishnet and overlay it with land use data. The entropy index (*ENT*) is usually applied to measure land use heterogeneity. The calculation procedure is as follows:

$$ENT = \frac{-\sum\left(A_{ij} \ln A_{ij}\right)}{\ln N_j} \tag{8}$$

where A_{ij} is the proportion of land use type i in cell j and N_j is the number of land use types in grid j. *ENT* ranges from zero to one.

Finally, the coordinate data of the trip origin and destination were imported to calculate *ENT* at each destination and origin.

Online Map Service Data. We introduce online map service data as multi-source data to improve detection performance [61]. Specifically, the Baidu application programming interface (API) is utilized to obtain the shortest travel path of individuals in seven transport modes: car, walk, bike, e-bike, bus, metro and public transport-including transfers.

The Baidu API is an open map solution interface for developers that provides route planning, real-time navigation, global positioning and other services. We mainly used the route planning function, which can query travel routes. The parameters include the latitude and longitude coordinates of the origin and destination of the trip, geographic coordinate system type and developer key. We consider that the four transport modes of route change less in a short period (including driving, riding a bike or e-bike and walking), while public transport needs to consider departure time. Therefore, the parameters of public transport include departure time, date and tactics. It is worth mentioning that the parameters named "Departure_date" and "Departure_time" respectively represent the date and time of the trip's departure. Therefore, it is possible to return the route parameters based on the specified date and time.

Table 1 lists the public transport API web addresses and parameters.

Table 1. Public transport API web address and parameters.

Parameters	Value	Description
Origin	32.048245, 118.796519	Trip origin (Latitude, longitude)
Destination	32.052371, 118.890583	Trip destination (Latitude, longitude)
Coord_type	WGS84	Type of geographic coordinate system
Departure_date	20150618	Date of trip origin
Departure_time	08:00	Time of trip origin
Tactics_incity	0	0: Recommended mode; 1: Minimal interchanges mode; 2: Minimal walking mode; 3: No metro mode; 4: Short time mode; 5: Metro priority mode.
URL	https://api.map.baidu.com/direction/v2/transit?	

These parameter values are input to a uniform resource locator (URL) and requested using Python to obtain the shortest travel path for each transport mode. The returned results include the travel time, distance and path for the entire trip. This method determines the transport mode of individual trips by comparing the travel time from the survey data with the time taken by the Baidu API.

2.4. Machine learning models for transport mode detection

Many machine learning models are used for classification. The task of the models is to detect the transport modes using key feature sets, with the entire process being supervised. Python and open-source libraries were used for all these models, and the parameters were transparent during the detection process.

Nonetheless, it is unclear which model is the best for our detection. Therefore, we tested several machine learning models. Specifically, we used logistic regression (LogisticReg), SVM, KNN, DT and RF. All the models were applied using the scikit-learn open-source library. In addition, we used artificial neural networks (ANN) [63], XGBoost [50], CatBoost [51] and DF [53]. Moreover, one or more hyperparameters must be optimized to achieve the best performance for each model. We tuned each model to the optimal parameters using a grid search to avoid the problem of unfairness due to improperly set hyperparameters. Unless otherwise stated, the default hyperparameter configuration of the respective models was maintained.

To verify the performance of the fine-grained detection models for the transport modes, we used the confusion matrix to calculate the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Then, we calculated four performance metrics through the classification results, including accuracy, recall, precision and F1-measure [64]. The performance metrics are calculated using the following equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 - measure = \frac{2TP}{2TP + FP + FN} \tag{12}$$

3. Dataset

The dataset used in this study is collected in the Nanjing Transport Development White Paper project. The collection date was June 18, 2015. The survey randomly selected 22,191 participants from 8547 households in 11 administrative districts of Nanjing.

Table 2. Descriptive statistics of extracted features.

Variables	Source of variables	Description	
Continuous Variables		Mean	Std.
Longitude of trip origin	Collected	118.801758	0.077547
Latitude of trip origin	Collected	32.025352	0.156764
Longitude of trip destination	Collected	118.801786	0.077544
Latitude of trip destination	Collected	32.025306	0.156759
Time of trip origin (hours, minutes)	Collected	11.63, 16.04	4.68, 16.70
Time of trip destination (hours, minutes)	Calculated	11.92, 24.71	4.69, 16.82
Trip duration (h)	Calculated	0.44	0.30
Distance l_i (km)	Calculated	4.30	6.60
GDP of administrative districts	Collected	781.53	336.50
Population of the administrative district	Collected	90.16	30.12
Center of mass $\vec{c}_m$	Calculated	13224.95, 3766.65	8281.63, 20296.36
Velocity of COM v_c (km/h)	Calculated	8.30	8.56
Acceleration of COM a_c (km/h^2)	Calculated	24.15	28.40
Radius of gyration r_g (km)	Calculated	2.28	3.33
Entropy index *ENT*	Calculated	0.46	0.35
Categorical variables			
Transport modes determined by Baidu API	From Baidu API	Car; walk; bike; e-bike; bus; metro; public transport (including transfers)	

In this study, the dataset was cleaned because the scope of our research included intra-city transport modes. The cleaned dataset comprises a total of 43,039 trips, including 13,301 walking trips, 4692 bike trips, 10,300 e-bike trips, 4996 bus trips, 8135 car trips, 335 public transport trips (including transfers) and 1280 metro trips. Public transport encompasses a broad category, including both buses and metro transfers. By adopting this classification approach, we can better capture the

diversity and nuances of travel behaviours, particularly concerning the usage of public transport involving bus and metro transfers. We only use trip information in the survey data, including coordinates of the trip origin, coordinates of the trip destination, time of the trip origin, time of the trip destination and the label of the transport mode. The extracted features of the proposed method are presented in Table 2. These features are used as inputs to the model. It is worth mentioning that the indicators of the calculation process are used as features, including $\vec{v}_{ix}$, $\vec{v}_{iy}$, $\vec{v}_{cx}$, $\vec{v}_{cy}$, $\vec{a}_{ix}$, $\vec{a}_{iy}$, $\vec{a}_{cx}$ and $\vec{a}_{cy}$.We used 80% of the dataset for training and 20% for testing.

Many studies have determined the radius of gyration distribution $P(r_g)$ using a truncated power-law statistic for all individuals to explore the statistical properties of individual trip features [65].We also used statistics to distribute the mobility of individuals in Nanjing:

$$P\left(r_g\right) = \left(r_g + r_g^0\right)^{-\beta_r} \exp\left(-r_g / k_r\right) \tag{13}$$

where $r_g^0 = 0.91$, $\beta_r = 0.81$, $k_r = 8$. The r-squared statistic is used to show how well the data fit the regression model. The r-squared statistic is calculated as follows:

$$r^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \bar{y}_i\right)^2} \tag{14}$$

where y_i is the observed values of the data. $\hat{y}_i$ represents the predicted values by the model. $\bar{y}_i$ represents the mean of the observed values.

The r-squared statistic for the truncated power-law function mentioned above is 0.999. It is evident that the radius of gyration decreases rapidly from 1 to 10 km, and the long tail part shows a discrete state, as shown in Figure 3. This phenomenon indicates that the daily activities of most individuals were limited to from 1 to 10 km. The share of long-distance travelers was relatively low.

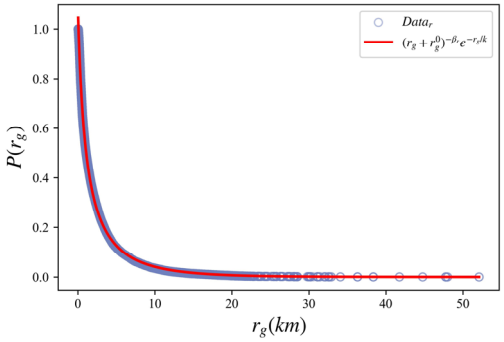


Figure 3. Distribution $P(r_g)$ of the radius of gyration measured for the individuals.

Meanwhile, we counted the velocity and acceleration of the COM introduced in this study to discover the patterns. Surprisingly, the distribution is also very close to the truncated power-law:

$$P\left(v_c\right) = \left(v_c + v_c^0\right)^{-\beta_v} \exp\left(-v_c / k_v\right) \tag{15}$$

$$P(a_c) = \left(a_c + a_c^0\right)^{-\beta_a} \exp\left(-a_c / k_a\right) \tag{16}$$

where v_c^0 = 1954.9, β_v = -0.015, k_v=7.03, r-squared = 0.998, a_c^0 = 0.9, β_a = -0.11, k_a = 16.4 and r-squared = 0.997. As previously depicted in Figures 4 and 5, slow-speed transport modes account for a relatively large proportion of all transport modes, whereas the acceleration of all travel modes is low. The main reason is that the study only considers intra-city transport modes excluding high-speed rail and airplanes, resulting in a lower velocity of the COM.

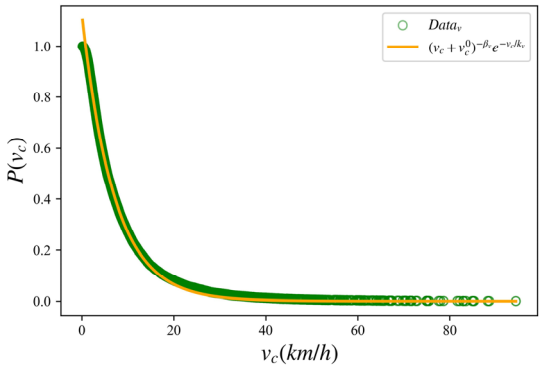


Figure 4. Distribution $P(v_c)$ of the velocity for the COM.

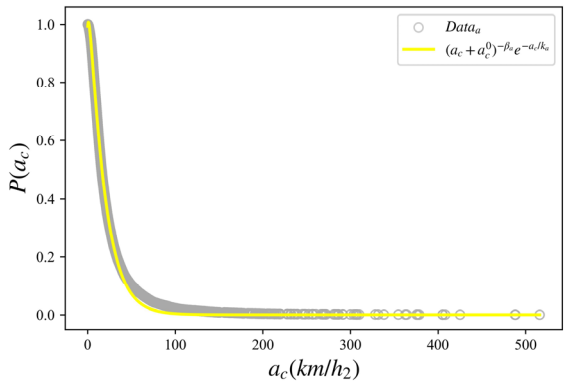


Figure 5. Distribution $P(a_c)$ of the acceleration for the COM.

4. Results

In order to validate the detection performance of various machine learning models, experiments were conducted using travel survey data from Nanjing, China in 2015. The primary focus of these experiments was to assess the accuracy and effectiveness of these models in transport mode detection. For the most promising model in terms of performance, further analysis was performed to provide detailed detection results for each transport mode, aiming to explore the reasons for any shortcomings in the detection performance of model. Furthermore, we explored the interpretability of the model, examining its role in the process of transport mode detection.

4.1. Performance results of machine learning models

The four standard metrics used to assess the performance of the detection process are accuracy, recall, precision and the F1-measure. The F1-measure is the harmonic average of precision and recall. Hence, the accuracy and F1-measure were selected as performance metrics in this study. Meanwhile, we selected several machine learning models for comparison. These models include logisticReg, SVM, KNN, DT, RF, ANN, CatBoost, XGBoost and DF. We also compared the results obtained directly from the Baidu API for detecting transport modes. The SVM used a radial basis function kernel to obtain the best results, and the cascade layer of the deep forest was automatically fitted to four layers. We used five layers (including three hidden layers) in the ANN to detect transport modes. The ANN includes 28 input neurons and 7 output neurons. Figure 6 shows the performance of all models.

Table 3. Performance results of machine learning models.

Models	Accuracy	Precision	Recall	F1-measure
Baidu API	0.45	0.37	0.39	0.35
logisticReg	0.58	0.35	0.34	0.31
SVM	0.64	0.43	0.73	0.41
KNN	0.68	0.6	0.56	0.58
DT	0.71	0.71	0.71	0.71
RF	0.8	0.67	0.83	0.73
ANN	0.63	0.52	0.64	0.56
CatBoost	0.74	0.62	0.72	0.65
XGBoost	0.75	0.64	0.76	0.68
DF	0.82	0.82	0.7	0.74

Table 3 demonstrates that the deep forest model achieves the best detection results among all the models, with an accuracy and F1-measure of approximately 0.82 and 0.74, respectively. While the RF model also performs well, it does not match the performance of the deep forest model. Our results suggest that the deep forest may have better applicability to our proposed detection model and further supports our decision to use the deep forest model for the remainder of the study.

4.2. Performance results of the DF model

In the previous experiments, we evaluated the performance of several machine learning models. These results suggest that the deep forest model is an effective and robust machine learning model for detecting fine-grained transport modes. The specific detection results for each transport mode were output further to explore the reasons for the imperfect detection performance of the DF model. The results are presented in Figure 6.

As seen in Figure 6, the performance of detecting walking is the best for either performance metric. The walking speed is lower than that of other transport modes, resulting in better detection through the velocity of the COM. Meanwhile, public transport has the worst detection performance, with an accuracy of only 0.31. Public transport may involve a transfer process that includes waiting and transfer times. Moreover, interchange transport modes are unknown, and detecting public transport (including transfers) is the most challenging.

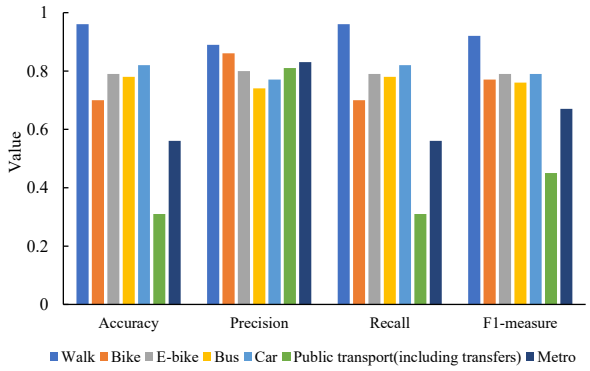
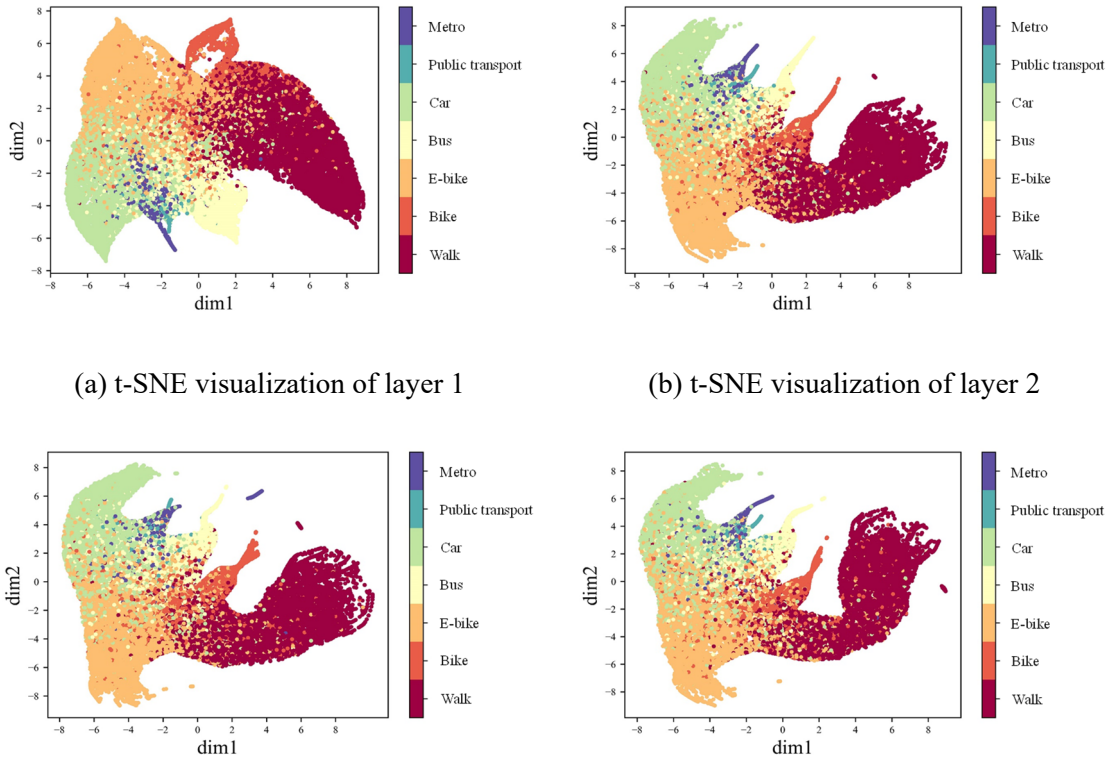


Figure 6. Performance results of each single-mode detection.

4.3. Interpretability of the DF model

In fact, it is unclear which individual transport mode is actually misidentified in the data or which features are more important. The reason is due to the need for more interpretability of machine learning models. We can find the reason for the detection error and improve detection accuracy if a model can capture the misidentification among different modes or obtain the importance ranking of features. Therefore, we performed t-SNE and ablation experiments to overcome the non-interpretability issue of the machine learning models. We also analysed the confusion matrix of the model.



(a) t-SNE visualization of layer 1

(b) t-SNE visualization of layer 2

(c) t-SNE visualization of layer 3

(d) t-SNE visualization of layer 4

Figure 7. t-SNE visualization of four layers.

First, we extracted augmented features of the deep forest detection process; there were four layers of features in the detection process of this study in total. They are challenging to understand and visualize because their features are highly dimensional. t-SNE, which is used for high-dimensional feature visualization, is used to reduce the dimensionality of the features [36,66] achieved by minimizing the Kullback-Leibler divergence between a joint probability distribution in high and low dimensional spaces. High-dimensional data are eventually mapped into 2D, which is more suitable for human observation. A visualization of t-SNE is shown in Figure 7. Subsequently, we extracted the confusion matrix of the model, as shown in Table 4.

Table 4. Confusion matrix of deep forest model.

		Actual result						
		walk	bike	e-bike	bus	car	public transport	metro
Predicted Result	walk	2578	26	47	18	10	0	0
	bike	133	637	90	33	30	0	1
	e-bike	154	43	1642	75	129	0	9
	bus	38	11	98	750	127	1	8
	car	35	12	158	77	1303	5	7
	public transport	0	0	2	13	21	24	2
	metro	1	1	26	28	60	2	143

Figure 7 and Table 4 illustrate the reasons for the imperfect detection of transport modes by the deep forest model. As depicted in the figure, some of the green, orange and blue-green dots are heavily intermingled. This is consistent with the results of the confusion matrix. This phenomenon means that the detection of cars, e-bikes and buses is biased. One potential explanation is that their travel speeds are not significantly different when urban traffic congestion occurs, resulting in inaccurate detection. Additionally, some of the red and yellow dots are also mixed, indicating a biased detection of walking and bikes due to similar travel speeds. The biased detection of transport modes is a result of the slight speed differences during actual travel process. Therefore, the results indicate that, although the velocity of the COM introduced by the paper can substantially enhance the detection performance, further improvements are necessary in the detection of details.

To further investigate the effectiveness of each feature and obtain the importance ranking of a feature, we designed an ablation experiment (including four features and two multi-source data) [67,68]. The results of the experiments are listed in Table 5. ENT and OMS respectively represent the entropy index of land use and the travel modes recognized by the Baidu API. The checked ones in the table are the reduced features. With each feature reduction, we kept the rest of the features and compared the performance metrics between the current model and our model. This step allows for an explicit analysis of the features that more significantly impact mode detection.

The accuracy decreases by 0.07 when the velocity of the COM is ignored. The second important feature is the acceleration of the COM, which decreases the accuracy by 0.04. The results directly indicate that the two mobility features introduced in this study play crucial roles in mode detection. Additionally, there is a relationship between transport mode and gyration radius, indicating that travel distance influences the choice of transport mode. Thus, individual mobility features are the most critical features that improve the detection performance, enhancing accuracy by 0.14. Furthermore, the land use mix improves the accuracy of detection by 0.01, implying little correlation between the urban

land use mix and the choice of transport modes. The potential reason may be that the dataset of our study includes weekdays and the effect of the land use mix is restricted [69].

Table 5. Results of the ablation experiment.

Features						Performance metrics			
$\vec{c}_m$	r_g	v_c	a_c	ENT	OMS	Accuracy	Precision	Recall	F1-Measure
Ours						0.82	0.82	0.70	0.74
					✓	-0.02	-0.02	-	-0.01
				✓		-0.01	-0.02	-0.01	-0.01
			✓			**-0.04**	**-0.03**	**-0.05**	**-0.04**
		✓				**-0.07**	**-0.04**	**-0.06**	**-0.06**
	✓					-0.01	-0.01	-0.02	-0.02
✓						-0.02	-0.02	-0.02	-0.02

In summary, our proposed model exhibits several advantages. Firstly, the DF model has demonstrated exceptional performance in accurately detecting between seven different transport modes, exhibiting a superior performance compared to alternative models. The high detection performance of the DF model emphasizes the model's robustness in effectively detecting various transport modes. Additionally, the DF model performs well in terms of interpretability, providing valuable insights into the factors that influence individuals' mode choices. However, the model also has some disadvantages. Specifically, the model faces challenges when it comes to detecting public transport modes, particularly those involving transfers. Furthermore, the influence of land use mix on the detection of the model is relatively modest. In conclusion, there is the potential for further research in the future.

4.4. Managerial implications

Precisely detecting various transport modes of individuals provides valuable insights for transportation management authorities. Tailored policies can be formulated to encourage the adoption of sustainable and low-carbon transportation modes for groups with specific transport modes, such as public transit, cycling and walking. This proactive approach helps address growing urban transportation challenges. For example, managers can adjust transportation policies, including fare structures, service frequencies and accessibility to offer more appealing public transportation options. This approach can help reduce urban traffic emissions, relieve congestion issues and promote the overall sustainability of cities.

Meanwhile, recent research has witnessed a surge in the development of innovative approaches for data collection, especially Global Positioning System (GPS) loggers or smartphone travel surveys. Once we detect individual transport modes through our proposed method, these approaches eliminate the necessity for extensive follow-up surveys or direct participant interactions, making data collection more efficient. Moreover, these data collection approaches emphasize the potential of smartphone and GPS travel surveys in providing highly precise transport mode data while preserving participant privacy. Managers can actively embrace these innovative data collection methods to better understand individuals' travel behavior, guide policy formulation and inform transportation planning.

5. Conclusions

Analysis of transport mode choice is a crucial element in developing and evaluating planning policies. Though many kinds of factors influence the transport mode choice of the individual, many of these factors involve the privacy of individuals. Meanwhile, detection models of transport modes widely use discrete choice models. In order to improve the limitations of existing discrete choice models, we propose an interpretable machine learning model for detecting transport modes by multi-source data. The mobility features of individuals, which innovatively introduce the velocity and acceleration of the center of mass (COM), are considered in the detection model. We combine these features with multi-source data as detection features and apply machine learning models to detect transport modes at the fine-grained level, which were divided into seven categories.

Four performance metrics are calculated through the classification results, including accuracy, recall, precision and F1-measure. The results show that the deep forest algorithm exhibits the best detection performance, achieving a detection accuracy of 0.82 in the test dataset. Meanwhile, recall, precision and F1-measure also achieved optimal performance. The evaluation shows that the detection performance for other transport modes is excellent apart from public transport detection. Compared to previous work, this study is a significant step forward in terms of granularity and detection performance. To address the issue of non-interpretability of machine learning models, we also perform t-distributed stochastic neighbor embedding (t-SNE) and ablation experiments and rank the importance of features. These experiments demonstrate a weak correlation between the land use mix and the choice of transportation modes. Moreover, the most critical features are the individual mobility features in the study, and the accuracy improves by 0.13.

Meanwhile, recent research has witnessed a surge in the development of innovative approaches for data collection, especially Global Positioning System (GPS) loggers or smartphone travel surveys. They leverage these travel surveys to collect accurate travel behavior details without involving the collection of individual and household attribute data. Our proposed approach enhances the potential of these data collection methods in maintaining privacy protection for survey participants without the need for extensive follow-up surveys. On the other hand, if individual travel modes can be accurately detected, transportation government agencies can intervene by targeting specific groups with specific transport modes and effectively promote sustainable and low-carbon transportation options. The results of the study can guide government transportation agencies in formulating effective policies to achieve long-term sustainable and efficient urban transportation goals.

In future research, the utilization of multi-source data will be further expanded to enhance the performance and applicability of the proposed model. The expansion may involve incorporating additional data, such as real-time traffic information, weather conditions or other factors that can influence transport mode choice. The objective is to continuously improve the accuracy and reliability of the model in detecting transport modes, and its effectiveness in guiding transportation policies and addressing urban transportation challenges.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. SJCX22_0061); the General Project of National Natural Science Foundation of China (Grant No. 52172317); and the Key Program of National Natural Science Foundation of China (Grant No. 52131203).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Z. Zhang, J. Zhang, Operating subsidies for urban rail transit PPP projects, *J. Tsinghua Univ*., **56** (2016), 1327–1332. https://doi.org/10.16511/j.cnki.qhdxxb.2016.25.046
2. S. Tscharaktschiew, F. Reimann, Less workplace parking with fully autonomous vehicles?, *J. Intell. Connect. Veh*., **5** (2022), 283–301. https://doi.org/10.1108/JICV-07-2022-0029
3. Y. Liu, C. Lyu, Z. Liu, J. Cao, Exploring a large-scale multi-modal transportation recommendation system, *Transp. Res. Part C Emerg. Technol*., **126** (2021), 103070. https://doi.org/10.1016/j.trc.2021.103070
4. K. Kim, K. Kwon, M. W. Horner, Examining the effects of the built environment on travel mode choice across different age groups in seoul using a random forest method, *Transp. Res. Record.*, **2675** (2021), 670–683. https://doi.org/10.1177/03611981211000750
5. M. Yang, D. Li, W. Wang, J. Zhao, X. Chen, Modeling gender-based differences in mode choice considering time-use pattern: Analysis of bicycle, public transit, and car use in Su Zhou, China, *Adv. Mech. Eng*., **2013** (2013), 706918. https://doi.org/10.1155/2013/706918
6. C. R. Bhat, S. Srinivasan, A multidimensional mixed ordered-response model for analyzing weekend activity participation, *Transp. Res. Part B Methodol*., **39** (2005), 255–278. https://doi.org/10.1016/j.trb.2004.04.002
7. L. Cheng, X. Chen, M. Wei, J. Wu, X. Hou, Modeling mode choice behavior incorporating household and individual sociodemographics and travel attributes based on rough sets theory, *Comput. Intell. Neurosci*., **2014** (2014), 26. https://doi.org/10.1155/2014/560919
8. C. Ding, Y. Chen, J. Duan, Y. Lu, J. Cui, Exploring the influence of attitudes to walking and cycling on commute mode choice using a hybrid choice model, *J. Adv. Transp*., **2017** (2017). https://doi.org/10.1155/2017/8749040
9. J. Jeong, J. Lee, T. H. T. Gim, Travel mode choice as a representation of travel utility: A multilevel approach reflecting the hierarchical structure of trip, individual, and neighborhood characteristics, *Pap. Reg. Sci*., **101** (2022), 745–765. https://doi.org/10.1111/pirs.12665
10. C. Ding, D. Wang, C. Liu, Y. Zhang, J. Yang, Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance, *Transp. Res. Part A Policy Pract*., **100** (2017), 65–80. https://doi.org/10.1016/j.tra.2017.04.008
11. P. van den Berg, T. Arentze, H. Timmermans, Estimating social travel demand of senior citizens in the Netherlands, *J. Transp. Geogr*., **19** (2011), 323–331. https://doi.org/10.1016/j.jtrangeo.2010.03.018

12. C. R. Bhat, S. Srinivasan, K. W. Axhausen, An analysis of multiple interepisode durations using a unifying multivariate hazard model, *Transp. Res. Part B Methodol*., **39** (2005), 797–823. https://doi.org/10.1016/j.trb.2004.11.002
13. X. Cao, S. L. Handy, P. L. Mokhtarian, The influences of the built environment and residential self-selection on pedestrian behavior: Evidence from Austin, TX, *Transportation*., **33** (2006), 1–20. https://doi.org/10.1007/s11116-005-7027-2
14. R. Ye, H. Titheridge, Satisfaction with the commute: The role of travel mode choice, built environment and attitudes, *Transp. Res. Part D Transp. Environ*., **52** (2017), 535–547. https://doi.org/10.1016/j.trd.2016.06.011
15. N. F. M. Ali, A. F. M. Sadullah, A. P. P. A. Majeed, M. A. M. Razman, R. M. Musa, The identification of significant features towards travel mode choice and its prediction via optimised random forest classifier: An evaluation for active commuting behavior, *J. Transp. Health*., **25** (2022) ,101362. https://doi.org/10.1016/j.jth.2022.101362
16. C. Ding, Y. Wang, T. Tang, S. Mishra, C. Liu, Joint analysis of the spatial impacts of built environment on car ownership and travel mode choice, *Transp. Res. Part D Transp*., **60** (2018), 28–40. https://doi.org/10.1016/j.trd.2016.08.004
17. L. Shen, P. R. Stopher, Review of GPS travel survey and GPS data-processing methods, *Transp. Rev*., **34** (2014) ,316–334. https://doi.org/10.1080/01441647.2014.903530
18. N. Caceres, L. M. Romero, F. G. Benitez, Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys, *Transportmetrica A: Transport Sci.*, **16** (2020), 574–601. https://doi.org/10.1080/23249935.2020.1720857
19. R. J. Lee, I. N. Sener, J. A. Mullins, An evaluation of emerging data collection technologies for travel demand modeling: From research to practice, *Transp. Lett*., **8** (2016), 181–193. https://doi.org/10.1080/19427867.2015.1106787
20. Y. Liu, E. Miller, K. N. Habib, Detecting transportation modes using smartphone data and GIS information: evaluating alternative algorithms for an integrated smartphone-based travel diary imputation, *Transp. Lett*., **14** (2022), 933–943. https://doi.org/10.1080/19427867.2021.1958591
21. K. Chin, H. Huang, C. Horn, I. Kasanicky, R. Weibel, Inferring fine-grained transport modes from mobile phone cellular signaling data, *Comput. Environ. Urban Syst*., **77** (2019), 101348. https://doi.org/10.1016/j.compenvurbsys.2019.101348
22. K. Gao, H. Wang, S. Wang, X. Qu, Data and code disclosure and sharing policy of communications in transportation research, *Commun. Transp. Res*., **2** (2022), 100055. https://doi.org/10.1016/j.commtr.2022.100055
23. L. Cheng, X. Chen, S. Yang, An exploration of the relationships between socioeconomics, land use and daily trip chain pattern among low-income residents, *Transp. Plan. Technol*., **39** (2016), 358–369. https://doi.org/10.1080/03081060.2016.1160579
24. S. A. O. Medina, Inferring weekly primary activity patterns using public transport smart card data and a household travel survey, *Travel Behav. Soc*., **12** (2018), 93–101. https://doi.org/10.1016/j.tbs.2016.11.005
25. Q. Yuan, X. Xu, T. Wang, Y. Chen, Investigating safety and liability of autonomous vehicles: Bayesian random parameter ordered probit model analysis, *J. Intell. Connect. Veh*., **5** (2022), 199–205. https://doi.org/10.1108/JICV-04-2022-0012

26. X. Zhao, X. Yan, A. Yu, P. Van Hentenryck, Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models, *Travel Behav. Soc*., **20** (2020), 22–35. https://doi.org/10.1016/j.tbs.2020.02.003
27. J. Hagenauer, M. Helbich, A comparative study of machine learning classifiers for modeling travel mode choice, *Expert Syst. Appl*., **78** (2017), 273–282. https://doi.org/10.1016/j.eswa.2017.01.057
28. T. Hillel, M. Bierlaire, M. Z. E. B. Elshafie, Y. Jin, A systematic review of machine learning classification methodologies for modelling passenger mode choice, *J. Choice Model*., **38** (2021), 100221. https://doi.org/10.1016/j.jocm.2020.100221
29. Y. Liu, F. Wu, Z. Liu, K. Wang, F. Wang, X. Qu, Can language models be used for real-world urban-delivery route optimization?, *Innovation*, 2023. https://doi.org/10.1016/j.xinn.2023.100520
30. L. Cheng, X. Chen, J. D. Vos, X. Lai, F. Witlox, Applying a random forest method approach to model travel mode choice behavior, *Travel Behav. Soc*., **14** (2019), 1–10. https://doi.org/10.1016/j.tbs.2018.09.002
31. P. Salas, R. de la Fuente, S. Astroza, J. A. Carrasco, A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity, *Expert Syst. Appl*., **193** (2022), 116253. https://doi.org/10.1016/j.eswa.2021.116253
32. L. Cheng, X. Lai, X. Chen, S. Yang, J. D. Vos, F. Witlox, Applying an ensemble-based model to travel choice behavior in travel demand forecasting under uncertainties, *Transp. Lett*., **12** (2020), 375–385. https://doi.org/10.1080/19427867.2019.1603188
33. W. Li, K. Xiao, Y. Ren, C. Li, Y. Fan, Path planning and control method for vehicle obstacle avoidance in pedestrian crossing scenes, *J. Automot. Saf. Energy*, **13** (2022), 489–501. https://doi.org/10.3969/j.issn.1674-8484.2022.03.010
34. Y. Hu, T. Jiang, X. Liu, Y. Shi, Pedestrian-crossing intention-recognition based on dual-stream adaptive graph-convolutional neural-network, *J. Automot. Saf. Energy*., **13** (2022), 325–332. https://doi.org/10.3969/j.issn.1674-8484.2022.02.013
35. I. Ullah, K. Liu, T. Yamamoto, M. Zahid, A. Jamal, Modeling of machine learning with SHAP approach for electric vehicle charging station choice behavior prediction, *Travel Behav. Soc*., **31** (2023), 78–92. https://doi.org/10.1016/j.tbs.2022.11.006
36. L. van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res*., **9** (2008).
37. S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process Syst.*, **30** (2017). https://doi.org/10.48550/arXiv.1705.07874
38. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *Adv. Neural Inf. Process Syst.*, 2015. https://doi.org/10.48550/arXiv.1506.01497
39. M. T. Kashifi, A. Jamal, M. S. Kashefi, M. Almoshaogeh, S. M. Rahman, Predicting the travel mode choice with interpretable machine learning techniques: A comparative study, *Travel Behav. Soc*., **29** (2022), 279–296. https://doi.org/10.1016/j.tbs.2022.07.003
40. Y. Zheng, J. Xiao, X. Hua, W. Wang, H. Chen, A comparative analysis of the robustness of multimodal comprehensive transportation network considering mode transfer: A case study, *Electron. Res. Arch*., **31** (2023), 5362–5395. https://doi.org/10.3934/era.2023272
41. A. A. Toorzani, A. A. Rassafi, Pro-environmental attitude and adherence to a travel mode in an integrated choice and latent variable (ICLV) model: results from a revealed preference survey, *Int. J. Civ. Eng*., **21** (2023), 235–249. https://doi.org/10.1007/s40999-022-00757-6

42. Y. Tran, N. Hashimoto, T. Ando, T. Sato, N. Konishi, Y. Takeda, et al., The indirect effect of travel mode use on subjective well-being through out-of-home activities, *Transportation*, **2023** (2023), 1–33. https://doi.org/10.1007/s11116-023-10408-x
43. J. De Vos, P. L. Mokhtarian, T. Schwanen, V. Van Acker, F. Witlox, Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility, *Transportation*, **43** (2016), 771–796. https://doi.org/10.1007/s11116-015-9619-9
44. M. C. González, C. A. Hidalgo, A. L. Barabási, Understanding individual human mobility patterns, *Nature*, **453** (2008), 779–782. https://doi.org/10.1038/nature06958
45. F. Xu, Y. Li, D. Jin, J. Lu, C. Song, Emergence of urban growth patterns from human mobility behavior, *Nat. Comput. Sci*., **1** (2021), 791–800. https://doi.org/10.1038/s43588-021-00160-6
46. Y. Hong, H. Martin, Y. Xin, D. Bucher, D. J. Reck, K. W. Axhausen, et al., Conserved quantities in human mobility: From locations to trips, *Transp. Res. Part C Emerg. Technol*., **146** (2023), 103979. https://doi.org/10.1016/j.trc.2022.103979
47. J. C. Xian-Yu, Travel mode choice analysis using support vector machines, in *ICCTP 2011: Towards Sustainable Transportation Systems*, (2011), 360–371. https://doi.org/10.1061/41186(421)37
48. G. Zhan, X. Yan, S. Zhu, Y. Wang, Using hierarchical tree-based regression model to examine university student travel frequency and mode choice patterns in China, *Transp. Policy*, **45** (2016), 55–65. https://doi.org/10.1016/j.tranpol.2015.09.006
49. H. Omrani, O. Charif, P. Gerber, A. Awasthi, P. Trigano, Prediction of individual travel mode with evidential neural network model, *Transp. Res. Record*., **2399** (2013), 1–8. https://doi.org/10.3141/2399-01
50. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 785–794. https://doi.org/10.1145/2939672.2939785
51. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: Unbiased boosting with categorical features, *Adv. Neural Inf. Process Syst.*, 2018. https://doi.org/10.48550/arXiv.1706.09516
52. H. Omrani, Predicting travel mode of individuals by machine learning, *Transp. Res. Procedia*, **10** (2015), 840–849. https://doi.org/10.1016/j.trpro.2015.09.037
53. Z. H. Zhou, J. Feng, Deep forest, *Natl. Sci. Rev*., **6** (2019), 74–86. https://doi.org/10.1093/nsr/nwy108
54. J. Qin, F. Liao, Space-time prisms in multimodal supernetwork-Part 2: Application for analyses of accessibility and equality, *Commun. Transp. Res*., **2** (2022), 100063. https://doi.org/10.1016/j.commtr.2022.100063
55. P. Widhalm, Y. Yang, M. Ulm, S. Athavale, M. C. González, Discovering urban activity patterns in cell phone data, *Transportation*, **42** (2015), 597–623. https://doi.org/10.1007/s11116-015-9598-x
56. H. Huang, Y. Cheng, R. Weibel, Transport mode detection based on mobile phone network data: A systematic review, *Transp. Res. Part C Emerg. Technol*., **101** (2019), 297–312. https://doi.org/10.1016/j.trc.2019.02.008
57. C. Chen, J. Ma, Y. Susilo, Y. Liu, M. Wang, The promises of big data and small data for travel behavior (aka human mobility) analysis, *Transp. Res. Part C Emerg. Technol*., **68** (2016), 285–299. https://doi.org/10.1016/j.trc.2016.04.005

58. Y. Song, L. Merlin, D. Rodriguez, Comparing measures of urban land use mix, *Comput. Environ. Urban Syst*., **42** (2013), 1–13. https://doi.org/10.1016/j.compenvurbsys.2013.08.001.
59. K. K. W. Yim, S. C. Wong, A. Chen, C. K. Wong, W. H. K. Lam, A reliability-based land use and transportation optimization model, *Transp. Res. Part C Emerg. Technol*., **19** (2011), 351–362. https://doi.org/10.1016/j.trc.2010.05.019
60. M. W. Horner, D. Schleith, Analyzing temporal changes in land-use-transportation relationships: A LEHD-based approach, *Appl. Geogr*., **35** (2012), 491–498. https://doi.org/10.1016/j.apgeog.2012.09.006
61. Z. Peng, G. Bai, H. Wu, L. Liu, Y. Yu, Travel mode recognition of urban residents using mobile phone data and MapAPI, *Environ. Plan. B Urban Anal. City Sci*., **48** (2021), 2574–2589. https://doi.org/10.1177/2399808320983001
62. P. Gong, B. Chen, X. Li, H. Liu, J. Wang, Y. Bai, et al., Mapping essential urban land use categories in China (EULUC-China): preliminary results for 2018, *Sci. Bull*., **65** (2020), 182–187. https://doi.org/10.1016/j.scib.2019.12.007
63. G. Xiao, Z. Juan, C. Zhang, Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization, *Transp Res Part C Emerg Technol*., **71** (2016), 447–463. https://doi.org/10.1016/j.trc.2016.08.008
64. M. Müller-Hannemann, R. Rückert, A. Schiewe, A. Schöbel, Estimating the robustness of public transport schedules using machine learning, *Transp. Res. Part C Emerg. Technol*., **137** (2022), 103566. https://doi.org/10.1016/j.trc.2022.103566
65. C. Song, Z. Qu, N. Blumm, A. L. Barabási, Limits of predictability in human mobility, *Science*, **327** (2010), 1018–1021. https://doi.org/10.1126/science.1177170
66. A. Chatzimparmpas, R. M. Martins, A. Kerren, t-visne: Interactive assessment and interpretation of t-sne projections, *IEEE Trans. Visual Comput. Graphics*, **26** (2020), 2696–2714. https://doi.org/10.1109/TVCG.2020.2986996
67. P. R. Anukrishna, V. Paul, A review on feature selection for high dimensional data, in *2017 International Conference on Inventive Systems and Control (ICISC)*, 2017. https://doi.org/10.1109/ICISC.2017.8068746
68. J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, *Neurocomputing*, **300** (2018), 70–79. https://doi.org/10.1016/j.neucom.2017.11.077
69. M. Zhang, The role of land use in travel mode choice: Evidence from Boston and Hong Kong, *J. Am. Plan. Assoc*., **70** (2004), 344–360. https://doi.org/10.1080/01944360408976383

AIMS Press