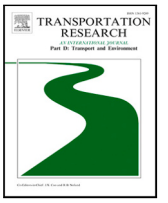
Transportation Research Part D 127 (2024) 104038

Contents lists available at ScienceDirect

Transportation Research Part D

journal homepage: www.elsevier.com/locate/trd



Examining active travel behavior through explainable machine learning: Insights from Beijing, China

Ganmin Yin [a,b], Zhou Huang [a,b,*], Chen Fu [a,b], Shuliang Ren [a,b], Yi Bao [a,b], Xiaolei Ma [c]

[a] Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, China
[b] Beijing Key Lab of Spatial Information Integration & Its Applications, Peking University, Beijing, China
[c] School of Transportation Science and Engineering, Beihang University, Beijing, China

ARTICLE INFO

Keywords:
Urban transportation
Active travel
Active mobility
Walking and cycling
Explainable machine learning
SHAP
Geospatial big data

ABSTRACT

Active travel, namely walking and cycling, is an eco-friendly and socially beneficial mode of sustainable transportation. However, existing research on active travel relies on limited survey data and generalized linear models. To fill the gap, our study integrates large-scale big trip data and data-driven machine learning to simultaneously predict active travel flow and probability. We employ SHapley Additive exPlanation to analyze the nonlinear effects of various characteristics (e.g., travel, socioeconomic, infrastructure, environment) on active travel. Gradient Boosting Decision Tree performs best for both prediction tasks. The overall importance of travel distance is over 50% to the model. Features like crow-fly distance, housing price, point-of-interest density, subway proximity, building area/road density, and urban greenery exhibit pronounced nonlinear effects. Local interpretability analysis reveals the determinants of specific trips, facilitating targeted optimization implications. Our study reveals the drivers and nonlinearities of active travel behavior and aids sustainable transportation planning.

1. Introduction

Urban transportation has always been an important topic in urban planning and sustainable development. With the acceleration of urbanization and the rapid increase of vehicles, problems such as traffic congestion, environmental pollution and energy consumption have become increasingly prominent (Erhardt et al., 2019; Zhang et al., 2019; Chai et al., 2016). In this context, active travel/mobility as a sustainable mode of transportation has attracted widespread attention and research. Compared to mechanized travel, such as private transportation and public transportation, active travel offers many advantages such as flexibility, cost-effectiveness, health benefits, and environmental friendliness (Iroz-Elardo et al., 2020; Frank et al., 2022; Schoner et al., 2018; Carlson et al., 2015). Therefore, active travel is well-known as a representative of sustainable transportation with positive implications for individuals, society and the environment (United Nations, 2021).

To better understand and promote active travel behavior, many researchers have conducted extensive research to investigate the behavior patterns of people's choice to walk and cycle (Xu et al., 2023; Tao et al., 2023; Wali et al., 2021). Nevertheless, there are still some limitations. First, most of them use questionnaires or survey data, which is small-scale and low-resolution (Yang et al., 2022a,b). The limited samples make it difficult to apply to advanced data-driven models, while the low resolution greatly limits its application in refined transportation planning (Liu et al., 2015; Chen et al., 2016). Second, the flow and probability of active travel are rarely considered at the same time (Shaer et al., 2021; Pisoni et al., 2022). Since the former represents the demand for active travel, the latter reflects the willingness of residents, ignoring either dimension will lead to insufficient understanding on

* Corresponding author.
E-mail address: huangzhou@pku.edu.cn (Z. Huang).

https://doi.org/10.1016/j.trd.2023.104038
Received 16 September 2023; Received in revised form 22 November 2023; Accepted 25 December 2023

active travel. Third, most studies use traditional regression models that presuppose linear or pre-defined patterns but ignore the pervasive nonlinearity (Kemperman and Timmermans, 2009; Liu et al., 2020). In this context, it would lead to overestimation or underestimation of the effects when the features fall into different intervals of values (Liu et al., 2021; Xiao et al., 2021).

To address the above issues, this study aims to comprehensively examine active travel behavior by integrating large-scale big trip data and data-driven machine learning models. Specifically, we first obtain the multi-dimensional characteristics of travel, socioeconomic, infrastructure and environment through various geospatial big data, and extract the flow and probability of active travel at the origin–destination (OD) level using mobile phone location data. Next, we use the above multisource features as input and employ machine learning models to accurately predict the flow and probability of active travel. Then, we introduce an explainable artificial intelligence method, called SHapley Additive exPlanation (SHAP), to explain the nonlinear and interaction effects of these features on active travel. Finally, we provide insights for policymakers to tailor optimized policies for specific OD trips, increase the willingness to actively travel, and improve the travel experience of city dwellers. By delving into active travel behavior, we can provide scientific evidence for urban planning and traffic management, and promote the development of sustainable urban transportation.

The main contributions of this research are as follows:

- The flow and probability of active travel are considered at the same time. To our best knowledge, it is the first to consider these two dimensions simultaneously.
- The nonlinear effects of multisource features on active travel are analyzed using explainable machine learning. It is a complement to traditional linear-based models.
- Taking Beijing as an example, we provide insights for targeted transportation optimization policies for specific OD trips, thus demonstrating the potential application value of the proposed method.

The rest of the paper is organized as follows. Section 2 reviews recent related work in active travel and machine learning. Section 3 describes the study area and data, and introduces the methodology framework of this study, including data processing, model prediction, and result analysis. Section 4 presents the important results of the study and offers some effective policy suggestions. Section 5 discusses the theoretical and practical implications of the research while acknowledging its limitations. Section 6 summarizes the main conclusions of the research.

2. Literature review

2.1. Behavior analysis of active travel

Active travel has been acknowledged as an environmentally friendly and sustainable mode of transportation (United Nations, 2021). In terms of the definition, active travel refers to a mode of travel in which the traveler needs to continuously expend physical energy to move (Burbidge and Goulias, 2009; Cook et al., 2022). And it is generally equivalent to walking and cycling in the past literature (Hankey et al., 2017; Pucher et al., 2010; Yang et al., 2022b,a), and has similar meanings to non-motorized travel (Lundberg and Weber, 2014; Rietveld, 2001) and low-speed travel (Rodier et al., 2003). Cook et al. (2022) also examined the concept of active travel, expanded the definition to include other physically exerting modes beyond walking and cycling. Notably, we state that active travel in this study refers specifically to walking and cycling in the commuting context.

The driving factors for active travel are multi-dimensional. Li et al. (2005) conducted a cross-sectional study to examine the relation between built environment factors (i.e., urban form of neighborhoods) and walking activity in older adults. Ma and Dill (2015) adopted binary logit and linear regression models to study how the objective environment (e.g., bike infrastructure, street connectivity) and perceived environment (e.g., safety, quietness, and easiness) affect bicycling behavior with random phone survey data. Shaer et al. (2021) study the relationship between socio-demographic factors (such as age, gender, and income), built environment factors (such as land use, street design, and transit accessibility) and active travel (purpose, duration, and frequency) under the COVID-19 pandemic and policies. Using travel survey data, Pisoni et al. (2022) explored the determinants of active travel choice, including demographic, socio-economic, and cultural factors. And they also quantified the financial benefits after increasing active travel shares. Gao et al. (2023) investigated the impact of urban greenness on the usage of a free-floating bike-sharing system (FFBS) in Beijing, China. They found that the greenery view index (GVI) and the normalized difference vegetation index (NDVI) had different impacts on FFBS, and they emphasized the importance of promoting urban greenness to benefit green traveling. Considering the previous study, we decide to consider multi-dimensional characteristics (including travel, socioeconomic, infrastructure, and environment) to fully understand the determinants of active travel.

Existing research considered multiple dimensions of active travel to better understand its behavior patterns. In these work, some focused on the flow/intensity/ridership of walking and cycling (Fu et al., 2023; Chai et al., 2018; Cao et al., 2019; Nourian et al., 2018), while others considered to predict the usage/odds/interest of active travel (Gao et al., 2023; Ferrari et al., 2020; Yang et al., 2022a,b). However, few studies have taken into account both dimensions of active travel (Wang et al., 2023). Obviously, the former reflects the demand for active travel among city dwellers, while the latter represents the willingness of people to choose walking and cycling. Therefore, our study accurately predicts both active travel flow and probability, fully modeling active travel behavior.

2.2. Machine learning methods for revealing nonlinearity in transportation research

In the field of travel behavior analysis, traditional regression models are dominant, which often presuppose linear or pre-defined patterns between the independent and dependent variables. For instance, what Ma and Dill (2015) have adopted for cycling behavior analysis are binary logit and linear regression models. Malambo et al. (2017) integrated a cross-sectional survey and multivariable logistic regression to investigate the associations between built environment attributes and leisure-time walking. Porter et al. (2018) also adopted the multivariable logistic regression model to uncover how the social and built environment factors affect bicycling behavior with an internet-based survey. Yin et al. (2023b) designed a nested generalized linear model called beta-binomial model to analyze the effects of travel efficiency and socioeconomic characteristics on public transportation usage. However, in many contexts, nonlinearity may be more common in travel behavior analysis, where these models become unreliable (Xiao et al., 2021; Yang et al., 2022a; Liu et al., 2021; Kemperman and Timmermans, 2009).

As a type of data-driven and black-box models, machine learning shines in various fields due to its excellent ability to model complex nonlinear patterns (Jordan and Mitchell, 2015). Many researchers are beginning to apply machine learning to the field of transportation, such as travel behavior analysis, travel mode selection (Hillel et al., 2021; Koushik et al., 2020). Rasouli and Timmermans (2014) used a random forest model to achieve an accurate prediction for travel mode choice. Lee et al. (2018) compared four artificial neural networks (ANNs) and a multinomial logit model for modeling travel mode selection, and they found that ANNs outperformed a lot. Yazdizadeh et al. (2019) designed an ensemble convolutional neural networks (CNNs) to infer the travel mode using a smartphone travel survey dataset.

In active travel modeling, machine learning models are also becoming popular. Pisoni et al. (2022) employed a gradient boosting machine learning approach to study the effects of demographic, socioeconomic, and cultural factors on active travel choice based on a travel survey with 26,500 responses. Yang et al. (2022a) used a random forest model to reveal the nonlinear relationship between built environment and active travel by comparing gender differences among older adults. The survey data used encompassed the trip information and socio-demographics of the participants, with a sample size of 2003. Xu et al. (2023) adopted multiple machine learning models (e.g., random forest, adaptive boosting, support vector machine, k-nearest neighbors, and artificial neural network) to investigate the determinants of post-pandemic active travel preference. They have used an online survey provided by the Alabama Transportation Institute (ATI) with a total of 1402 respondents. Guo et al. (2023) also integrated travel survey data and random forest model to examine the nonlinear effects of social and built environment factors on people's choice of walking and cycling. The resident travel survey was collected in Wuhan, 2020, with 30,174 samples. In summary, most of them used small-scale questionnaires or surveys, so the strong data mining capabilities of machine learning were not fully exploited. In our study, we leverage large-scale (over 1 million users) big trip data instead, thus better uncovering the complex and nonlinear relationship using advanced machine learning.

3. Methodology

3.1. Study area

This study focuses on the travel patterns of commuters living or working in Dongcheng and Xicheng districts in Beijing, as shown in Fig. 1a. Beijing is the capital of the People's Republic of China and is the political, cultural, scientific and technological center. Beijing subordinates 16 districts, with a total area of 16,410.54 square kilometers and a permanent population of about 21.84 million (Beijing Municipal Bureau of Statistics, 2023). Dongcheng District and Xicheng District are the two most central urban districts in Beijing, almost all located within the Third Ring Road, with a total area of 92.54 square kilometers and a permanent population of about 1.80 million. Dongcheng and Xicheng districts are densely populated, economically developed, and have a large number of commuters, making them ideal study areas for transportation research.

3.2. Data

The commuter flow data used in this study is provided by Amap company, China's largest e-mapping company. This data records travel information for commuters who live or work in the study area. The data spans from January to June 2019 with a spatial resolution of 500 m after spatial aggregation. The commuters' homes and workplaces are inferred through the random forest based on the location information from millions of location-based services (LBS) users. After validation with the ground truth location of the registered users, the inference accuracy exceeds 90%, so it is highly reliable (Yin et al., 2023a,b). The data is illustrated in Table 1, where each row represents a pair of ODs, including ID, coordinates of home and workplace (i.e., origin and destination), number of commuters by various modes (including car, bus, subway, and active travel). In this study, we mainly focus on active travel.

According to statistics, the data has 122,047 pairs of ODs, a total of 1,335,769 commuters, which undoubtedly has a high coverage. The distribution of the commuting flow is shown in Fig. 1b, with a heavy-tail distribution. We further counted the number and proportion of people choosing various travel modes, as shown in Fig. 1c. It can be found that in the study area, the subway is the most important travel mode, accounting for 43.3%, and about 17.3% of people choose to walk or cycle for commuting. The choice of travel mode has obvious geographical differences, and Fig. 1d shows the statistical results of commuting between each area by ring road and the Dongcheng & Xicheng. Obviously, the proportion of private car is highest between the 4th and 5th ring roads, the bus usage does not show spatial differences, the farther from the city center the proportion of the subway is higher and the proportion of active travel is lower, which may be closely related to the distance factor.

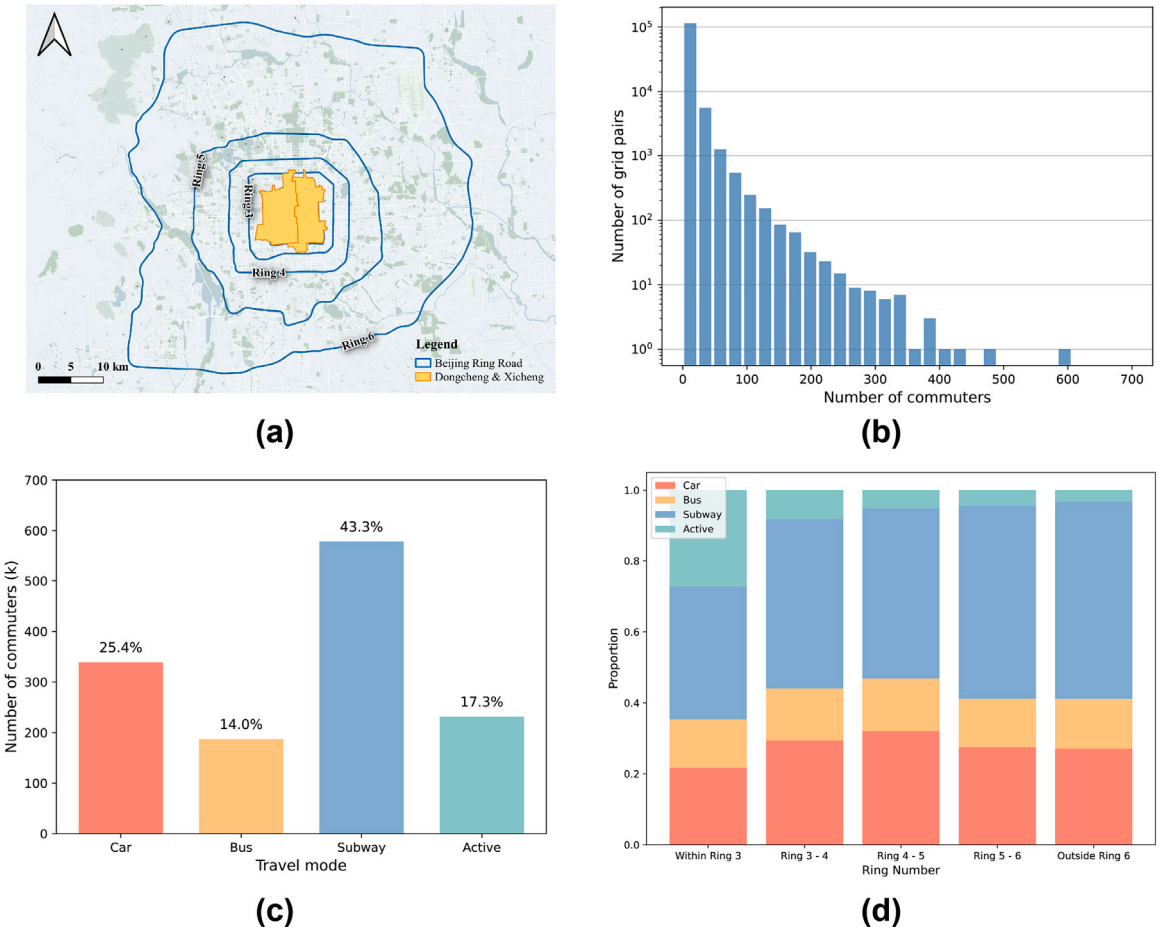


Fig. 1. Study area and data: (a) Study area, (b) The distribution of commuting flow, (c) The distribution of travel modes, (d) Geographical differences in the proportion of travel modes.

Table 1
The field information of commuting flow data.

Field name	Type	Description
OD ID	String	Unique ID of each OD grid pair.
Home/workplace coordinates	Float	The longitude and latitude of home and workplace.
Car	Int	The number of commuters by car.
Bus	Int	The number of commuters by bus.
Subway	Int	The number of commuters by subway.
Active	Int	The number of commuters by active mobility, i.e., walking and cycling.

3.3. Methods

The research framework is illustrated in Fig. 2. Firstly, *data processing* is conducted where we aggregate multisource features (including travel, socioeconomic, infrastructure, and environment) as driving factors and obtain the flow and probability of active travel between any OD pairs. Secondly, *model prediction* is performed, we split the dataset and employ four machine learning models to accomplish the tasks of flow and probability prediction. Finally, *result analysis* is carried out, we select the optimal model and utilize the SHAP method to interpret the model results. Furthermore, targeted policy implications are provided based on the findings.

3.3.1. Data processing

Human travel behavior is undoubtedly influenced by multiple driving factors (Chen et al., 2016; Lenormand et al., 2015). Taking into account the studies of Javaid et al. (2020) and Casali et al. (2022), we have decided to summarize the determinants from four dimensions: travel, socioeconomic, infrastructure, and environment. By integrating these dimensions, it will help us develop a comprehensive understanding of active travel behavior. Additionally, for the dependent variable, we calculate the flow and

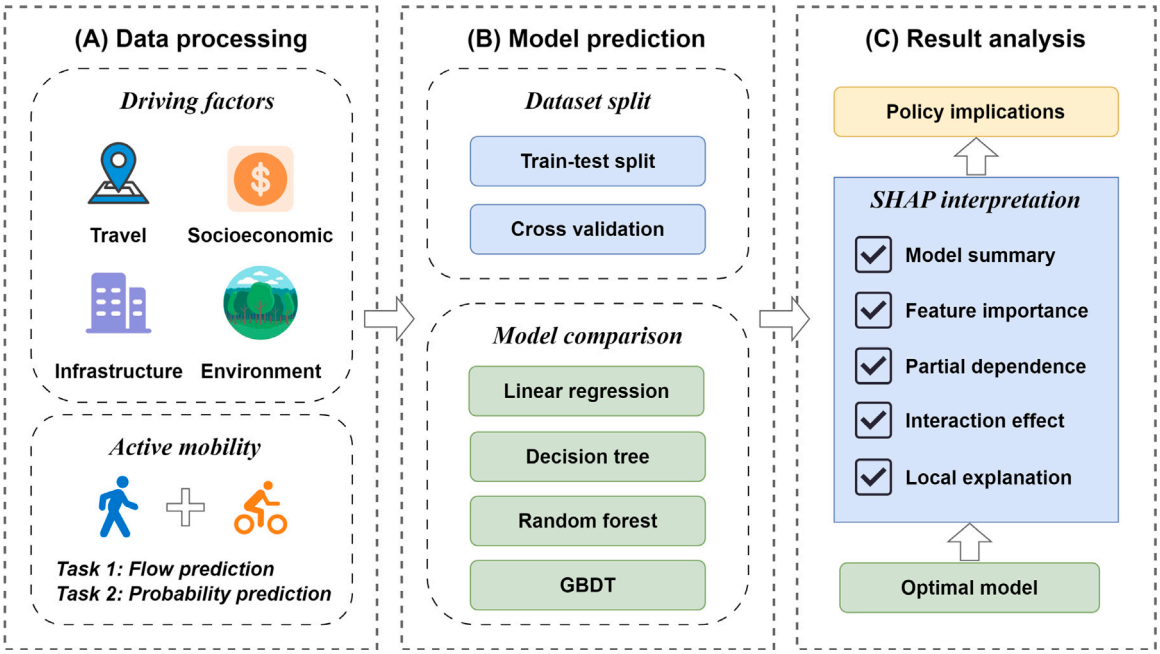


Fig. 2. The framework of the study: (a) Data processing, (b) Model prediction, (c) Result analysis.

probability of active travel between each OD pair based on commuting flow data, which forms the two prediction tasks in this study. Specifically, the flow of active travel is an existing field of the data used (Table 1), and the probability is the proportion of this flow to the total flow of all four travel modes.

For the *travel* dimension, we have calculated the crow-fly distance and road network distance for a given OD pair using the Amap API (https://lbs.amap.com/api/webservice/summary/). Distance is considered a key factor that influences human travel decisions and has been validated in numerous studies (Yin et al., 2023b; Holz-Rau et al., 2014; Scheiner, 2010). Since crow-fly distance and network distance separately offer an intuitive and realistic representation of travel distance, considering them together will provide a more nuanced understanding of how distance affects active travel (Yin et al., 2023a; Frank et al., 2017; Forsyth et al., 2012; Berke et al., 2007). And the Amap API considers the real-world characteristics of local road network, such as walkability and rideability, so the results are more accurate and reliable. Considering the *socioeconomic* dimension, we have calculated features such as population density, average housing price, and the number of Point-of-Interests (POIs) within a 1 km buffer around home and workplace. These features can be seen as key indicators reflecting the local socioeconomic development level (Mirkatouli et al., 2018; Dong et al., 2019; Shi et al., 2020; Huang et al., 2023b). The POIs here mainly retained the commercial-related facilities, while removing the transportation infrastructure such as bus stops, subway stations and parking lots. The *infrastructure* dimension is characterized by two sub-dimensions: the accessibility of other travel modes and the development level of the built environment. For the former, we calculate the distances from home and workplace to the nearest bus stop, subway station, as well as the number of parking lots within a 1 km buffer. These features measure the convenience of bus, subway, private cars, indirectly reflecting competition and substitution effects on active travel (Pisoni et al., 2022). For the latter, we calculate the building area within the grid and the road length within a 1 km buffer. Building and road are the two main elements of urban infrastructure (Huang et al., 2023a), effectively capturing the development level of urban built environment. In terms of the natural *environment*, we use the normalized difference vegetation index (NDVI) and elevation difference between home and workplace, representing the characteristics of the greenspaces and terrain in the natural environment. The greenness has been widely proven to have a significant positive impact on people's willingness to walk or cycle (Gao et al., 2023; Yang et al., 2021), while the terrain's undulation should not be overlooked too (Ospina et al., 2020). Notably, the buffers in this study were calculated from crow-fly distances using QGIS with the 500 m grids as references. The statistics for independent variables are shown in Table 2.

The above data was from the following sources in 2020: Population data was from WorldPop (https://www.worldpop.org/), with a resolution of 100 m. The housing price data was from Lianjia company (https://m.lianjia.com/). POIs data (including bus stops, subway stations, parking lots, etc.) was also obtained using the Amap API. Beijing's building footprint data was provided by Baidu Maps company (https://map.baidu.com/). The road network data was obtained from OpenStreetMap (https://www.openstreetmap.org/). The NDVI and DEM data were downloaded from the National Tibetan Plateau Data Center (https://data.tpdc.ac.cn/) and the Resource and Environmental Science Data Center (https://www.resdc.cn/), Chinese Academy of Sciences, with a resolution of 250 m.

Table 2
Statistics for the independent variables. Note: The suffixes "_o" and "_d" mean the origin and destination grids, "Pop" means the population, "Unit_price" means the average housing price per square meter, "Cnt" is count or the number of, "Dist2bus" and "Dist2sub" mean the distances to the nearest bus stop or subway station.

Category	Variable	Description	Min	Max	Mean	Std
Travel	fly_dist	Crow-fly distance (km)	0.00	71.70	6.63	5.95
	net_dist	Network distance (km)	0.00	83.22	8.15	6.81
Socioeconomic	Pop_mean_o	Population density (O)	9.83	184.23	143.76	34.26
	Pop_mean_d	Population density (D)	0.00	184.23	162.03	16.98
	Unit_price_o	Housing price (O, CNY)	3362.00	147,827.00	80,614.67	29,476.43
	Unit_price_d	Housing price (D, CNY)	3362.00	147 827.00	96,838.98	28,580.88
	Cnt_poi_o	# POIs (O, 1 km buffer)	0.00	1917.00	305.20	210.96
	Cnt_poi_d	# POIs (D, 1 km buffer)	0.00	1917.00	491.95	344.86
Infrastructure	Dist2bus_o	Nearest distance (O, km)	0.00	1.46	0.18	0.12
	Dist2bus_d	Nearest distance (D, km)	0.00	1.09	0.15	0.10
	Dist2sub_o	Nearest distance (O, km)	0.03	36.86	0.71	1.03
	Dist2sub_d	Nearest distance (D, km)	0.03	35.75	0.48	0.45
	Cnt_parkin_o	# parking (O, 1 km buffer)	0.00	680.00	351.38	145.23
	Cnt_parkin_d	# parking (D, 1 km buffer)	0.00	680.00	427.03	109.01
	Building_o	Building area (O, m^2)	3221.46	135,657.21	59,100.12	18,702.26
	Building_d	Building area (D, m^2)	3221.46	135,657.21	62,429.91	14,543.16
	Road_lengt_o	Road length (O, km)	24.72	203.57	115.45	24.57
	Road_lengt_d	Road length (D, km)	7.36	203.57	131.87	20.50
Environment	NDVI_mean_o	NDVI value (O)	0.21	0.73	0.38	0.05
	NDVI_mean_d	NDVI value (D)	0.21	0.79	0.35	0.05
	Delta_DEM	Elevation difference (m)	0.00	446.00	6.75	8.94

3.3.2. Model prediction

In this study, four common machine learning models, namely linear regression (LR), decision tree (DT), random forest (RF), and gradient boosting decision trees (GBDT) (Jordan and Mitchell, 2015), are selected for predicting the flow and probability of active travel. LR assumes a strict linear relationship between the independent variables and the dependent variable, making it less effective when modeling complex nonlinear patterns. DTs are tree-like algorithms that group data by splitting input features. They consist of nodes and edges, where internal nodes represent feature-based decisions, edges represent feature values, and leaf nodes represent predicted values. The algorithm for DTs is as follows:

$$DT(x) = \sum_{j=1}^{J} b_j I\left(x \in R_j\right) \tag{1}$$

where J denotes the number of regions divided by the decision tree, b_j is the predicted value of the region R_j. I is the indicator function, i.e., if $x \in R_j$, then $I = 1$, otherwise, $I = 0$.

RF and GBDT are both tree-based ensemble learning models (Breiman, 2001; Friedman, 2001). The basic idea behind these models is to construct a series of weak learners (i.e., decision trees) and combine them into a strong learner to perform prediction tasks. However, they differ in terms of their ensembling strategies. RF belongs to the Bagging algorithm, while GBDT adopts the Boosting approach. Specifically, RF constructs each decision tree using randomly sampled data subsets and randomly selected feature subsets. The regression result is obtained by averaging the predictions from all decision trees. On the other hand, GBDT employs an iterative process to improve prediction performance by gradually building decision trees. In each iteration, a new regression tree is constructed to correct the errors made by the previous trees to minimize the objective function. Assuming there are M decision trees, the equations for RF and GBDT are as follows:

$$RF(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x), f_m(x) = \sum_{j=1}^{J} b_{jm} I\left(x \in R_{jm}\right) \tag{2}$$

where M is the number of decision trees, f_m is the m-th tree of the RF model. Other variables have similar meanings to the DT model in Eq. (1).

$$GBDT(x) = h_M(x), h_m(x) = h_{m-1}(x) + \sum_{j=1}^{J} b_{jm} I\left(x \in R_{jm}\right) \tag{3}$$

where M is the number of decision trees, h_m is the GBDT model with m trees. Other variables have similar meanings to the DT model in Eq. (1).

LR, DT, RF are implemented using the Scikit-learn package, while the GBDT algorithm utilizes the CatBoost package, all in Python 3.7 version. Firstly, the dataset is randomly divided into a training set (70%) and a testing set (30%). The former is used to learn the model parameters, while the latter is used to evaluate the model's performance. To improve accuracy and avoid overfitting, we further utilize 10-fold cross-validation on the training set to fine-tune the model's parameters and obtain the best model. The mean squared error (MSE) is chosen as the loss function during model training. The evaluation metrics selected are the mean absolute

error (MAE) and coefficient of determination (R^2). The MSE and MAE reflect the deviation between the predicted value and the true value, and the smaller the better. R^2 measures the degree to which the independent variables explain the variation in the dependent variable, with the larger the better. These metrics can be calculated as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \tag{6}$$

where n is the number of samples, $\hat{y}_i$ and y_i are the predicted and true values of the sample i, and $\overline{y}$ is the average of all samples.

3.3.3. Result analysis

Based on the comparison of the above models, we select the optimal model, namely GBDT, for further analysis. A local explanation method called SHAP (SHapley Additive exPlanations) is employed to analyze the local impact of each feature in individual predictions. SHAP is a game-theoretic approach originally used to calculate the contribution of each player in achieving a collective goal (Shapley, 1953). Subsequently, this method is introduced to the field of machine learning to quantify the contribution of each feature to the model's output (Lundberg and Lee, 2017; Lundberg et al., 2020). Simply put, the SHAP value of each feature represents the average marginal contribution it makes when participating in different combinations with other variables. The SHAP value is defined as follows:

$$\phi(X_k) = \sum_{S\subseteq\mathcal{X}\setminus X_k}\frac{|S|!\cdot(|\mathcal{X}| - |S| - 1)!}{|\mathcal{X}|!}\left(f\left(S\cup\left\{X_k\right\}\right) - f(S)\right) \tag{7}$$

where $\phi(X_k)$ is the SHAP value of the k-th feature X_k. $\mathcal{X}$ is the complete set of all features and S is the subset of $\mathcal{X}$ excluding X_k. $|\cdot|$ is the cardinality of a set, $f(S)$ is the model prediction with the features in S as the input. Furthermore, each model prediction is broken down into multiple additive terms ($\phi(\cdot)$) as shown below, each of which represents the impact of a certain feature on the model outcome.

$$\hat{y}_i = \phi_0 + \sum_{k=1}^{|\mathcal{X}|}\phi(X_{ki}) \tag{8}$$

where $\hat{y}_i$ is the predicted value for the i-th sample, $\phi_0 = E(\hat{y})$ is the average of all predictions, $|\mathcal{X}|$ is the number of features, and $\phi(X_{ki})$ is the SHAP value of k-th feature for sample i. Therefore, the SHAP value of each feature causes the actual prediction to shift from the average prediction. In addition, the overall importance of each feature can be measured by averaging absolute SHAP values of this feature in all samples, as shown below.

$$I(X_k) = \frac{1}{n}\sum_{i=1}^{n}\left|\phi_i(X_k)\right| \tag{9}$$

where $I(X_k)$ is the importance of the feature X_k, $\phi_i(X_k)$ is the SHAP value of the feature X_k for the i-th prediction, and n is the number of samples.

In this study, SHAP values can measure the contribution of each feature to the flow and probability prediction of active travel between individual OD pairs. A positive SHAP value indicates a positive impact of the relevant feature on active travel. The magnitude of the SHAP value reflects the extent of its influence on the model prediction. Therefore, the sign and magnitude of SHAP values can be utilized to analyze the local effects of determinants for active travel. We employ the TreeExplainer method from the SHAP package (Python 3.7) to interpret the GBDT model. To uncover the patterns of active travel, we not only use global measures of the overall importance of multisource features, but also locally interpret their nonlinear and interaction effects. Furthermore, we provide targeted policy implications for active travel-oriented urban planning.

4. Results

4.1. Predictions of active travel flow and probability

In this study, a comparison of four machine learning models is shown in Table 3, with GBDT being the best-performing model. LR performs poorly on both prediction tasks, indicating a strong nonlinear relationship between active travel and determining factors. DT, as a non-linear model, achieves better results than linear regression. RF and GBDT, as ensemble learning methods, unsurprisingly outperform other single-learner models. Specifically, GBDT achieves an MAE of 2.08 and an R^2 of 0.72 for flow prediction, with 0.07 and 0.73 for probability prediction, respectively. Considering the superior performance of GBDT on both tasks, it will be used for further analysis in subsequent studies.

The predictions of GBDT implemented by CatBoost for both tasks are shown in Figs. 3 and 4. GBDT performs well in flow prediction, as shown in Fig. 3a. When the predicted values are less than 100, the residuals are small, while the predictions become

Table 3
The comparison results of four models in predicting flow and probability.

Model	MAE	R^2
Flow prediction		
GBDT	2.08	0.72
RF	2.36	0.61
DT	3.12	0.17
LR	5.27	0.07
Probability prediction		
GBDT	0.07	0.73
RF	0.07	0.71
DT	0.10	0.42
LR	0.12	0.16

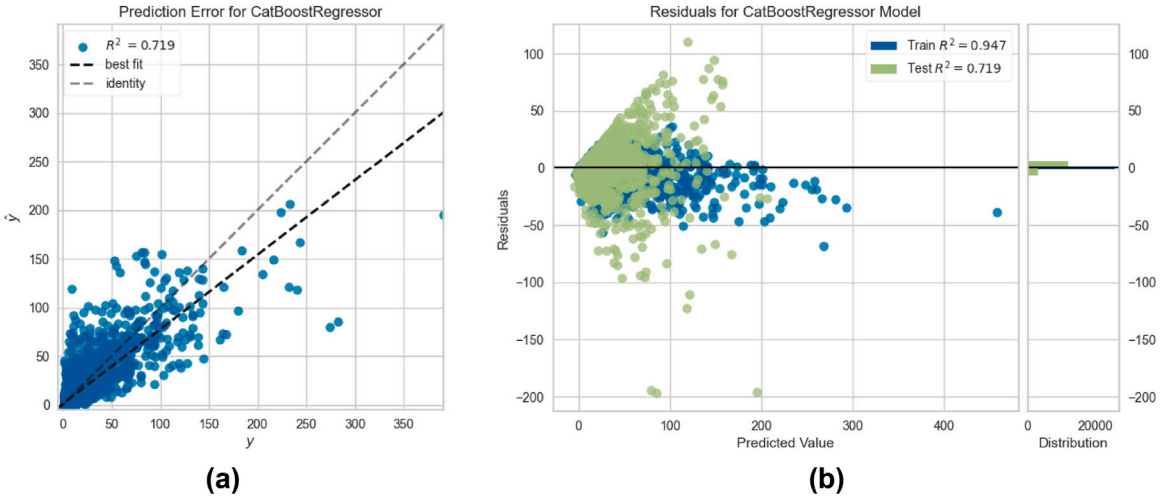


Fig. 3. The results of flow prediction: (a) Fitting performance. (b) Prediction residuals.

more unstable for higher values, as depicted in Fig. 3b. This indicates that the model is more accurate in predicting low-intensity flows. GBDT also demonstrates excellent performance in probability prediction, as shown in Fig. 4a. Predicted values from tree-based models tend to fall within the range of labels in the training set (Jordan and Mitchell, 2015; Breiman, 2001). Since probabilities are bounded between 0 and 1, a clear truncation of residuals can be observed in Fig. 4b. For instance, a predicted value of 0.8 as the label would have a residual greater than −0.2 but less than 0.8. It is worth noting that the residuals tend to be greater than 0 (Fig. 4b), indicating the possible overestimation of the model.

4.2. Relative importance of multisource features

The absolute SHAP values quantify the relative importance of variables for the model. Table 4 presents the relative contributions of four categories of features in flow and probability prediction. Firstly, distance-based travel features have the largest contributions to both tasks, with an average impact exceeding 50%. This indicates that distance is the most dominant variable affecting active travel intensity and willingness, which aligns with previous research findings (Kaplan et al., 2016; Guo et al., 2023; Wu et al., 2021). The impacts of socioeconomic and infrastructure characteristics rank second after travel characteristics. In flow prediction, their average contributions are 22.21% and 20.37%, respectively. In probability prediction, their contributions are 15.40% and 21.80%, respectively. While the environment dimension has been proven to be a key factor influencing active travel in many studies (Panter et al., 2008; Gao et al., 2023), it has the smallest impact when other factors are controlled, with average impacts of 3.38% and 5.94%, respectively.

Figs. 5 and 6 depict the relative importance of the top ten variables in flow and probability prediction. On the left, the variables are sorted in descending order by the global importance, while on the right, the impacts of these features on active travel for each OD pair are displayed. For flow prediction (Fig. 5), crow-fly distance and road network distance are the most dominant variables, collectively changing the predicted values by an average of 5.52. Additionally, it can be observed that the number of POIs at the workplace is the third most important variable. When it is large (red), it has a positive impact on flow prediction, and vice versa. Furthermore, socioeconomic variables (i.e., housing prices at the home and workplace, the number of POIs at the home, population density at the workplace) and infrastructure variables (distance from home to the nearest subway station, building area, and parking

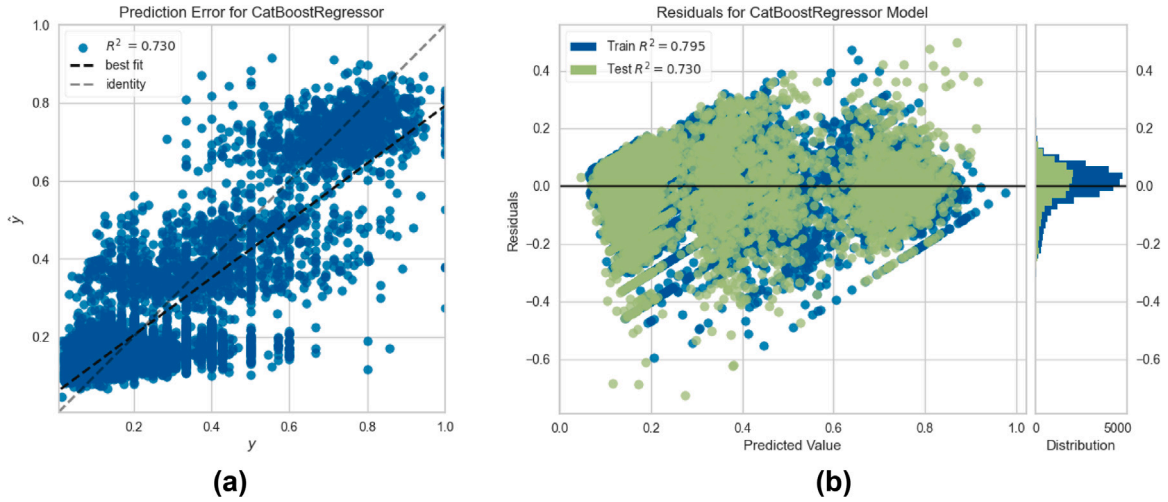


Fig. 4. The results of probability prediction: (a) Fitting performance. (b) Prediction residuals.

Table 4
The contribution of each category of features in predicting flow and probability.

Feature	Mean (|SHAP|)	Ratio (|SHAP|)
Flow prediction		
Travel	5.52	54.04%
Socioeconomic	2.27	22.21%
Infrastructure	2.08	20.37%
Environment	0.35	3.38%
Probability prediction		
Travel	0.09	56.86%
Socioeconomic	0.03	15.40%
Infrastructure	0.04	21.80%
Environment	0.01	5.94%

lot density) all contribute significantly to flow prediction and remain within the top ten variables. For probability prediction (Fig. 6), the impact of crow-fly distance stands out prominently. Extremely low values of distance (blue) substantially increase residents' willingness to choose active travel. Apart from distance, socioeconomic variables (i.e., population, housing prices, and the number of POIs), infrastructure variables (i.e., distance to the nearest subway station, parking lot density, road length), and environmental variables (i.e., NDVI) all have certain impact on the model, but their effects are more complex. Furthermore, by comparing flow prediction and probability prediction, we can observe some differences in the determinants of the two tasks. Firstly, the top ten contributing variables are distinct for the two tasks (left panels of Figs. 5 and 6). Secondly, some variables may have opposite local effects for the predictions (right panels of Figs. 5 and 6). For example, a low value in *Cnt_poi_d* reduces the flow but increases the probability for active travel. These differences highlight the necessity of separately modeling flow and probability prediction for active travel behavior. The variations in these local effects will be further explored and analyzed in the next section.

4.3. Nonlinear and interaction effects of multisource features

Compared to traditional models that reflect global effects, the SHAP-based dependency plots provide a clear revelation of the local effects of each feature on the model across all samples (Li, 2023; Xiao et al., 2021). In other words, by controlling for other variables, we can analyze the marginal effect between each variable and the output at a local level or for each individual sample. Furthermore, SHAP offers a tool for computing interaction effects, allowing us to decompose the impact of a feature in each sample into its interaction effects with other features (Lundberg et al., 2020). The SHAP interaction value has a similar definition to the standard SHAP value, except that the object changes from a single feature to a feature pair when measuring the marginal contribution (Lundberg et al., 2020; Fujimoto et al., 2006). This enables us not only to calculate the impact of a single feature on each prediction, but also to quantify the interaction effects between two features on the model. Thus, the SHAP-based dependency plots will facilitate our analysis of the intricate patterns between various characteristics and active travel.

Figs. 7 and 8 depict partial dependence plots for selected representative features on the prediction of active travel flow and probability. The vertical dispersion of the SHAP values in different colors represents the strength of the interaction effects (Lundberg et al., 2020). And the greater the dispersion, the stronger the interaction. These features are chosen to represent four major categories

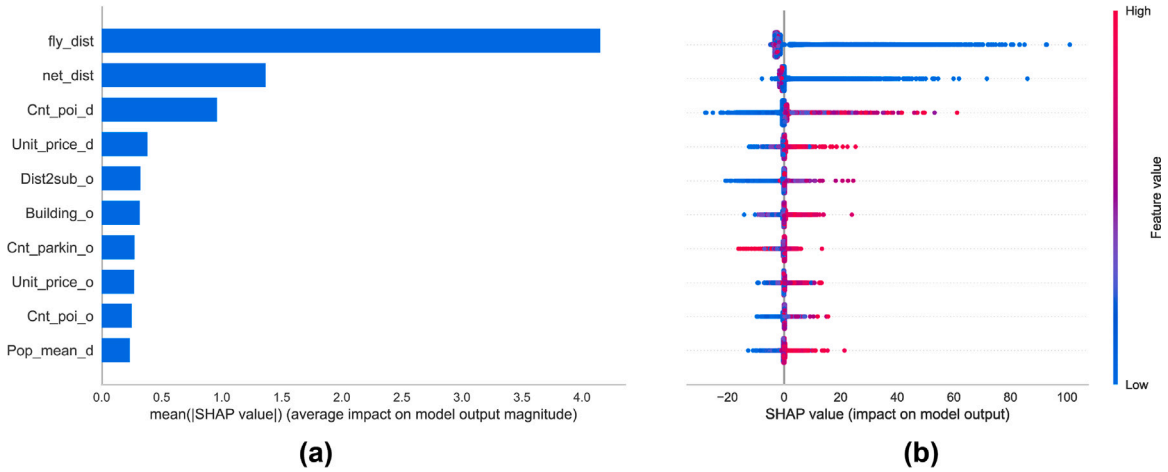


Fig. 5. SHAP based global analysis for flow prediction: (a) Relative importance of main features. (b) SHAP distribution of main features, each point represents an OD pair, the x-axis denotes the SHAP value, the y-axis is associated with the features on the left, and the points are colored by the feature values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

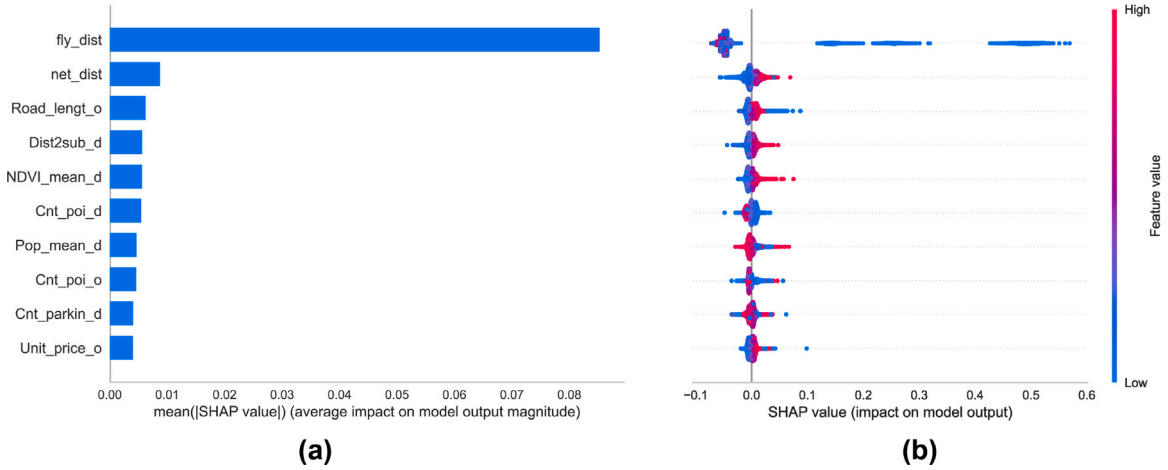


Fig. 6. SHAP based global analysis for probability prediction: (a) Relative importance of main features. (b) SHAP distribution of main features, each point represents an OD pair, the x-axis denotes the SHAP value, the y-axis is associated with the features on the left, and the points are colored by the feature values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and also appear in the top ten contributing features list (Figs. 5 and 6). For flow prediction, we select *fly_dist*, *Unit_price_o*, *Cnt_poi_d*, *Dist2sub_o*, *Building_o*, and *NDVI_mean_o* for nonlinear and interaction effect analysis. Firstly, for *fly_dist* (Fig. 7a), we observe a distinct truncation phenomenon. When the crow-fly distance is less than about 2 km, it has a significant positive impact on flow prediction. However, as the crow-fly distance increases, the impact rapidly weakens. When the crow-fly distance exceeds 2 km, the SHAP value becomes a weak negative value and remains almost unchanged. This reflects the high competitiveness of walking and cycling for short-distance travel (Ji et al., 2022; Rahul and Verma, 2014). Regarding the remaining five categories, we find that road network distance (*net_dist*) has the strongest interaction effect with them. Moreover, when the road network distance is too large (e.g., >10 km), the impact of these features on the model approaches zero because residents rarely choose active travel for long-distance trips. Specifically, the impact of *Unit_price_o* on active travel volume exhibits an increasing relationship (Fig. 7b). When *Unit_price_o* is below 80,000 RMB, it has a negative impact on the traffic volume of active travel. However, when it exceed this threshold, it shows a positive impact. This indicates that the number of high-income individuals choosing active travel is increasing (Buehler and Pucher, 2017; Foster et al., 2018). *Cnt_poi_d* has a negative impact on the model when it is below 500, but a positive impact when it exceeds this threshold (Fig. 7c). A higher number of workplace POIs implies a prosperous economy with more job opportunities, resulting in more commuters and a corresponding increase in the number of individuals choosing active travel. When *Dist2sub_o* increases, the number of active travelers sharply increases (Fig. 7d). When this distance is less than 1 km, it has a negative impact on flow prediction, whereas distance greater than 1 km exhibits a positive impact. This reflects the competitive effect of the subway for active travel (Fung et al., 2021; Ettema et al., 2016). The more convenient the subway is, the fewer individuals choose active travel. As

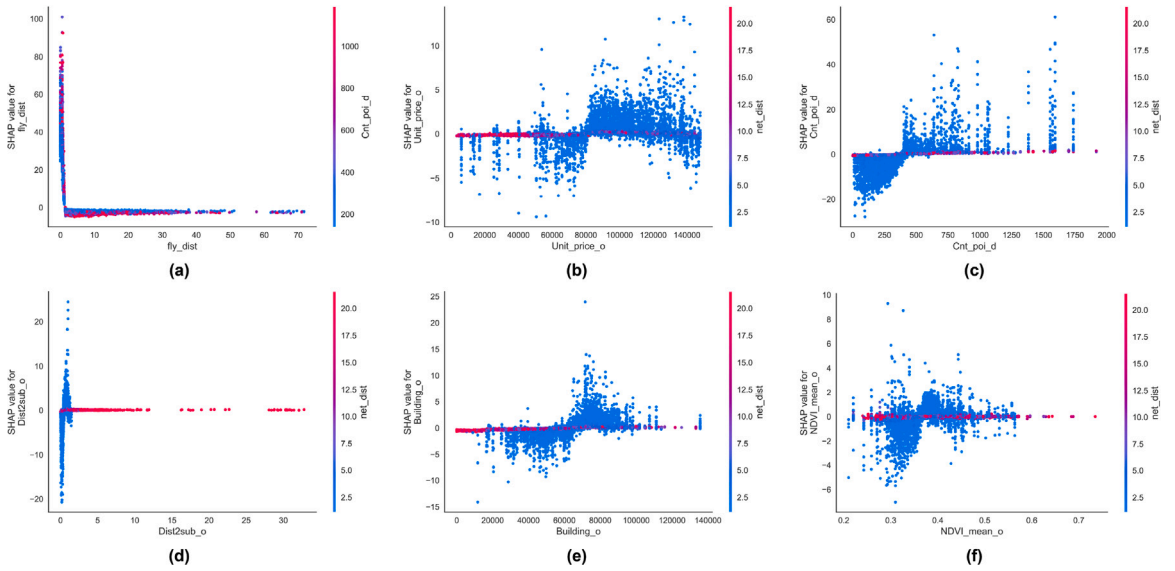


Fig. 7. SHAP dependence plots of six representative features for flow prediction. Each point represents an OD pair, the x-axis denotes the variable value, the y-axis denotes the SHAP value for the variable, and the color is encoded by the feature with the strongest interaction effect. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

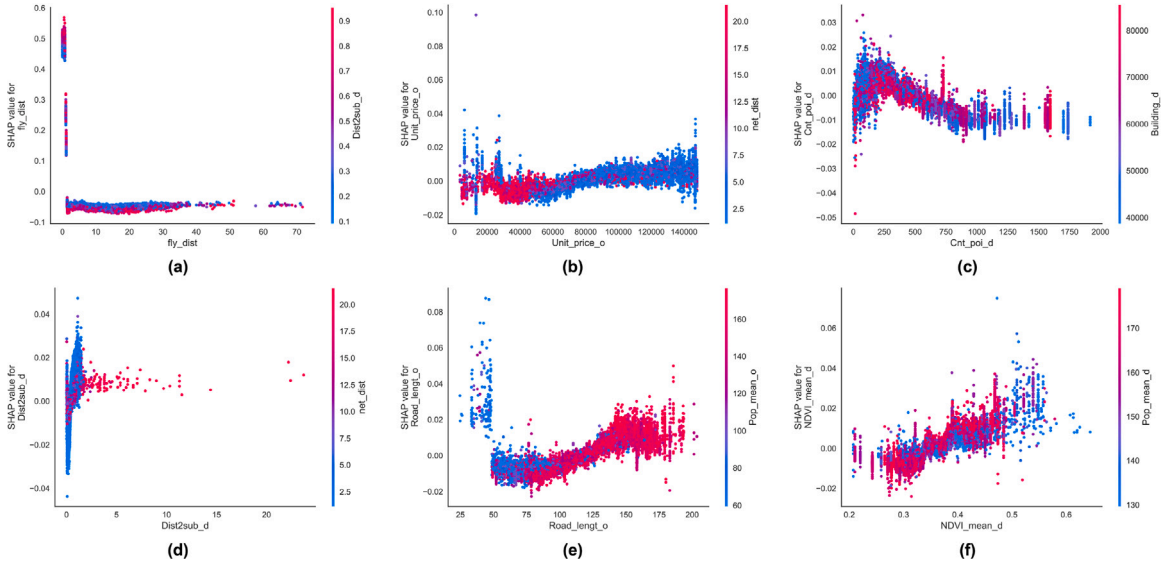


Fig. 8. SHAP dependence plots of six representative features for probability prediction. Each point represents an OD pair, the x-axis denotes the variable value, the y-axis denotes the SHAP value for the variable, and the color is encoded by the feature with the strongest interaction effect. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Building_o increases, the number of individuals choosing active travel also increases, with a notable threshold at 70,000 square meters (Fig. 7e). This can be attributed to a larger commuter base in areas with more residential buildings, resulting in higher number of active travelers. Overall, the impact of *NDVI_mean_o* on the number of active travelers shows an increasing trend, with a shift from negative to positive when the NDVI exceeds 0.35 (Fig. 7f). Areas with higher NDVI values typically have more greenspaces, such as parks and streets with abundant vegetation. These green spaces may provide better environments and conditions, thus attracting more individuals to choose active travel (Gao et al., 2023; Panter et al., 2008).

For the probability prediction of active travel, we have selected the following features for nonlinear and interaction effect analysis: *fly_dist*, *Unit_price_o*, *Cnt_poi_d*, *Dist2sub_d*, *Road_lengt_o*, and *NDVI_mean_d*. The impact of *fly_dist* exhibits a similar pattern to flow prediction, displaying a truncation phenomenon around 2 km. This also indicates a short-distance preference of active travel (Fig. 8a). Regarding *Unit_price_o*, the strongest interaction is observed with road network distance (*net_dist*). Samples with longer road

network distances (>10 km) tend to concentrate in areas with lower housing prices (<60,000 RMB) (red points in Fig. 8b), reflecting a more pronounced jobs-housing separation among low-income populations (Ta et al., 2017; Cervero, 1989). In such long-distance commutes, the drawbacks of long time, low convenience, and inadequate comfort may reduce the willingness of low-income people to choose active travel. However, for short-distance travel (blue points), the local impact of *Unit_price_o* gradually increases with their rise, indicating a growing acceptance of active travel among high-income individuals. Walking and cycling, as healthy modes of transportation that also offer environmental benefits, are gaining popularity among the affluent class (Burbidge and Goulias, 2009). When *Cnt_poi_d* is less than 500, it positively influences the likelihood of choosing active travel. However, when it exceeds 500, the effect becomes negative (Fig. 8c). The willingness to choose walking or cycling decreases with the growth of workplace POI density, while the number of active travelers increases (Fig. 7c). A larger number of POIs indicates a more developed economy, with both an increase in job opportunities and improved infrastructure. While the former attracts a significant number of commuters, leading to an increase in active travelers, the latter prompts the working class to opt for public or private transportation due to factors such as time, convenience, and comfort, thereby reducing their inclination towards active travel (Javaid et al., 2020). *Dist2sub_d* exhibits the strongest interaction with *net_dist* (Fig. 8d). When the road network distance is excessively large (about >10 km, red points), *Dist2sub_d* has negligible impact on the model. For short-distance travel (blue points), if *Dist2sub_d* is less than about 1 km, it negatively affects the probability of choosing walking or cycling, as the convenience of the subway leads people to prefer it as the travel mode. However, when *Dist2sub_d* exceeds 1 km, a decrease in subway accessibility actually promotes active travel popularity. *Road_lengt_o* demonstrates the strongest interaction with *Pop_mean_o* (Fig. 8e). When the population density is below 100 and the road network length is less than 50 km, a synergistic effect is clearly observed in the model. This may be attributed to walking or cycling being the conventional modes of travel in sparsely populated areas with weak infrastructure (Burbidge and Goulias, 2009; Cook et al., 2022). However, as the road network length surpasses 50 km, its impact gradually shifts from negative to positive. This reflects the improvement in walking or cycling conditions due to enhancements in road infrastructure, such as the addition of bicycle lanes, pedestrian pathways, and improved road connectivity, leading to an increased probability of active travel. When *NDVI_mean_d* is too low (<0.35), it negatively impacts the choice of active travel. However, when it exceeds 0.35, the effect becomes positive (Fig. 8f). This finding is consistent with the results of the flow prediction task (Fig. 7f) and also reflects the positive influence of green spaces on walking or cycling.

4.4. Policy implications for traffic optimization using local SHAP analysis

SHAP analysis offers insights into the determinants of individual travel behavior, enabling a focused examination of specific trips. Such analysis aids in understanding the local effects of active travel behavior for each OD pair, thereby facilitating targeted improvements and optimizations for active travel-oriented transportation. OD trips with high demand but low probability signify substantial travel needs with limited inclination towards active travel, warranting higher priority in optimization efforts. The occurrence of such instances can stem from multifaceted factors, including long distances, inadequate environmental comfort, competition from alternative transport modes, and more (Javaid et al., 2020; Cook et al., 2022). Fortunately, SHAP provides a sample-level tool for such localized analysis. Thus, we have randomly selected three OD trips meeting the criteria of high volume (>10) and low probability (<0.1) for further analysis. These OD trips serve as the cases using local SHAP analysis for policy implications and can be applied to other samples. Fig. 9a depicts the spatial distribution of these three OD pairs, while Fig. 9b–d present waterfall plots illustrating the local explanations. The x-axis represents the SHAP values for each feature, while the y-axis denotes the corresponding feature and its values. And the features are sorted from top to bottom by their SHAP values, with the bottom row representing the total effect of the remaining features.

For OD #1, the origin is located in Yongtieyuan community near Beijing South Railway Station, while the destination is in the north of Taoranting Park. Firstly, the remarkably short crow-fly distance contributes to a predicted value that is 0.15 higher than the mean, making it the most influential factor. However, the predicted probability of active travel for this OD trip is 0.337, whereas the actual value is below 0.1. This discrepancy may indicate an overestimation of the positive impact of short distances by the model, as reflected in the negative effect of road network distance. Additionally, features such as workplace housing price, workplace subway accessibility, and residential building area exert minimal influence on the model. Noting that this OD trip is a short distance (*net_dist* = 1.649 km) and the accessibility of other travel modes is low (*Dist2sub_d* > 1 km), it is more feasible to improve the convenience of active travel. Therefore, improving road connectivity, such as adding bike lanes or walking lanes on certain key road segments, could become a crucial aspect of transportation optimization. Lastly, this sample highlights the occasional distortion of the model and emphasizes the need for specific analyses in conjunction with real-world considerations.

For OD #2, the origin is Guangqumen Community, and the destination is Wangfujing Commercial Street. Compared to OD #1, this trip has a longer crow-fly distance and a road network distance exceeding 2.5 km, which negatively impacts the willingness for active travel due to the long distance (as shown in Fig. 8a), reducing the probability by approximately 0.06. Excessive POIs and population density at both the origin and destination may contribute to increased traffic congestion, thereby reducing the likelihood of walking or cycling. In such cases, a decentralized urban development strategy can help alleviate traffic pressure and mitigate pedestrian congestion, thereby enhancing the willingness for active travel. The low NDVI value (0.278) indicates limited green coverage, which diminishes the comfort of walking or cycling. The government can promote urban greening initiatives by strategically increasing public greenspaces along this trip to enhance the attractiveness of green traveling (Gao et al., 2023). The notable proximity of the workplace to the nearest subway station (*Dist2sub_d* = 0.293 km) makes residents more inclined to use subway services. A well-utilized subway system is desirable, and in such cases, the government can improve walking and cycling facilities around subway stations, such as providing convenient bicycle parking, enhancing sidewalks, and ensuring adequate

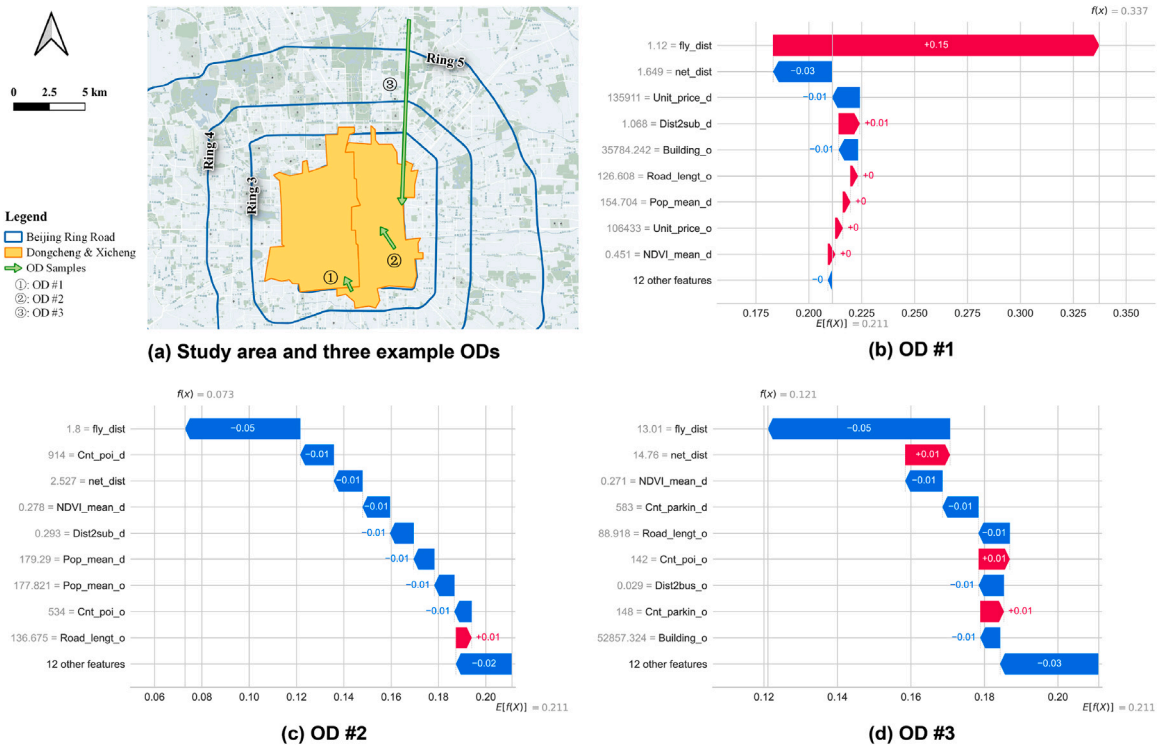


Fig. 9. Explanation of three example ODs based on local SHAP analysis: (a) Study area and three example ODs, (b) Explanation for OD #1, (c) Explanation for OD #2, (d) Explanation for OD #3.

transportation signage, to enhance the multimodal travel experience to connect to subway. Additionally, a well-developed road network (*Road_lengt_o* = 136.675 km) provides sufficient infrastructure for walking and cycling, and ensuring that pedestrian and bicycle paths are seamlessly connected to the road system is crucial for promoting walking and cycling among residents.

For OD #3, the origin is Beiyuan, a large community, and the destination is Chaoyangmen Bridge near the CBD. Clearly, the long commuting distance is the most significant negative factor affecting the willingness for active travel, reducing the probability by approximately 0.04 overall. Despite the long commuting distance, there are still many residents who choose active travel modes. Therefore, it becomes crucial to alleviate this portion of travel demand. The low NDVI value (0.271) contributes to a decrease in residents' willingness to walk or cycle. Implementing measures such as increasing parks and other greenspaces along the trip can help change this situation. The insufficient parking lots at the residential area (*Cnt_parkin_o* = 148) offset the impact of abundant parking lots at the workplace (*Cnt_parkin_d* = 583), making the influence of private transportation on active travel relatively weak. Additionally, many of the top-ranking features are related to the residential area, indicating that the focus of optimization for this OD trip should be on the origin location. Specifically, the number of POIs (142), road length (88.918 km), and building area (52,857.324 m^2) in the residential area are significantly lower than the mean, indicating a lack of job opportunities and infrastructure in that location. Therefore, promoting mixed-use development for the community to achieve a better jobs-housing balance, and further improving convenience facilities to alleviate the pressure for long-distance commuting could be important measures. It is worth noting that the residential bus accessibility is high (*Dist2bus_o* = 0.029 km), indicating that buses may already be carrying a significant number of commuters. In this case, improving subway accessibility could be an excellent option. With fixed subway station locations, enhancing the connection experience between walking/cycling and the subway becomes an effective solution. For example, the government can increase convenient bicycle parking areas, construct well-connected pedestrian roads and footbridges, strengthen the sharing of traffic information and navigation guidance, encourage integration between shared mobility (such as shared bicycles, electric scooters, etc.) and the subway, and provide diverse last-mile solutions.

5. Discussions

This study holds both theoretical and practical significance. Theoretically, it endeavors to incorporate large-scale and high-resolution travel big data to investigate active travel behavior (Fig. 1). On one hand, it serves as a supplement to traditional small-sample data like questionnaires or surveys, while on the other hand, it empowers data-driven models to fully leverage their modeling capabilities. Furthermore, this study decomposes the active travel prediction task into two sub-tasks: flow prediction and probability prediction. Through experiments, it reveals numerous differences in the influencing factors between the two sub-tasks. For instance, the top ten contributing features are different (Figs. 5 and 6). For another instance, a lower value of workplace

POIs decreases the number of active travelers but increases their willingness (Figs. 7 and 8). This demonstrates the necessity of decomposing the prediction task and contributes to a more comprehensive analysis of active travel behavior. Lastly, this study introduces an explainable machine learning method (SHAP), to thoroughly explore the determinants of active travel. It uncovers evident nonlinear relationships (Figs. 7 and 8), which complement traditional generalized linear models. Moreover, it analyzes the local effects of each feature at the individual level (Fig. 9), enabling targeted adjustments and optimizations for each OD trip.

In practice, the local analysis tools provided by SHAP allow us to investigate the influencing factors of each OD trip, identify their shortcomings, and make targeted improvements. The importance of different features varies within the same OD sample, and the impact of the same feature differs across different OD samples. For certain samples, enhancing road connectivity and improving bus capacity may be of paramount importance for transportation optimization (Fig. 9b). In some cases, a decentralized urban layout, abundant urban greenery, and convenient subway connections can create a transportation system that is favorable to active travel (Fig. 9c). In other instances, policy interventions should focus on enhancing the connection experience between walking/cycling and the subway, and provide diversified last-mile solutions through various measures (Fig. 9d). Hence, policymakers can utilize the optimization methods proposed in this study to devise effective improvement policies at the individual OD level, thereby promoting sustainable urban transportation systems.

Nevertheless, this study still has certain limitations. Firstly, the current set of driving factors considered is not exhaustive, and future research could explore additional factors such as travel costs, consumption levels, the number of bicycle lanes, the level of bus services, and eye-level urban greenery. Secondly, there is room for improvement in the prediction models, and the introduction of deep learning models like graph neural networks could enhance prediction accuracy. Thirdly, the buffer construction does not consider the structure of the actual road network, such as the connectivity and accessibility of walking and cycling. A more complete active travel network needs to be considered in the future. Fourthly, a 500 m grid is used as the study unit, which may affect the calculation of distances and the derivative results, especially for ODs over short distances. In addition, the impact of crow-fly distance on active travel is significantly higher than that of network distance, and the reasons for this need to be explored in the future. Lastly, although this study utilizes big trip data from over a million residents, the research area is limited and may not represent the entire city of Beijing. Future studies should expand the scope to achieve more representative experimental results.

6. Conclusion

Active travel, i.e., walking and cycling, is widely acknowledged as a sustainable and eco-friendly mode of travel. To comprehensively investigate active travel behavior, this study leverages big trip data and machine learning models to accurately predict both the flow and probability, achieving R^2 values of 0.72 and 0.73, respectively. By introducing an explainable artificial intelligence method (SHAP), we conduct a focused analysis on the local effects of multisource characteristics on active travel, such as travel, socioeconomic, infrastructure, and environment. Our findings reveal that distance stands as the most influential factor, contributing to over 50% of the observed effects on active travel. Moreover, features such as housing prices, POI density, distance to the nearest subway station, building area or road length, and urban greenery exhibit notable nonlinear and interaction effects on active travel, exhibiting significant threshold phenomena. Through localized interpretability analysis, we unveil the individual variations in the impact of these features on active travel, providing valuable insights for adjusting and optimizing transportation planning. The results of this study offer policymakers a profound understanding of active travel behavior, serving as a basis for formulating sustainable and effective transportation strategies. By promoting active travel, these strategies can ultimately enhance individual well-being, social welfare, and urban environments.

CRediT authorship contribution statement

Ganmin Yin: Conceptualization, Methodology, Visualization, Writing – original draft. **Zhou Huang:** Methodology, Writing – review & editing. **Chen Fu:** Data curation, Validation. **Shuliang Ren:** Validation, Visualization. **Yi Bao:** Validation, Visualization. **Xiaolei Ma:** Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

We acknowledge the financial support from the National Natural Science Foundation of China (42271471, U2344216). We thank the editors and reviewers for their suggestions to improve the quality of this paper. We are also grateful to Mr. Wang Han, Mr. Zheng Jiangpeng and Dr. Dong Quanhua for their help.

References

Beijing Municipal Bureau of Statistics, 2023. Beijing 2022 national economic and social development statistical communiqué. http://tjj.beijing.gov.cn/bwtt_31461/202303/t20230321_2940949.html.

Berke, E.M., Koepsell, T.D., Moudon, A.V., Hoskins, R.E., Larson, E.B., 2007. Association of the built environment with physical activity and obesity in older persons. Amer. J. Public Health 97 (3), 486–492.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Buehler, R., Pucher, J., 2017. Trends in walking and cycling safety: recent evidence from high-income countries, with a focus on the United States and Germany. Amer. J. Public Health 107 (2), 281–287.

Burbidge, S., Goulias, K., 2009. Active travel behavior. Transp. Lett. 1 (2), 147–167.

Cao, M., Cai, B., Ma, S., Lü, G., Chen, M., 2019. Analysis of the cycling flow between origin and destination for dockless shared bicycles based on singular value decomposition. ISPRS Int. J. Geo-Inf. 8 (12), 573.

Carlson, J.A., Saelens, B.E., Kerr, J., Schipperijn, J., Conway, T.L., Frank, L.D., Chapman, J.E., Glanz, K., Cain, K.L., Sallis, J.F., 2015. Association between neighborhood walkability and GPS-measured walking, bicycling and vehicle time in adolescents. Health Place 32, 1–7.

Casali, Y., Aydin, N.Y., Comes, T., 2022. Machine learning for spatial analyses in urban areas: a scoping review. Sustain. Cities Soc. 85, 104050.

Cervero, R., 1989. Jobs-housing balancing and regional mobility. J. Amer. Plan. Assoc. 55 (2), 136–150.

Chai, J., Lu, Q.-Y., Wang, S.-Y., Lai, K.K., 2016. Analysis of road transportation energy consumption demand in China. Transp. Res. D 48, 112–124.

Chai, D., Wang, L., Yang, Q., 2018. Bike flow prediction with multi-graph convolutional networks. In: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 397–400.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. C 68, 285–299.

Cook, S., Stevenson, L., Aldred, R., Kendall, M., Cohen, T., 2022. More than walking and cycling: What is 'active travel'? Transp. Policy 126, 151–161.

Dong, L., Ratti, C., Zheng, S., 2019. Predicting neighborhoods' socioeconomic attributes using restaurant data. Proc. Natl. Acad. Sci. 116 (31), 15447–15452.

Erhardt, G.D., Roy, S., Cooper, D., Sana, B., Chen, M., Castiglione, J., 2019. Do transportation network companies decrease or increase congestion? Sci. Adv. 5 (5), eaau2670.

Ettema, D., Friman, M., Gärling, T., Olsson, L.E., 2016. Travel mode use, travel mode shift and subjective well-being: Overview of theories, empirical findings and policy implications. In: Mobility, Sociability and Well-Being of Urban Living. Springer, pp. 129–150.

Ferrari, G., Oliveira Werneck, A., Rodrigues da Silva, D., Kovalskys, I., Gómez, G., Rigotti, A., Yadira Cortés Sanabria, L., García, M.C.Y., Pareja, R.G., Herrera-Cuenca, M., et al., 2020. Association between perceived neighborhood built environment and walking and cycling for transport among inhabitants from Latin America: The ELANS study. Int. J. Environ. Res. Public Health 17 (18), 6858.

Forsyth, A., Van Riper, D., Larson, N., Wall, M., Neumark-Sztainer, D., 2012. Creating a replicable, valid cross-platform buffering technique: the sausage network buffer for measuring food and physical activity built environments. Int. J. Health Geogr. 11 (1), 1–9.

Foster, C., Kelly, P., Reid, H.A., Roberts, N., Murtagh, E.M., Humphreys, D.K., Panter, J., Milton, K., 2018. What works to promote walking at the population level? A systematic review. Br. J. Sports Med. 52 (12), 807–812.

Frank, L.D., Fox, E.H., Ulmer, J.M., Chapman, J.E., Braun, L.M., 2022. Quantifying the health benefits of transit-oriented development: Creation and application of the san diego public health assessment model (SD-PHAM). Transp. Policy 115, 14–26.

Frank, L.D., Fox, E.H., Ulmer, J.M., Chapman, J.E., Kershaw, S.E., Sallis, J.F., Conway, T.L., Cerin, E., Cain, K.L., Adams, M.A., et al., 2017. International comparison of observation-specific spatial buffers: maximizing the ability to estimate physical activity. Int. J. Health Geogr. 16, 1–13.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 1189–1232.

Fu, C., Huang, Z., Scheuer, B., Lin, J., Zhang, Y., 2023. Integration of dockless bike-sharing and metro: Prediction and explanation at origin-destination level. Sustainable Cities Soc. 104906.

Fujimoto, K., Kojadinovic, I., Marichal, J.-L., 2006. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. Games Econom. Behav. 55 (1), 72–99.

Fung, C.M., McArthur, D.P., Hong, J., 2021. Examining the effects of a temporary subway closure on cycling in glasgow using bike-sharing data. Travel Behav. Soc. 25, 62–77.

Gao, J., Ma, S., Wang, L., Shuai, L., Du, H., 2023. Does greenness bring more green travelling? Evidence from free-floating bike-sharing in Beijing. J. Transp. Geogr. 109, 103586.

Guo, L., Yang, S., Peng, Y., Yuan, M., 2023. Examining the nonlinear effects of residential and workplace-built environments on active travel in short-distance: A random forest approach. Int. J. Environ. Res. Public Health 20 (3), 1969.

Hankey, S., Lindsey, G., Marshall, J.D., 2017. Population-level exposure to particulate air pollution during active travel: planning for low-exposure, health-promoting cities. Environ. Health Perspect. 125 (4), 527–534.

Hillel, T., Bierlaire, M., Elshafie, M.Z., Jin, Y., 2021. A systematic review of machine learning classification methodologies for modelling passenger mode choice. J. Choice Model. 38, 100221.

Holz-Rau, C., Scheiner, J., Sicks, K., 2014. Travel distances in daily travel and long-distance travel: what role is played by urban form? Environ. Plan. A 46 (2), 488–507.

Huang, Z., Bao, Y., Mao, R., Wang, H., Yin, G., Wan, L., Qi, H., Li, Q., Tang, H., Liu, Q., et al., 2023a. Big geodata reveals spatial patterns of built environment stocks across and within cities in China. Engineering.

Huang, Z., Yin, G., Peng, X., Zhou, X., Dong, Q., 2023b. Quantifying the environmental characteristics influencing the attractiveness of commercial agglomerations with big geo-data. Environ. Plan. B: Urban Anal. City Sci. 23998083231158370.

Iroz-Elardo, N., Schoner, J., Fox, E.H., Brookes, A., Frank, L.D., 2020. Active travel and social justice: Addressing disparities and promoting health equity through a novel approach to regional transportation planning. Soc. Sci. Med. 261, 113211.

Javaid, A., Creutzig, F., Bamberg, S., 2020. Determinants of low-carbon transport mode adoption: systematic review of reviews. Environ. Res. Lett. 15 (10), 103002.

Ji, S., Wang, X., Lyu, T., Liu, X., Wang, Y., Heinen, E., Sun, Z., 2022. Understanding cycling distance according to the prediction of the xgboost and the interpretation of SHAP: a non-linear and interaction effect analysis. J. Transp. Geogr. 103, 103414.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349 (6245), 255–260.

Kaplan, S., Nielsen, T.A.S., Prato, C.G., 2016. Walking, cycling and the urban form: A heckman selection model of active travel mode and distance by young adolescents. Transp. Res. D 44, 55–65.

Kemperman, A., Timmermans, H., 2009. Influences of built environment on walking and cycling by latent segments of aging population. Transp. Res. Rec. 2134 (1), 1–9.

Koushik, A.N., Manoj, M., Nezamuddin, N., 2020. Machine learning applications in activity-travel behaviour research: a review. Transp. Rev. 40 (3), 288–311.

Lee, D., Derrible, S., Pereira, F.C., 2018. Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. Transp. Res. Rec. 2672 (49), 101–112.

Lenormand, M., Louail, T., Cantú-Ros, O.G., Picornell, M., Herranz, R., Arias, J.M., Barthelemy, M., Miguel, M.S., Ramasco, J.J., 2015. Influence of sociodemographic characteristics on human mobility. Sci. Rep. 5 (1), 10075.
Li, Z., 2023. Leveraging explainable artificial intelligence and big trip data to understand factors influencing willingness to ridesharing. Travel Behav. Soc. 31, 284–294.
Li, F., Fisher, K.J., Brownson, R.C., Bosworth, M., 2005. Multilevel modelling of built environment characteristics related to neighbourhood walking activity in older adults. J. Epidemiol. Commun. Health 59 (7), 558–564.
Liu, Z., Kemperman, A., Timmermans, H., 2020. Correlates of older adults' walking trip duration. J. Transp. Health 18, 100889.
Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015. Social sensing: A new approach to understanding our socioeconomic environments. Ann. Assoc. Amer. Geogr. 105 (3), 512–530.
Liu, J., Wang, B., Xiao, L., 2021. Non-linear associations between built environment and active travel for working and shopping: An extreme gradient boosting approach. J. Transp. Geogr. 92, 103034.
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Lundberg, B., Weber, J., 2014. Non-motorized transport and university populations: an analysis of connectivity and network perceptions. J. Transp. Geogr. 39, 165–178.
Ma, L., Dill, J., 2015. Associations between the objective and perceived built environment and bicycling for transportation. J. Transp. Health 2 (2), 248–255.
Malambo, P., Kengne, A.P., Lambert, E.V., De Villers, A., Puoane, T., 2017. Association between perceived built environmental attributes and physical activity among adults in South Africa. BMC Public Health 17 (1), 1–16.
Mirkatouli, J., Samadi, R., Hosseini, A., 2018. Evaluating and analysis of socio-economic variables on land and housing prices in Mashhad, Iran. Sustain. Cities Soc. 41, 695–705.
Nourian, P., Rezvani, S., Valeckaite, K., Sariyildiz, S., 2018. Modelling walking and cycling accessibility and mobility: The effect of network configuration and occupancy on spatial dynamics of active mobility. Smart Sustain. Built Environ. 7 (1), 101–116.
Ospina, J.P., Botero-Fernández, V., Duque, J.C., Brussel, M., Grigolon, A., 2020. Understanding cycling travel distance: The case of medellin city (Colombia). Transp. Res. D 86, 102423.
Panter, J.R., Jones, A.P., Van Sluijs, E.M., 2008. Environmental determinants of active travel in youth: a review and framework for future research. Int. J. Behav. Nutr. Phys. Activ. 5 (1), 1–14.
Pisoni, E., Christidis, P., Cawood, E.N., 2022. Active mobility versus motorized transport? User choices and benefits for the society. Sci. Total Environ. 806, 150627.
Porter, A.K., Kohl, III, H.W., Perez, A., Reininger, B., Gabriel, K.P., Salvo, D., 2018. Perceived social and built environment correlates of transportation and recreation-only bicycling among adults. Prev. Chronic Dis. 15.
Pucher, J., Buehler, R., Bassett, D.R., Dannenberg, A.L., 2010. Walking and cycling to health: a comparative analysis of city, state, and international data. Amer. J. Public Health 100 (10), 1986–1992.
Rahul, T., Verma, A., 2014. A study of acceptable trip distances using walking and cycling in bangalore. J. Transp. Geogr. 38, 106–113.
Rasouli, S., Timmermans, H.J., 2014. Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. Eur. J. Transp. Infrastr. Res. 14 (4), 412–424.
Rietveld, P., 2001. Biking and walking: the position of non-motorized transport modes in transport systems. In: Handbook of Transport Systems and Traffic Control. Emerald Group Publishing Limited, pp. 299–319.
Rodier, C., Shaheen, S.A., Chung, S., 2003. Unsafe at any speed?: what the literature says about low-speed modes.
Scheiner, J., 2010. Interrelations between travel mode choice and trip distance: trends in Germany 1976–2002. J. Transp. Geogr. 18 (1), 75–84.
Schoner, J., Chapman, J., Brookes, A., MacLeod, K.E., Fox, E.H., Iroz-Elardo, N., Frank, L.D., 2018. Bringing health into transportation and land use scenario planning: Creating a national public health assessment model (N-PHAM). J. Transp. Health 10, 401–418.
Shaer, A., Rezaei, M., Rahimi, B.M., Shaer, F., 2021. Examining the associations between perceived built environment and active travel, before and after the COVID-19 outbreak in Shiraz city, Iran. Cities 115, 103255.
Shapley, L.S., 1953. A value for n-person games. Contrib. Theory Games 2 (28), 307–317.
Shi, K., Chang, Z., Chen, Z., Wu, J., Yu, B., 2020. Identifying and evaluating poverty using multisource remote sensing and point of interest (POI) data: A case study of Chongqing, China. J. Clean. Prod. 255, 120245.
Ta, N., Chai, Y., Zhang, Y., Sun, D., 2017. Understanding job-housing relationship and commuting pattern in Chinese cities: Past, present and future. Transp. Res. D 52, 562–573.
Tao, T., Wu, X., Cao, J., Fan, Y., Das, K., Ramaswami, A., 2023. Exploring the nonlinear relationship between the built environment and active travel in the twin cities. J. Plann. Educ. Res. 43 (3), 637–652.
United Nations, 2021. Sustainable Transport, Sustainable Development. Interagency Report for Second Global Sustainable Transport Conference.
Wali, B., Frank, L.D., Chapman, J.E., Fox, E.H., 2021. Developing policy thresholds for objectively measured environmental features to support active travel. Transp. Res. D 90, 102678.
Wang, X., Liu, Y., Yao, Y., Zhou, S., Zhu, Q., Liu, M., Luo, W., Helbich, M., 2023. Associations between streetscape characteristics at Chinese adolescents' activity places and active travel patterns on weekdays and weekends. J. Transp. Health 31, 101653.
Wu, J., Wang, B., Wang, R., Ta, N., Chai, Y., 2021. Active travel and the built environment: A theoretical model and multidimensional evidence. Transp. Res. D 100, 103029.
Xiao, L., Lo, S., Liu, J., Zhou, J., Li, Q., 2021. Nonlinear and synergistic effects of TOD on urban vibrancy: Applying local explanations for gradient boosting decision tree. Sustain. Cities Soc. 72, 103063.
Xu, N., Nie, Q., Liu, J., Jones, S., 2023. Post-pandemic shared mobility and active travel in Alabama: A machine learning analysis of COVID-19 survey data. Travel Behav. Soc. 32, 100584.
Yang, L., Ao, Y., Ke, J., Lu, Y., Liang, Y., 2021. To walk or not to walk? Examining non-linear effects of streetscape greenery on walking propensity of older adults. J. Transp. Geogr. 94, 103099.
Yang, Y., Sasaki, K., Cheng, L., Liu, X., 2022a. Gender differences in active travel among older adults: Non-linear built environment insights. Transp. Res. D 110, 103405.
Yang, Y., Sasaki, K., Cheng, L., Tao, S., 2022b. Does the built environment matter for active travel among older adults: Insights from Chiba city, Japan. J. Transp. Geogr. 101, 103338.
Yazdizadeh, A., Patterson, Z., Farooq, B., 2019. Ensemble convolutional neural networks for mode inference in smartphone travel survey. IEEE Trans. Intell. Transp. Syst. 21 (6), 2232–2239.
Yin, G., Huang, Z., Bao, Y., Wang, H., Li, L., Ma, X., Zhang, Y., 2023a. ConvGCN-RF: A hybrid learning model for commuting flow prediction considering geographical semantics and neighborhood effects. GeoInformatica 27 (2), 137–157.
Yin, G., Huang, Z., Yang, L., Ben-Elia, E., Xu, L., Scheuer, B., Liu, Y., 2023b. How to quantify the travel ratio of urban public transport at a high spatial resolution? A novel computational framework with geospatial big data. Int. J. Appl. Earth Obs. Geoinf. 118, 103245.
Zhang, L., Long, R., Chen, H., Geng, J., 2019. A review of China's road traffic carbon emissions. J. Clean. Prod. 207, 569–581.