Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 6629507, 17 pages
https://doi.org/10.1155/2023/6629507

Research Article

What Affects Bus Passengers' Travel Time? A View from the Built Environment and Weather Condition

Xiaowei Li, Lanxin Shi, Haotian Li, Qian Liu, and Jun Chen

School of Civil Engineering, Xi'an University of Architecture & Technology, Xi'an 710055, China

Correspondence should be addressed to Xiaowei Li; lixiaowei@xauat.edu.cn

Received 7 June 2023; Revised 4 September 2023; Accepted 25 September 2023; Published 9 October 2023

Academic Editor: Muqing Du

The study aims to examine the impact of the built environment and weather conditions on travel time for bus passengers in Weinan, China. Various sources of data, including smart card data, bus GPS data, bus station data, road information data, and smart card swiping time, were integrated and analyzed. The study employed the light gradient boosting machine (LightGBM) model and SHapley Additive exPlanations (SHAP) value to assess the feature importance and nonlinear effects of different types of POI density, weather conditions, and time series on bus passengers' travel time. The study findings indicate that several factors are associated with bus passengers' travel time, including destination residential density, destination diversity, destination life service density, origin science and education density, origin residential density, origin diversity, humidity, visibility, boarding time between 7 and 8 a.m., and precipitation. This study also reveals nonlinear threshold effects. The study findings provide valuable insights that can be utilized to optimize the bus network and develop low-carbon-oriented land-use planning.

1. Introduction

The increasing number of motor vehicles has brought about challenges such as congestion, noise, and pollution to urban transportation systems. Public transportation systems have become essential components of the transportation systems in most cities worldwide [1]. Travel time, the duration required to complete a trip, is influenced by travel distance, travel speed, and various nontravel factors. For bus passengers, managers, planners, and operators, bus travel time is an essential transportation parameter that helps in the development of bus route planning and optimization of bus operation networks [2]. In many small- and medium-sized cities located in the underdeveloped regions of western China, the bus is often the only mode of transportation available to connect different parts of the city. Consequently, the bus travel time largely represents the commuting time of residents in these cities. Therefore, exploring the evolution of bus passengers' travel time and identifying the factors that influence it can help relevant departments to predict residents' bus travel time, plan the bus operation network reasonably, and provide a scientific and reasonable basis for the development of native low-carbon strategies.

In recent years, various studies have been conducted on travel time prediction using relevant models [3–5]. Several factors that influence bus travel time have been analyzed, including congestion within the bus, bus section length, traffic signal density, land use, and departure delay relative to the expected departure time [6–8]. Moreover, previous studies have shown that the built environment [9–11] and weather conditions [12–15] significantly impact bus travel. However, most of these studies have focused on large cities in developed countries. As small- and medium-sized cities differ in size, population, land development, and climate region, travel time and the related influencing factors of passengers in such cities may differ from those in large cities. Therefore, the study of bus travel time and its influence factors in small- and medium-sized cities remains largely unexplored.

Moreover, most of the current studies have utilized traditional small sample data to analyze bus travel time and its related factors [16, 17]. However, small sample survey data cannot fully and accurately obtain the origin, destination, and departure time of bus passengers, thereby making it challenging to accurately reveal the joint influence of the built environment at the origin and destination, weather, and temporal and spatial characteristics on travel time. Furthermore, previous studies have shown that the impact of the built environment on travel behavior is often nonlinear [18, 19]. Therefore, the nonlinear threshold effect is essential in understanding how the regulation of the built environment and weather influence can be utilized to reduce the travel time of passengers. This can help planners determine what type of urban planning controls can effectively reduce travel time. However, the nonlinear effects of factors related to bus travel time have not been fully revealed.

The aim of this study is to examine how theweather conditions and built environment at both the origin and destination affect the travel time of bus passengers in Weinan, a medium-sized city located in western China. To achieve this goal, multisource spatiotemporal big data for two consecutive weeks from November 12 to November 30, 2018, was collected from Weinan Bus Company. In addition, POI and weather data were obtained through web crawler technology. The interpretable LightGBM model was used to explore the feature importance and nonlinear effects of multidimensional factors.

The study makes two main contributions. Firstly, it reveals the effects of the built environment, weather, and spatiotemporal characteristics on bus travel time. This provides new insights into the resilience of the bus transit system. Secondly, it employs an interpretable machine learning approach using LightGBM and SHAP to uncover the feature importance and nonlinear effects of significant influencing factors.

2. Literature Review

The urban built environment is a complex system that comprises land use patterns, urban design, and transportation systems. One widely adopted framework for characterizing the built environment is the 6D elements, which include density, diversity, design, distance to transit, destination accessibility, and demand management [20]. However, to obtain a more accurate understanding of travel behavior, scholars have introduced point of interest (POI) data as an additional indicator to characterize the built environment [21]. POI data are an emerging geographic location big data that offer the advantages of large data volume, low cost of acquisition, fast update speed, and comprehensive information coverage. A plethora of studies have utilized POI data to analyze and understand the built environment. For instance, Krumm and Mummidi [22] utilized user-generated POI data to analyze access to POIs. Xie and Yan [23] proposed a network kernel density analysis that integrated the line element into kernel density analysis to improve operational efficiency. Kwan [24] analyzed travel characteristics by using POI data to obtain activity density and spatial distribution of residents and conducted spatial and temporal simulations using GIS methods. McKenzie et al. [25] explored regional variability in POIs by analyzing the differences in characteristics of POIs in various regions. Becker et al. [26] analyzed population migration characteristics by examining POI point data of call locations in Morristown. Lian et al. [27] weighted POI check-in information, expanded the influence area of the POI through user check-in frequency, and provided users with POI recommendations. Yue et al. [28] found that mixed land use had a significant impact on travel behavior, and residential, employment, entertainment, and life service facilities could reduce travel distances. Rich diversity can provide more job opportunities for nearby residents and reduce the proportion of working residents who use private cars [29]. The impact of diversity is not limited to regional diversity but can also reflect individual attribute diversity. For instance, higher-income groups have more complex and longer travel chains and more flexible travel patterns compared to lower-income groups [30]. Sun et al. [31] found that better-designed urban roads with a better road network and fewer parking spaces encouraged residents to use green modes of travel, such as public transportation and bicycles. Loo et al. [32] argued that the design, rationality, and perfection of the public transport network have a significant impact on the proportion of bus travel. When dedicated corridors are provided between bus and railway stations, the proportion of railway travel increases significantly. Wells and Hutchinson [33] emphasized that improving station accessibility is crucial to enhancing the quality of public transport services and increasing the attractiveness of public transport travel. Choi and Zhang [34] highlighted that accessibility can be assessed based on the distance from the residence to the city center, with shorter distances indicating better accessibility and thereby reducing travel distances to some extent, which could influence travel mode choice in favor of public transportation.

Weather is a real-time environmental factor that can have a significant impact on urban traffic travel. Research into the mechanisms by which weather affects travel can provide valuable insights into the relationship between weather and transportation, ultimately leading to the development of more effective and efficient transportation systems [35]. Numerous studies have explored the influence of various weather parameters such as temperature, humidity, and rainfall on public transit travel behaviors, including traffic volume, travel mode choice, travel time, and distance traveled, among others [12–15, 36–42]. For instance, with regard to traffic volumes, Ngo [12] found that bus ridership decreases significantly during extreme weather conditions, particularly during periods of very hot, cold, or heavy precipitation. Kalkstein et al. [13] surveyed three regions in the United States (the Bay Area, Chicago, and northern New Jersey) and found that public transportation travels were greater during periods of dry, comfortable weather compared to wetter and cooler weather conditions. Similarly, Arana et al. [14] studied the impact of weather conditions on buses in Guipuzcoa, Spain, and found that strong winds and heavy rain led to a reduction in bus travel. Singhal et al. [15] investigated the impact of rainfall on travel volumes on different days of the week and found that, during weekends, the negative impact of rainfall is consistent, while during working days, rainfall has a greater impact on travel volume during morning and evening rush hours. Regarding travel time, Kamga and Yazıcı [39] found that during severe weather, such as heavy rainfall, taxi passengers' trips tend to be work-oriented and necessary, and these trips generally take longer to complete, while nonnecessary trips such as short-distance shopping and fun trips are reduced due to the occurrence of severe weather. Tsapakis et al. [40] noted that, in London, light, moderate, and heavy rainfall resulted in increases in travel time for motor vehicles of 0.1–2.1%, 1.5–3.8%, and 4.0–6.0%, respectively, while the effect of temperature on travel time was almost negligible. Li et al. [41] analyzed the characteristics of travel time under different weather conditions to determine the change rule of travel time and speed of motor vehicles on urban highways under different rainfall intensities and visibility. The results showed that the new model for predicting travel time with consideration of weather conditions has less error and can effectively improve the calculation accuracy compared with the Bureau of Roads and Highways (BPR) function model.

The implementation of machine learning algorithms for travel time prediction mainly focuses on two areas: predicting vehicle arrival times and estimating traffic travel times. Vehicle arrival time prediction involves collecting arrival information of buses at each station and then using time series analysis, Kalman filtering, support vector machines (SVM), and other methods to process the data. For example, Abidin and Kolberg [43] used Kalman filtering to process real-world input data, allowing their model to overcome existing models' data processing limitations. This approach predicts arrival times by utilizing information obtained from social networks, particularly Twitter, and simulates the vehicle arrival time through urban traffic simulation software. The results demonstrated excellent performance. He et al. [44] proposed a multi-index evaluation method for bus arrival time prediction based on SVM. This method uses three new metrics, including GPS coverage, release rate, and accuracy, to evaluate the prediction service. The SVM model is then trained using these metrics to evaluate the accuracy of the prediction. The study concluded that the SVM-based multi-index evaluation method is intuitive and comprehensive, allowing for the accurate identification of issues based on the three new indicators. Chen et al. [45] developed a neural network-based bus arrival time prediction model that uses automated passenger count data collected by the New Jersey Transportation Bureau. The model was tested and found to predict travel time fairly accurately.

In the domain of travel time prediction, machine learning methods have gained considerable attention in recent years. Zhang and Ge [46] introduced a Takagi–Sugeno–Kang fuzzy neural network- (TSKFNN-) based online prediction method for expressway corridor travel time. In another study, Ran et al. [47] utilized convolutional neural networks (CNNs) for predictive analysis of short pass times. The study considered the stochastic nature of traffic demand, spatiotemporal dependence between traffic flows, and other periodic and nonperiodic factors to develop a relevant predictive model. Gupta et al. [48] employed random forest and gradient boosting to analyze one-year taxi travel time data in Porto City and compared the accuracy of the two models. The results showed that gradient boosting slightly outperformed random forest in predicting travel times. He et al. [49] developed a passenger travel time prediction model for multiple road sections and starting points in Singapore based on bus swipe data. The model used long- and short-term memory networks (LSTM) and was compared with other models such as time series and SVM. The study validated the superiority of LSTM for passage-time prediction.

In recent years, LightGBM has emerged as a popular choice for interpretable machine learning models due to its numerous advantages. Chen and Guestrin [50] found that LightGBM and XGBoost support parallel algorithms, but LightGBM is more powerful, is faster to train, and consumes less memory than XGBoost, thus reducing the communication cost of parallel learning. Compared to deep learning methods, statistical models (e.g., SVM), and graphical models (e.g., Bayesian belief networks), LightGBM is more predictable [51]. Bentéjac et al. [52] compared existing gradient boosting models and demonstrated that LightGBM has unique advantages in terms of training speed during algorithm optimization, especially for larger datasets. Moreover, Wen et al. [53] utilized LightGBM to quantify the influence of hazard factors on the probability of vehicle crashes. Using Texas vehicle crash data, they analyzed and compared the effect of key crash factors on the total number of crashes and the number of crashes of different types. Their results indicate that LightGBM outperforms other models in terms of mean absolute error (MAE) and root mean square error (RMSE).

Machine learning models are not only capable of analyzing variables with significant effects but are also highly interpretable in terms of nonlinearities and threshold effects. In their recent study, Zhang et al. [54] demonstrated the usefulness of nonlinear analysis in examining the effective sphere of influence of urban and suburban centers, which can provide planning guidelines for polycentric development. In addition, threshold effects can assist planners in determining the normative range of land use settings for planning pedestrian-scale neighborhoods, such as the 15-minute living circle scheme in China. Many local land-use variables exhibit unique threshold effects, including local indicators of daily facilities and density. These findings highlight the importance of nonlinear and threshold analyses and suggest that polycentric development and neighborhood living circle planning can help reduce vehicle use in Beijing. Similarly, Yang et al. [55] explored the nonlinear and threshold effects of the built environment on e-scooter sharing using two spatial analysis units (census area and census block groups) and four temporal units of analysis (spring, summer, autumn, and winter).

After conducting a thorough literature review, there is a lack of analysis regarding the impact of the built environment and weather factors on bus travel time and the threshold effects of these influencing factors remain undisclosed, especially in small and medium-sized cities in developing countries. However, with the increasing maturity of LightGBM technology and SHAP analysis, it is expected to explore the feature importance of built environment and weather factors on bus travel time and their nonlinear threshold effects. This study employs big data and LightGBM models to investigate the mechanistic analysis of weather and the built environment on bus travel time.

3. Data

3.1. Research Subjects. Weinan is a prefecture-level city situated in Shaanxi Province and is renowned for its historical significance as the birthplace of the Chinese nation. It is also known as the root of China, the source of culture, the hometown of three saints, and the town of generals. The city is located in the eastern part of the Guanzhong Plain and the eastern part of Shaanxi Province, with the Wei River serving as the primary terrain axis. It encompasses five major geomorphological zones, including the two mountains in the north and south, the two plateaus, and the central flat river, and experiences a temperate monsoon climate. Weinan has a total area of 13,030 square kilometers and comprises two districts, seven counties, and two county-level cities. The city is an important ecological protection barrier zone and a crucial economic zone in the Yellow River section of Shaanxi. Figure 1 provides a general layout of Weinan City.

As of the end of 2018, the main urban area of Weinan (Linwei District) had a total of 24 bus operating lines with a bus network length of 187.6 km and an average bus line length of 15.7 km. This includes 19 unmanned ticket lines and 5 suburban lines. The urban areas have 377 bus stations, with a coverage rate of 64.5% for a 300-meter radius covering 35.10 square kilometers and a coverage rate of 91.4% for a 500-meter radius covering 49.71 square kilometers. The urban bus-sharing rate is about 10.31%, and a total of 418 buses were put into use, including 492 standard buses operating in the main urban area, resulting in a bus ownership rate of approximately 9.98 standard buses/10,000 people. Furthermore, 100% of electric vehicles are used for state-owned bus lines in the central urban area. Figure 2 illustrates the public transport network in Weinan City.

3.2. Data Collection. The Advanced Public Transportation Systems (APTS) data [56] used in this study are bus multisource data provided by the Weinan bus corporation for two consecutive weeks from November 12 to November 30, 2018. The data includes bus smart card data, bus GPS data, bus station data, road information data, bus station POI data, and weather data. Due to the certain correlation between APTS data, multiple tables can be associated with data fusion technology. Smart card data, bus GPS data, and bus station data are associated based on the related fields. The data source-association relationship is shown in Figure 3.

3.2.1. Bus Smart Card Data. The data collected from the smart card system contain various fields such as user name, card swiping number, amount, card swiping time, rate, and others. To simplify data processing and minimize interference, redundant fields are removed, and only relevant information such as users' travel date, boarding swiping time, bus route, self-numbering, bus registration number, terminal number, and smart card number are retained.

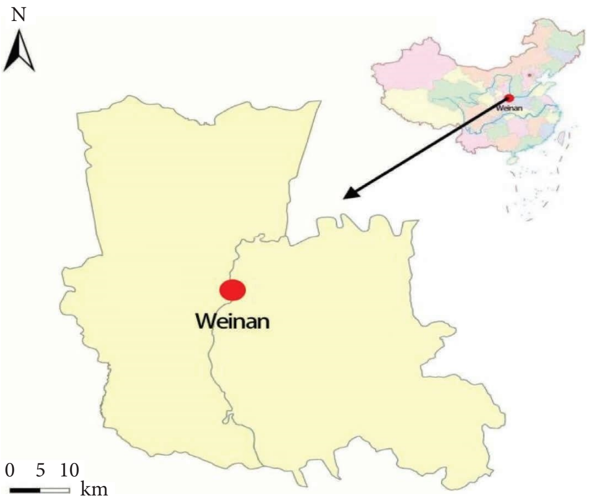


Figure 1: General layout of Weinan in China.

3.2.2. Bus GPS Data. The onboard GPS data covers all vehicles of the 24 bus lines operated by the Weinan Bus Company. The data include the date, time, bus route, self-numbering, bus registration number, running mileage, instantaneous bus speed, longitude, and latitude of the bus lines.

3.2.3. Bus Station Data. The bus station data include information on the GPS coordinates of 1280 bus stations in Weinan. The data provide details such as the station name, number, route, latitude, and longitude, as well as the direction of the upline and downline.

3.2.4. Road Information Data. To obtain basic road information, the Baidu Map "API Console" was utilized, which provided data on coordinate location, road name, unique identification number, road level, one-way status, bridge status, and other related information.

3.2.5. Bus Station POI Data. The Amap open platform API was utilized to collect POI distribution information near bus stations in Weinan City. The Amap platform subdivides POI interest points into more than 20 small types, which we classified into five large types based on their use: commercial land, science and education land, residential land, office land, and life service land. We connected the station coordinate data to the Amap API interface and collected the above five types of POI data within a 500-meter radius of the bus station, according to the radius size and POI coding rules. The station name in the bus passenger boarding and unloading data was matched with the corresponding POI value, and the direction of the bus was determined. We selected ten feature values, including the number of

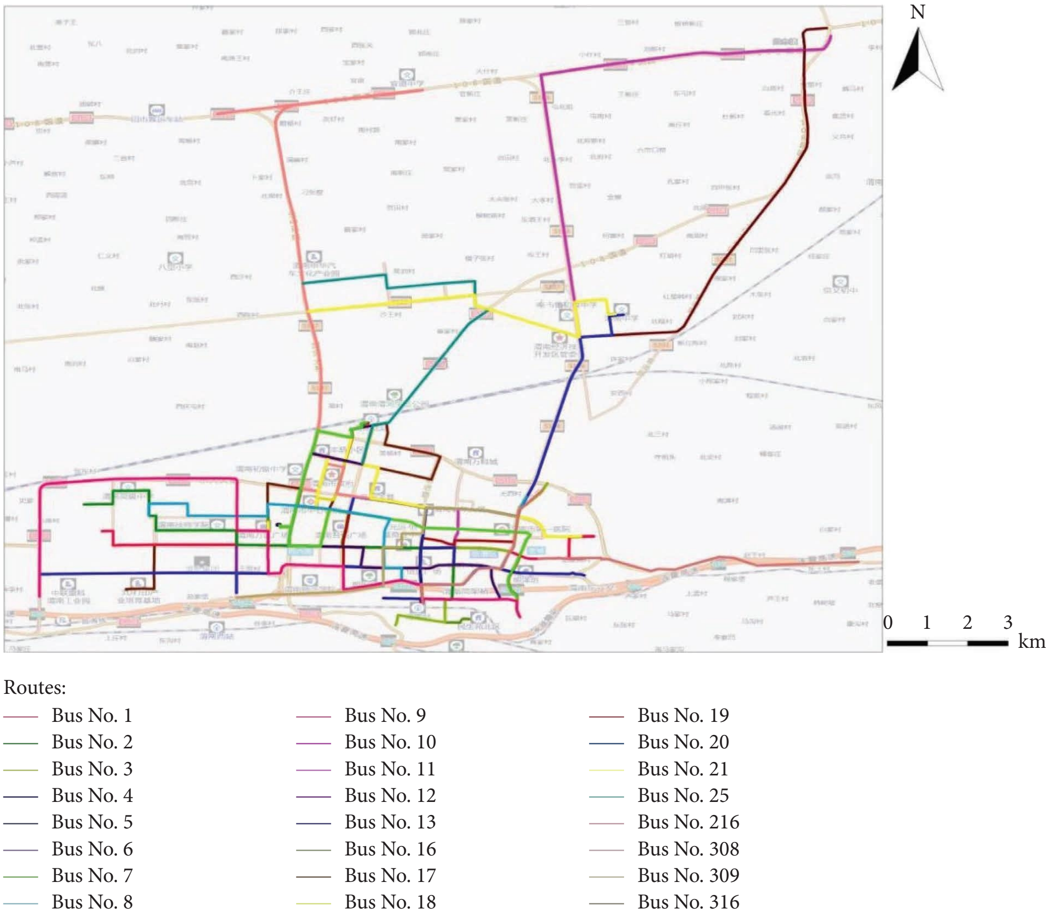


Figure 2: Bus routes' network of Weinan.

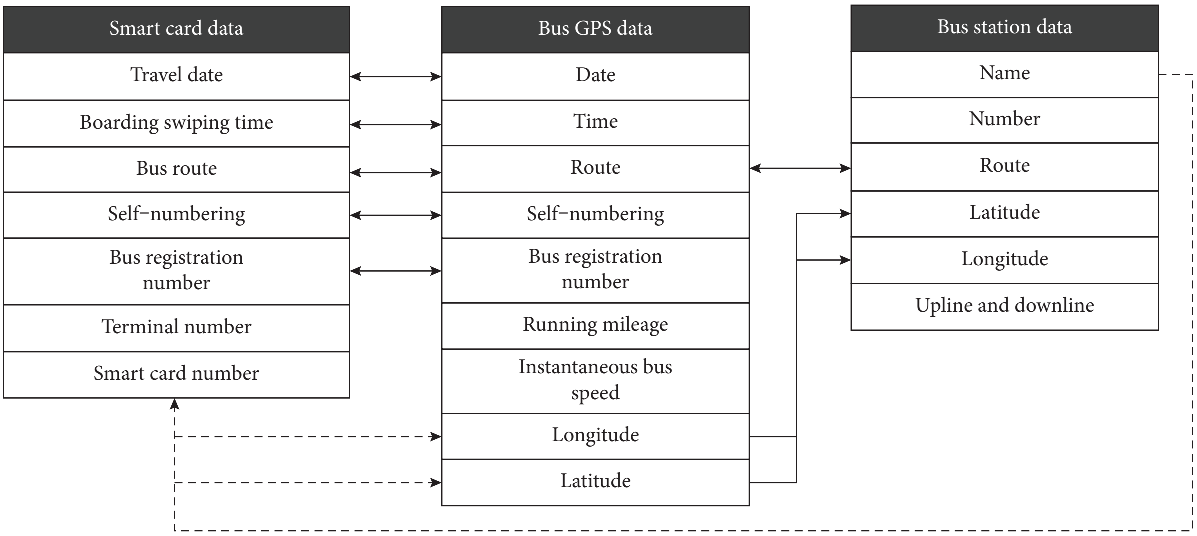


Figure 3: APTS data association relationship.

commercial POI, science and education POI, residential POI, office POI, and life service POI in the off-boarding station and within 500 meters of the off-boarding station. To characterize the built environment of the city when modeling, we used density, defined as the number of POI on a 500 by 500 square meter block area. These features were denoted successively as "origin commercial density," "origin science and education density," "origin residential density," "origin office density," "origin life service density," "destination commercial density," "destination science and education density," "destination residential density," "destination office density," and "destination life service density."

3.2.6. Weather Data. The historical weather information was obtained by using Python crawlers on the website (https://lishi.tianqi.com/weinan/201811.html). Temperature, humidity, precipitation, and visibility were extracted as variables for measuring the weather conditions.

3.3. Data Judgment. In Weinan, residents swipe their cards only when boarding the bus, not when alighting. Since there is a correlation between the various APTS datasets, data fusion techniques can be used to obtain more comprehensive information than single datasets. The method used to determine the boarding and alighting stations and the bus travel time is primarily based on existing data processing methods, with modifications and adjustments made to the calculation methods to suit the characteristics of the existing data.

3.3.1. Boarding Station Judgment. By utilizing data from bus smart cards, valuable information can be extracted for effective decision-making in bus system planning and management. Moreover, accurately identifying the boarding stations of passengers via smart card data is fundamental for utilizing such data in analyzing bus travel time, as established in previous studies by Chen et al. [57] and Chen and Yang [58].

Step 1: Determine the coordinates of the bus when swiping a smart card

Preliminary Screening. The study utilizes the GPS data of Weinan buses and smart card swiping data to match the route number, vehicle number, swiping date, and swiping time of the smart card data with the bus GPS data by truncating the records. Furthermore, as buses primarily follow fixed routes and there are no identical bus routes with the same paths in the city, this study employs the discrete Freixian distance (DFD) algorithm, which is a distance-based trajectory similarity measurement algorithm, to match the current Weinan bus onboard GPS data with the Weinan bus station data to obtain alternative bus GPS data [59].

Final Determination. The study matches the time of smart card swiping with the operating time of buses, which is from 6 a.m. to 10 p.m. in Weinan. The swiping records are sorted by time and aligned with the GPS operating records. The coordinates of the bus vehicle at the time of swiping are determined by taking the "longitude" and "latitude" values of the lowest instantaneous speed in the alternative GPS data. If multiple GPS data points have the same minimum instantaneous speed, the one closest to the 30-second mark within 1 minute of the swiping time is selected.

Step 2: Determine the direction of the upline or downline of the bus

The bus smart card data are correlated with the operation record data to determine the direction of the bus, whether it is upline or downline, at the time of card swiping. Specifically, the direction is determined based on the location of the station, where the card was swiped. If the station is a secondary station and follows a main station, then the bus is considered to be traveling in the downline direction. Conversely, if the station is not following a main station, then the bus is considered to be traveling in the upline direction.

Step 3: Match the bus coordinates with station coordinates when swiping smart cards

When a passenger swipes their smart card while boarding a bus, the system calculates the distance between the current coordinates of the bus and the coordinates of all the stations listed in the bus station data table. This calculation is performed to identify the nearest station to the bus. The station with the minimum distance is considered to be the one where the passenger gets on the bus. This process is based on the principle of identifying the nearest neighbor using distance calculations in mathematics and geography.

3.3.2. Alighting Station Judgment. The process of determining a bus passenger's alighting station involves analyzing the time and space relationship of their card number within the travel chain of the bus. By examining the sequence of bus stops visited by the passenger, their final destination can be calculated. The travel chain-based station estimation method is based on the following four assumptions [60].

Hypothesis 1. The same route matches the same direction. Assuming that a passenger takes two bus trips on the same route and direction and boards at the same or nearby stations, we can infer that their alighting station on the current day is likely to be the same or similar to their alighting station on the previous day.

Hypothesis 2. The same route matches the opposite direction. If a passenger takes two bus trips on the same route but in opposite directions, and these trips occur on the same day and a previous day, we can infer that their alighting station for the current day's travel is likely to be the same as their boarding station or a nearby station on the previous day's travel.

Hypothesis 3. Different routes match the same direction. When a passenger takes two bus trips on different routes but in the same direction, and these trips occur on the same day and a previous day, we can predict that the downstream station on the current day's route, which is the same or close to the alighting station on the previous day's travel, is likely to be the alighting station for the current day's travel.

Hypothesis 4. Different routes match the opposite direction. Suppose a passenger takes two bus trips on different routes, one on a previous day and the other on the same day. If the direction of travel on the two routes is opposite and the routes themselves are different, then we can infer that the station where the passenger alighted on the previous day is likely to be the same or close to the downstream station on the route taken on the same day.

The basic steps are shown as follows:

Step 1: Deletion of single smart card records

Once the boarding station has been identified, any records that have been determined to be incorrect or impossible are removed from the dataset. In addition, any consecutive records that are identical, indicating multiple swipes at the same station, are also removed. Finally, any single swiping records throughout the day that do not represent a complete trip are removed as well.

Step 2: Traversal verification of whether the hypotheses are satisfied

The card numbers are sorted and examined one by one to determine if they satisfy the four hypotheses mentioned above. If a card fails to meet these criteria or if the distance between two consecutive boarding stations is outside the reasonable distance range, the card is considered unprojectable and excluded from further analysis.

3.3.3. Travel Time Calculation. The time difference between the boarding and alighting swiping times is used to calculate the passenger's travel time by bus. The swiping records are sorted in chronological order, and the time interval between each swiping time and the GPS running time immediately before or after it is calculated. The data point that is closest to both the swiping time and the adjacent running time is selected based on this interval, ensuring that the swiping time and its corresponding running time are as consistent as possible.

Step 1: Add the "Travel Time" column to the database

A new column named "Travel Time" is added to the database to record the travel time.

Step 2: Judge the boarding time

The operating hours for buses in Weinan are from 6 a.m. to 10 p.m. Therefore, the boarding time for each passenger can be assumed to fall within this time frame. The boarding time is determined by identifying the boarding station and retrieving the corresponding smart card swiping record for that station.

Step 3: Judge the alighting time

The process of determining the alighting time primarily involves analyzing the time and spatial relationship between consecutive swiping records associated with the same card number. By examining the travel chain formed by the passenger's bus boarding and alighting stations, the alighting time can be calculated with reasonable accuracy.

Step 4: Correct smart card swiping time and association analyze bus GPS data

To ensure the accuracy of time data and avoid discrepancies between bus travel time data and bus routes, we use a combination of the time similarity method and the time average deviation method to calculate the similarity between vehicle GPS track data and smart card data. Through several days of data verification, we can fully establish the corresponding relationship and make necessary time corrections [59].

Step 5: Calculate the interval between boarding and alighting time to obtain passengers' bus travel time

3.4. Data Description. The collected data are subjected to normalization to ensure its reliability and suitability for subsequent analysis. This involves two steps: (1) outlier processing, which involves checking for missing and duplicate values that could adversely affect subsequent studies and (2) text labeling, which involves digitizing text information in the data and labeling data such as boarding and alighting stations, smart card numbers, bus running directions, and weather.

Table 1 shows the results of the basic data descriptive analysis. For this study, travel time is selected from two consecutive weeks of 7, 8, and 9 a.m. peak data in November 2018 in Weinan City.

4. Methodology

4.1. LightGBM

4.1.1. Overview of LightGBM. LightGBM is a framework based on XGBoost, which was released by Microsoft in 2017 [61]. Both LightGBM and XGBoost [50] support parallel algorithms, but LightGBM is more powerful than XGBoost due to its faster training speed and lower memory requirements, which reduce the communication cost of parallel learning. LightGBM has several main features, including gradient-based one-side sampling (Goss), exclusive feature bundling (EFB), and a histogram and leaf-oriented growth strategy with depth limitation. Goss can achieve a balance between the number of samples and the precision of LightGBM's decision tree. During the training process, downsampling will give more weight to the samples with larger gradients, which have a greater impact on information acquisition. When the feature space is sparse, LightGBM can use EFB to group mutually exclusive features into new features, thereby reducing the dimensionality of the feature space.

Table 1: Descriptive statistics of original variables.

Variables	Units	Count	Mean	Std.	Min	Max
Travel time	min	7822	20.58	14.60	4	70
Longitude	°	7822	109.47	0.03	109.41	109.53
Latitude	°	7822	34.50	0.01	34.48	34.53
Temperature	˚C	7822	3.8	2.42	−0.8	7.9
Humidity	%rh	7822	75.45	15.16	44	98
Precipitation	mm	7822	0.01	0.05	0	0.3
Visibility	m	7822	5.78	6.09	0	26
Origin road density	km/km^2	7822	4.12	0.95	0.93	10.57
Origin business density	pcs/500 m^2	7822	550.95	471.06	1	1920
Origin science and education density	pcs/500 m^2	7822	55.16	45.48	1	239
Origin office density	pcs/500 m^2	7822	70.08	48.86	1	266
Origin life service density	pcs/500 m^2	7822	256.03	198.26	1	921
Origin residential density	pcs/500 m^2	7822	53.68	48.65	1	497
Origin diversity	pcs/500 m^2	7822	0.75	0.09	0.38	0.98
Destination road density	km/km^2	7822	4.12	0.95	0.93	10.57
Destination business density	pcs/500 m^2	7822	791.92	609.04	1	1920
Destination science and education density	pcs/500 m^2	7822	75.89	55.78	1	194
Destination office density	pcs/500 m^2	7822	91.43	67.18	3	266
Destination life service density	pcs/500 m^2	7822	359.02	259.73	1	921
Destination residential density	pcs/500 m^2	7822	73.47	58.65	1	326
Destination diversity	pcs/500 m^2	7822	0.73	0.09	0.38	0.96

LightGBM performs better than existing decision tree methods such as deep learning methods (e.g., neural networks), statistical models (e.g., SVM), and graphical models (e.g., Bayesian belief networks) [51]. However, existing decision tree methods have low performance and require the use of specific algorithms such as GBDT, Goss, and EFB. Therefore, the LightGBM method is used in this study. Its basic idea is to linearly combine N weak regression trees into a strong regression tree [62]. The calculation formula is given as follows [63]:

$$\mathrm{G}(x) = \sum_{n=1}^{N} \mathrm{g}_{\mathrm{n}}(x), \tag{1}$$

where $\mathrm{G}(x)$ is the final output and $\mathrm{g}_n(x)$ is the output of the weak regression tree.

The LightGBM model has made significant improvements in two key areas: the histogram algorithm and the leaf-wise strategy with depth limitation. The histogram algorithm converts continuous data into K integers and constructs a histogram with a width of K. During the traversal process, the discrete values are accumulated as indexes in the histogram, and the optimal decision tree segmentation points are searched. The leaf-wise strategy with depth limitation means that during each split, the leaf with the greatest gain is selected to split and cycle. Furthermore, the complexity of the model is reduced, and overfitting is avoided by limiting the depth of the tree and the number of leaves.

4.1.2. Model Building and Hyperparameter Tuning.

Initially, we utilized the LightGBM model with default parameters, and subsequently, we employed the bagging algorithm to perform hyperparameter tuning. Bagging is a well-known ensemble learning method that combines the outcomes of multiple weak learners to collaborate on a shared learning task. This method involves generating N sets of samples by repeatedly sampling M samples. Each sample set is used to train a separate learning model, resulting in N weak learners.

To begin the tuning process, we first focus on adjusting the "*n*_estimators" and "learning_rate" parameters. The default value for "*n*_estimators" is 100, and we explore a broad range of values to identify an optimal number before narrowing down to a threshold range. The default value for "learning_rate" is 0.1, and we adjust this based on the specific needs of the data. Next, we adjust the "num_leaves" and "max_depth" parameters, with "num_leaves" defaulting to 31, but ultimately determined by the data. These parameters can be adjusted simultaneously, using a "coarse tuning, then fine-tuning" strategy. Finally, we use a larger "*n*_estimators" value to train the data using the optimized parameters obtained from the tuning process.

4.1.3. Model Evaluation.

Regression models were evaluated using the common statistical functions: mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), root mean squared logarithmic error (RMSLE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). It should be noted that the ideal value for R^2 is 1, and the ideal values for MAE, MSE, RMSE, RMSLE, and MAPE are all 0.

4.2. SHAP.

SHapley Additive exPlanation (SHAP) is a model that utilizes the principles of additive explanation inspired by cooperative game theory [64]. The SHAP value is a technique that determines the relative contribution of each input variable in generating the final output variable. This concept is similar to parametric analysis, where one variable is changed while the others remain constant to observe the effect of the changed variable on the target attribute [65].

In the SHAP method, all features in the dataset are considered contributors to the predictions made by the machine learning model [66]. For each sample that is predicted by the model, SHAP generates a value called the SHAP value, which is assigned to each feature in the sample. SHAP maps the original input value r to a simplified input value z and creates a simplified function of $f(z)$. Using the additive feature attribution method, SHAP constructs a binary linear function to estimate the objective function f. The explanation process is as follows [67]:

$$f(r) = u(z) = \varphi_0 + \sum_{m=1}^{M} \varphi_m z_m, \tag{2}$$

where z_m equals one when feature m is involved in the prediction process and zero otherwise. M is the number of features and φ_m is the contribution of feature m. The contribution of feature m is described in the following [67]:

$$\varphi_m = \sum_{S \subseteq F/\{m\}} \frac{|S|!(M - |S|! - 1)!}{M!} \left[f_r(S \cup \{m\}) - f_r(\mathrm{S})\right], \tag{3}$$

where F is the set of all input features, S is a subset of different feature combinations without feature m, and $f_r(C)$ is a prediction model with input instance r conditional on the subset C of feature combination. For SHAP, $f_r(C)$ is defined as $E[f(r)|r_C]$ that this means that the function f has an exceptional value in terms of the input characteristic subset C [68].

The conventional feature importance method only measures the importance of features without considering their effect on the prediction results. However, SHAP provides both the magnitude and direction of the impact of each feature on the prediction results for each sample. The SHAP value is used to calculate feature importance, and visualization results are generated. We present SHAP dependency plots for the most important features to describe their impact on the predicted output. For analyzing the nonlinear effects of a single variable, partial dependence plots are used to show the partial dependence of travel time [69].

5. Results

We implemented the LightGBM regression model using Python 3.8. The available dataset was partitioned into two disjoint sets, namely, the training set and the testing set. Random sampling was applied to split the dataset, where 80% of the samples were assigned to the training set and the remaining 20% were assigned to the testing set.

5.1. Models Construction. Before establishing the model, we conducted univariate regression to identify significant factors contributing to travel time. We also examined the correlations among these significant factors. However, we found strong correlations among specific pairs of variables. To address multicollinearity issues, we removed one variable from each highly correlated pair based on their variance inflation factor (VIF) values [70]. As a result, the following variables were included from the variable set used for modeling: destination diversity, destination residential density, destination life service density, origin diversity, origin residential density, origin science and education density, visibility, precipitation, and humidity.

We utilized the same dataset to compare the performance of LightGBM models against other machine learning models. Our findings suggest that the LightGBM model outperformed other models in terms of accuracy and fitting performance, as presented in Table 2. We constructed LightGBM models with and without hyperparameter tuning using the bagging method. The important hyperparameters of the two models are presented in Table 3, with the ones that were altered after tuning highlighted in red. To assess their performance, we conducted a 5-fold cross-validation separately for each model, as shown in Table 4. The results indicate that the LightGBM model after hyperparameter tuning exhibited significantly better 5-fold cross-validation performance compared to the model before tuning. The process of generating LightGBM models and their prediction performances are presented in Figure 4.

5.2. Explicability of Variables. Figure 5 displays the impact directionality of the top 10 features on travel time, where the overall impact of each variable on travel time is represented by its average SHAP value across all samples, indicating the average impact of each variable on travel time. Each data point in the graph represents a sample of data, where DateTime_hour_7 represents boarding time between 7 and 8 a.m. Figure 6 depicts the effect of POI and weather variables on passenger travel time in a nonlinear relationship with a threshold effect. This graph can be utilized to analyze the threshold effect of a single independent variable on the dependent variable while holding other variables constant. Moreover, the black rugs on the horizontal axis represent the actual data distribution.

5.2.1. Residential Density. In the POI variables, we find both the residential density at the origin and destination play important roles in influencing travel time during the morning peak hours. In particular, destination residential density is the primary factor affecting travel time, and this influence is predominantly inhibitory. This is consistent with the findings of Feng et al. [71], highlighting the suppressive role of residential density in travel time. This could be because higher residential population density leads to the growth of supporting industries in the community. Consequently, residents in such areas spend less time on essential travel due to the availability of amenities within proximity.

In Figure 6(a), it is evident that the overall reduction in travel time for residents persists with an increase in destination residential density. However, a noteworthy upward rebound in travel time occurs when the destination residential density falls within the range of 135 to 160 pcs/

Table 2: Comparison of five machine learning algorithms.

Model	MAE	MSE	RMSE	RMSLE	MAPE	R^2
LightGBM	**7.1327**	**90.8591**	**9.5316**	**0.4672**	**0.5140**	**0.5620**
GBDT	9.2712	146.6968	12.1110	0.5743	0.6848	0.2930
AdaBoost	11.6253	195.2151	13.9703	0.7054	1.0033	0.0592
Bayesian ridge	10.9980	196.8434	14.0292	0.6680	0.8451	0.0513
LR	10.9960	196.8575	14.0297	0.6679	0.8448	0.0513

We have used bold formatting to emphasize that the LightGBM model outperforms these performance metrics (MAE, MSE, RMSE, RMSLE, MAPE, and R^2) among these models, in order to enhance the readability of the article.

Table 3: Hyperparameter for LightGBM models.

Hyperparameters	Values	
	Before tuning	After tuning
boosting_type	GBDT	GBDT
class_weight	None	None
colsample_bytree	1.0	1.0
importance_type	Split	Split
learning_rate	0.1	0.3
max_depth	−1	−1
min_child_samples*	20	86
min_child_weight	0.001	0.001
min_split_gain*	0.0	0.2
*n*_estimators*	100	280
*n*_jobs	−1	−1
num_leaves*	31	90
Objective	None	None
random_state	5734	5734
reg_alpha*	0.0	0.05
reg_lambda*	0.0	0.005
Silent	Warn	Warn
Subsample	1.0	1.0
subsample_for_bin	200000	200000
subsample_freq	0	0

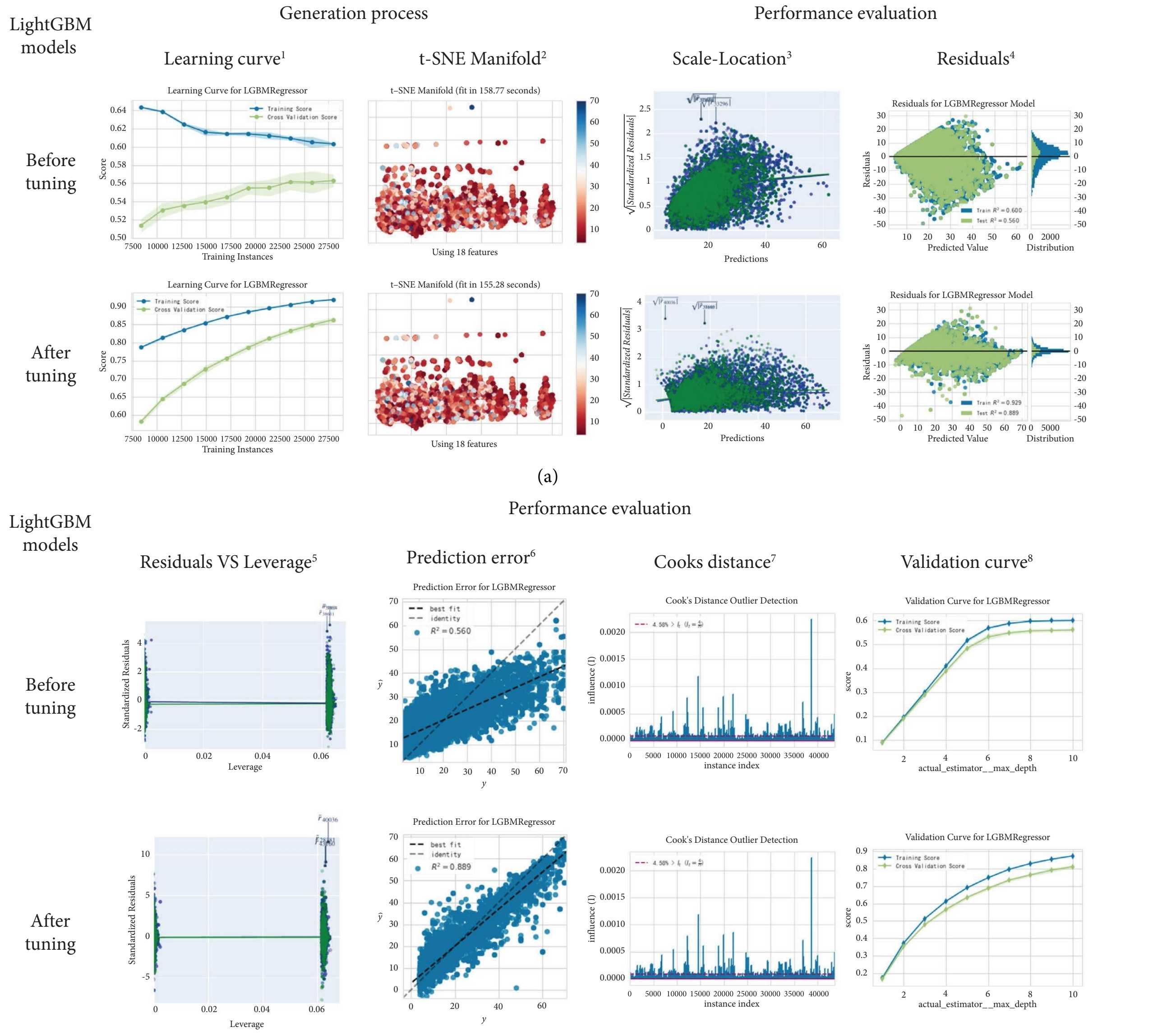
Table 4: 5-fold cross-validation results.

LightGBM models	5-fold	MAE	MSE	RMSE	RMSLE	MAPE	R^2
Before tuning	0	7.1199	91.5006	9.5656	0.4656	0.5120	0.5703
	1	7.0385	87.4810	9.3531	0.4639	0.5074	0.5666
	2	7.2264	91.9952	9.5914	0.4662	0.5127	0.5633
	3	7.1034	91.3756	9.5591	0.4672	0.5143	0.5571
	4	7.1751	91.9433	9.5887	0.4731	0.5237	0.5529
	Mean	**7.1327**	**90.8591**	**9.5316**	**0.4672**	**0.5140**	**0.5620**
	SD	0.0640	1.7062	0.0901	0.0031	0.0054	0.0063
After tuning	0	3.4854	28.9521	5.3807	0.2823	0.2369	0.8641
	1	3.2938	24.9931	4.9993	0.2774	0.2235	0.8762
	2	3.5073	28.8135	5.3678	0.2896	0.2359	0.8632
	3	3.4933	30.0568	5.4824	0.2920	0.2374	0.8543
	4	3.5205	29.4615	5.4278	0.2975	0.2468	0.8567
	Mean	**3.4601**	**28.4554**	**5.3316**	**0.2878**	**0.2361**	**0.8629**
	SD	0.0840	1.7855	0.1710	0.0071	0.0074	0.0076

We have used bold formatting to emphasize that the LightGBM models mean performance after 5-fold validation, in order to enhance the readability of the article.

500 m^2. One possible reason is that when the residential density reaches this threshold range, road traffic congestion can be severe during peak hours, potentially increasing the bus travel time. Similar findings were found for the relationship between origin residential density and bus travel time. Figure 6(e) also shows that bus travel time increases slightly when the origin residential density is approximately 85–100 pcs/500 m^2.

5.2.2. Land Diversity. Figure 5 shows that both the diversity of land at the destination and the origin are crucial factors affecting bus travel time. This can be attributed to the greater variety of land types (e.g., business, office, science and education, life service, and residence) that have a significant impact on bus travel time both at the origin and at the destination. Figures 6(b) and 6(f) indicate that the diversity of land at the destination and origin have different effects on bus travel time.



(a)

(b)

Figure 4: Predictive performance and generation process of LightGBM models. (1) Learning curve is a representation of the model's performance metrics against varying sizes of training data. It helps analyze the impact of dataset size on model accuracy and generalization. (2) t-Distributed stochastic neighbor embedding (t-SNE) is a technique used for dimensionality reduction and visualization of high-dimensional data in a lower-dimensional space. (3) Scale location helps assess the assumption of homoscedasticity, where the spread of residuals remains consistent across the range of predicted values. (4) Residuals represent the differences between observed and predicted values. (5) Leverage indicates the influence of individual data points on the regression model. (6) Prediction error refers to the discrepancy between predicted values and observed outcomes (R^2). (7) Cook distance is able to assess the influence of individual data points on the regression analysis. (8) Validation curve illustrates how the performance of a model changes with variations in a specific hyperparameter.

Travel time shows a nonlinear three-stage variation pattern: initially increasing, then decreasing, and finally increasing again as the degree of destination diversity rises (Figure 6(b)). In the previous stage, as destination diversity increases within the range of 0.5 to 0.6 pcs/500 m^2, the probability of residents opting for longer journeys also rises. This could be because when the industrial function of the destination area reaches a certain small scale, the comprehensive experience of small business offices and residential complexes attracts long travel. Then, with an improvement in destination diversity, travel time gradually decreases. The strongest inhibitory effect on travel time is observed when destination diversity is at 0.85 pcs/500 m^2. At this point, the travel purpose of residents near the medium-sized business office complex has been adequately fulfilled, and long travel is unnecessary. In the final stage, when destination diversity is greater than 0.9 pcs/500 m^2, travel time tends to increase again. This occurs because large business, office, and residential complexes attract greater traffic volumes, involving nearby and even whole city residents, leading to longer travel times.

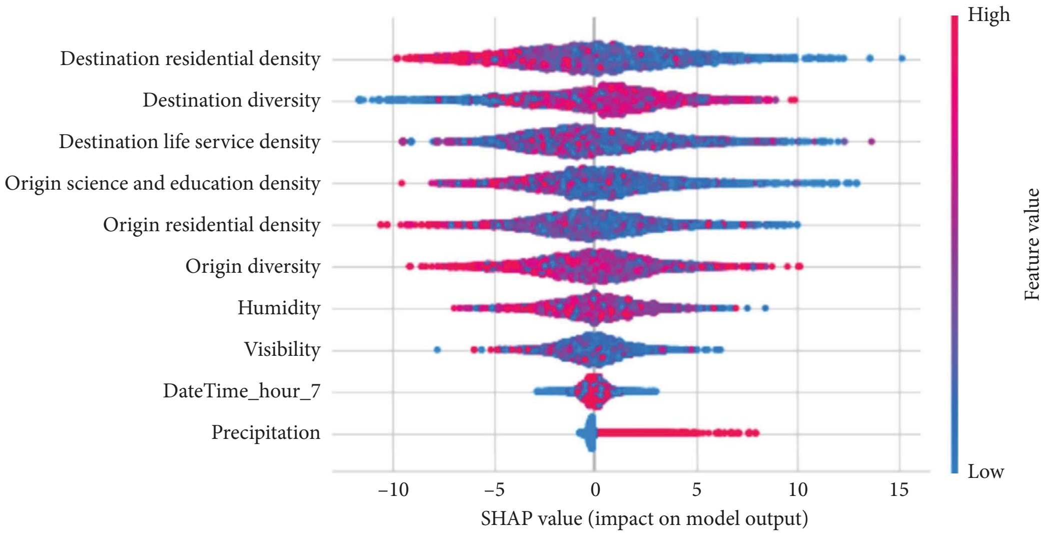


Figure 5: The SHAP values of variables.

However, Figure 6(f) shows that the nonlinear effect of origin diversity on bus travel time is significantly different. When the origin diversity exceeds 0.7 pcs/500 m^2, as the origin diversity increases, residents' bus travel time becomes increasingly shorter. Especially when origin diversity is between 0.8 and 0.9 pcs/500 m^2, higher origin diversity significantly shortens bus travel time. The result is consistent with the findings of Hu et al. [72] and Ji et al. [65], who also indicated that the higher the land mix, the shorter the electric vehicle and cycling time for residents.

5.2.3. Life Service Density. Figure 5 indicates that the destination life service density also has a slightly positive impact on bus travel time. One possible reason is that with the increase of the density of life services, the influence range of the area on the surrounding residents is increasing, and then the farther residents are attracted to the area, and the bus travel time of the residents is longer.

Furthermore, the promotion effect of destination life service density on bus travel time reaches its maximum when the density is around 350 pcs/500 m^2 (Figure 6(c)). Once the density of life services exceeds this threshold, the effect of this factor on bus travel time remains stable. It is possible that once the commercial service facilities reach a certain scale, the increase in the density of life service facilities will not lead to the continuous expansion of the attraction range of residents. The corresponding residents' bus travel time will not change greatly.

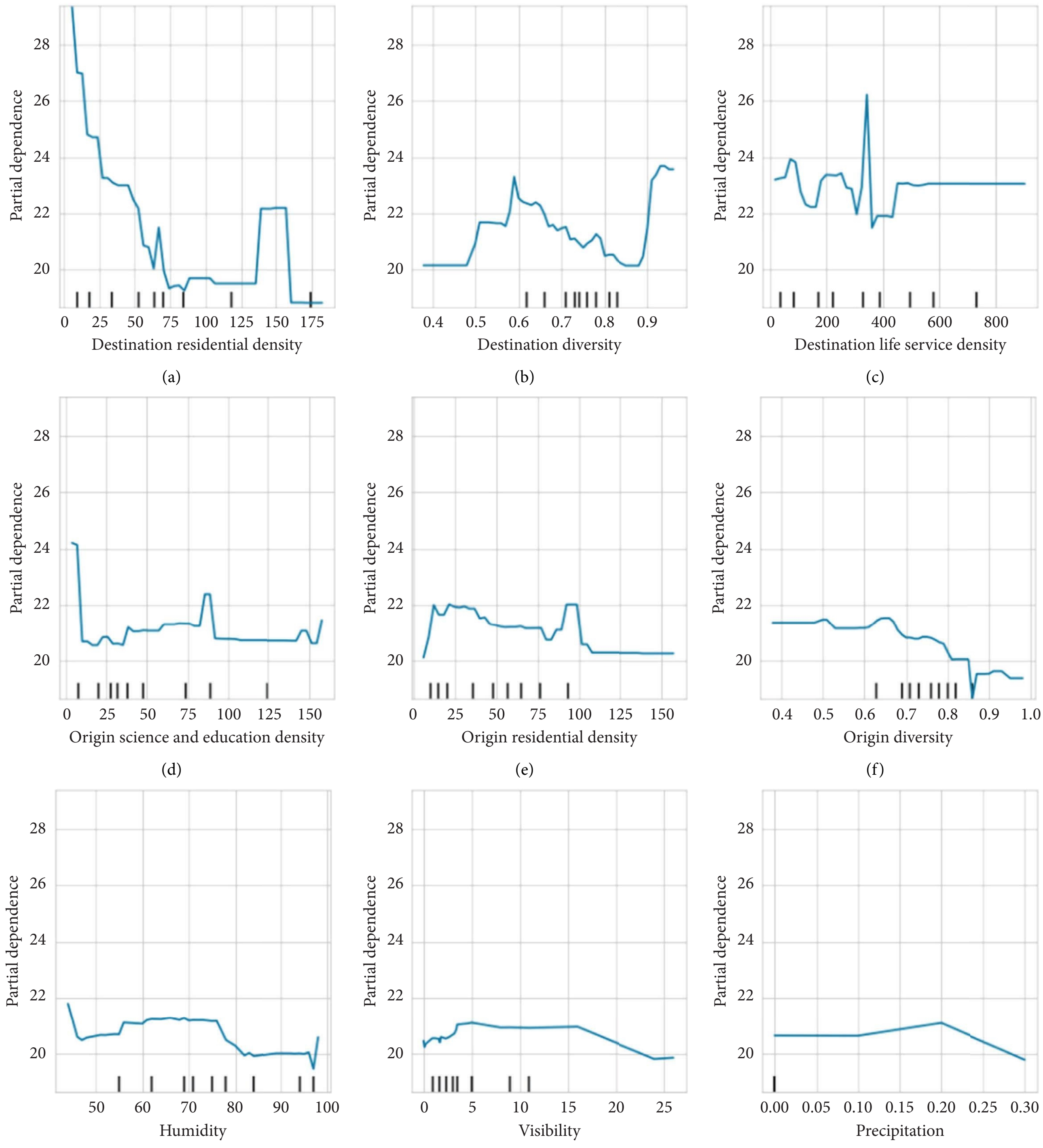
5.2.4. Science and Education Density. As shown in Figure 5, origin scientific and education density is an important factor that negatively affects bus travel time. One possible reason is that, with the continuous increase of science and education density at the origin, the supporting facilities of the region are more perfect. This allows passengers to complete their travel purposes in a relatively short time, such as school and shopping.

Figure 6(d) shows that when the origin scientific and education density is low, especially when the density of science education is below 10 pcs/500 m^2, residents travel longer by bus. In contrast, when the origin scientific and education density exceeds this threshold, residents' bus travel time is relatively short. This shows that the centralized use of land for science and education can shorten the bus travel time of residents to a great extent.

5.2.5. Humidity and Precipitation. Humidity and precipitation are significant weather variables that affect travel time. Among them, the positive impact of precipitation on travel time is quite evident. This kind of impact was expected, as various factors such as reduced visibility, slippery road surfaces, and increased traffic congestion, which are caused by increases in humidity or precipitation, collectively contribute to longer bus travel time. Similarly, Mathew and Pulugurtha [73] also found rainfall leads to reduced driving speeds.

Figures 6(g) and 6(i) both demonstrate the macroscopic nonlinear effects of humidity and precipitation on bus travel time, showing a pattern of increasing first and then decreasing. This is possible because increasing humidity and precipitation will result in longer travel times as mentioned above. However, when the humidity increases to 75% rh or the precipitation increases to 0.20 mm, the bad weather will make residents give up the bus and choose the taxi that can reach the destination point to point, which will potentially shorten the bus travel time of residents to some extent.

5.2.6. Visibility. We found visibility is also a crucial weather factor affecting residents' bus travel time, which is supported by Mathew and Pulugurtha [73]. Furthermore, this influence exhibits a slight inhibitory tendency. This is possible because low visibility significantly reduces the ability of drivers to see obstacles, road signs, and other vehicles, which can cause longer bus travel time.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 6: Nonlinear effects of significant factors.

As shown in Figure 6(h), the influence of visibility on bus travel time showed a nonlinear trend of increasing first and then decreasing. When the visibility is lower than 16 meters, the lower visibility will lead to a significantly longer bus travel time. However, when the visibility exceeds this threshold, the continuous increase in visibility will instead make the bus travel time significantly shorter.

5.2.7. Boarding Time. Among time series variables, the boarding time between 7 and 8 a.m. has the most important impact on passengers' bus travel time in Weinan. This is possible because the period from 7 to 8 a.m. is the peak time for Weinan residents. During this period, the road traffic congestion will be more serious, which will lead to a longer bus travel time for passengers.

6. Conclusion and Discussion

This study utilized a multisource big data approach to investigate the effect of the built environment and weather on passengers' bus travel time, in Weinan, a small- and medium-sized city in western China. Data source included bus smart card, bus operation information, bus station information, card swiping data, road information data, bus station data, POI data, and weather data. Through data fusion and mining, the study constructed a spatiotemporal database of bus travel in Weinan. Using the interpretable tuned-LightGBM model and SHAP values, the study analyzed the nonlinear threshold effects of the built environment and weather on bus travel time. It has been found that destination residential density, destination diversity, destination life service density, origin science and education density, origin residential density, origin diversity, humidity, visibility, boarding time between 7 and 8 a.m., and precipitation are important factors potentially affecting passengers' bus travel time. What's more, the built environment and weather factors have a threshold effect on the bus travel time.

The findings of this study offer valuable insights to bus companies in the development of bus routes. Firstly, the peak hour in the morning, particularly at 7 a.m., has a substantial positive impact on travel time, as commuting constitutes a significant portion of residents' travel. Accordingly, optimizing bus routes, providing customized signal timing schemes for intersections, and reducing bus travel intervals during peak hours may be very important measures that can significantly reduce bus travel time. Secondly, poor weather, such as higher humidity, precipitation, and low visibility, can also increase the bus travel time for individual passengers. Therefore, the traffic management department can issue relevant travel warning information according to the weather forecast to provide more accurate travel services for individual passengers, especially when a particular weather factor reaches its threshold.

Furthermore, the study's results provide valuable and novel evidence for urban land planning authorities with low-carbon goals to plan small- and medium-sized urban sites in the context of high-growth urbanization challenges. The application of POI indicators can promote comprehensive land-use and traffic planning in developing countries' small- and medium-sized cities. For instance, the comprehensive office, science, education, and business complex diversity near the residential gathering area should reach 0.85 pcs/500 m^2, as this satisfies the residents' travel purposes, significantly reducing travel time when the origin diversity is between 0.8 and 0.9 pcs/500 m^2 and the destination diversity is between 0.7 and 0.9 pcs/500 m^2. Urban land-use planning authorities should consider balancing occupancy and residence in their land-use planning. Areas with a residential density of 75–135 pcs/500 m^2 had the lowest total travel time cost. Thus, residential planning in the future should consider a density within this range to inhibit travel time.

This study has some limitations that should be taken into consideration. Firstly, the analysis is limited to the small- and medium-sized city of Weinan, and the behavioral characteristics of passengers in other urban contexts may vary. Secondly, the study relies on a dataset of only fifteen days, which may not provide a comprehensive understanding of the impact of weather and other factors on bus travel time. In addition, the use of POI amount without considering POI size does not fully represent the built environment. Therefore, future studies should include data from other cities to verify the transferability of the model and should aim to obtain more comprehensive bus-related data and POI data covering a longer period to more accurately investigate the impact of related factors on bus travel time.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest that could have appeared to influence the work reported in this paper.

Authors' Contributions

Xiaowei Li conceptualized the study, contributed to funding acquisition, and wrote, reviewed, and edited the article. Lanxin Shi contributed to data curation, provided software, and wrote the original drafts. Haotian Li performed validation and investigated the study. Qian Liu contributed to the methodology and investigated the study. Jun Chen performed supervision and investigated the study.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (No. 71701046), Shaanxi Natural Science Foundation Youth Project (No. 2017JQ5086), Shaanxi Education Department Special Science and Technology Project Science (No. 19JK0477), Foundation Project of Xi'an University of Architectural Science and Technology (Nos. ZR18021 and QN1711), and Scientific Research Program funded by the Research Institute of New Urbanization and Human Settlement in Shaanxi Province of XAUAT (No. 2023SCZH18).

References

[1] B. Büchel and F. Corman, "Modeling conditional dependencies for bus travel time estimation," *Physica A: Statistical Mechanics and Its Applications*, vol. 592, Article ID 126764, 2022.

[2] S. Banik, L. Vanajakshi, and D. M. Bullock, "Mapping of bus travel time to traffic stream travel time using econometric modeling," *Journal of Intelligent Transportation Systems*, vol. 26, no. 2, pp. 235–251, 2022.

[3] P. He, G. Jiang, S. K. Lam, and Y. Sun, "Learning heterogeneous traffic patterns for travel time prediction of bus journeys," *Information Sciences*, vol. 512, pp. 1394–1406, 2020.

[4] T. Nimpanomprasert, L. Xie, and N. Kliewer, "Comparing two hybrid neural network models to predict real-world bus travel time," *Transportation Research Procedia*, vol. 62, pp. 393–400, 2022.

[5] J. Ma, J. Chan, G. Ristanoski, S. Rajasegarar, and C. Leckie, "Bus travel time prediction with real-time traffic information," *Transportation Research Part C: Emerging Technologies*, vol. 105, pp. 536–549, 2019.

[6] M. Shao, C. Xie, T. Li, and L. Sun, "Influence of in-vehicle crowding on passenger travel time value: insights from bus transit in Shanghai, China," *International Journal of Transportation Science and Technology*, vol. 11, no. 4, pp. 665–677, 2022.

[7] V. J. M. Low, H. L. Khoo, and W. C. Khoo, "Quantifying bus travel time variability and identifying spatial and temporal factors using Burr distribution model," *International Journal of Transportation Science and Technology*, vol. 11, no. 3, pp. 563–577, 2022.

[8] E. Mazloumi, G. Currie, and G. Rose, "Using GPS data to gain insight into public transport travel time variability," *Journal of Transportation Engineering*, vol. 136, no. 7, pp. 623–631, 2010.

[9] Y. Cai, Y. Zhao, J. Yang, and C. Wang, "A bus passenger flow estimation method based on POI data and AFC data fusion," *Communications in Computer and Information Science*, vol. 1210, pp. 352–367, 2019.

[10] C. Ding, D. Wang, C. Liu, Y. Zhang, and J. Yang, "Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance," *Transportation Research Part A: Policy and Practice*, vol. 100, pp. 65–80, 2017.

[11] R. Ye and H. Titheridge, "Satisfaction with the commute: the role of travel mode choice, built environment and attitudes," *Transportation Research Part D: Transport and Environment*, vol. 52, pp. 535–547, 2017.

[12] N. S. Ngo, "Urban bus ridership, income, and extreme weather events," *Transportation Research Part D: Transport and Environment*, vol. 77, pp. 464–475, 2019.

[13] A. J. Kalkstein, M. Kuby, D. Gerrity, and J. J. Clancy, "An analysis of air mass effects on rail ridership in three US cities," *Journal of Transport Geography*, vol. 17, no. 3, pp. 198–207, 2009.

[14] P. Arana, S. Cabezudo, and M. Peñalba, "Influence of weather conditions on transit ridership: a statistical study using data from Smartcards," *Transportation Research Part A: Policy and Practice*, vol. 59, pp. 1–12, 2014.

[15] A. Singhal, C. Kamga, and A. Yazici, "Impact of weather on urban transit ridership," *Transportation Research Part A: Policy and Practice*, vol. 69, pp. 379–391, 2014.

[16] E. Dütschke, L. Engel, A. Theis, and D. Hanss, "Car driving, air travel or more sustainable transport? Socio-psychological factors in everyday mobility and long-distance leisure travel," *Travel Behaviour and Society*, vol. 28, pp. 115–127, 2022.

[17] S. Yoo, A. Cho, F. Salman, and Y. Yoshida, "Green paradox: factors affecting travel distances and fuel usages, evidence from Japanese survey," *Journal of Cleaner Production*, vol. 273, Article ID 122280, 2020.

[18] H. Yang, R. Zheng, X. Li, J. Huo, L. Yang, and T. Zhu, "Nonlinear and threshold effects of the built environment on e-scooter sharing ridership," *Journal of Transport Geography*, vol. 104, Article ID 103453, 2022.

[19] H. Yang, P. Luo, C. Li, G. Zhai, and A. G. Yeh, "Nonlinear effects of fare discounts and built environment on ridesplitting adoption rates," *Transportation Research Part A: Policy and Practice*, vol. 169, Article ID 103577, 2023.

[20] R. Ewing and R. Cervero, "Travel and the built environment: a meta-analysis," *Journal of the American Planning Association*, vol. 76, no. 3, pp. 265–294, 2010.

[21] D. Wang, S. Thunéll, U. Lindberg, L. Jiang, J. Trygg, and M. Tysklind, "Towards better process management in wastewater treatment plants: process analytics based on SHAP values for tree-based machine learning methods," *Journal of Environmental Management*, vol. 301, Article ID 113941, 2022.

[22] J. C. Krumm and L. N. Mummidi, *U.S. Patent No. 8*, U.S. Patent and Trademark Office, Washington, DC, USA, 2013.

[23] Z. Xie and J. Yan, "Kernel density estimation of traffic accidents in a network space," *Computers, Environment and Urban Systems*, vol. 32, no. 5, pp. 396–406, 2008.

[24] M. P. Kwan, "GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns," *Geografiska Annaler- Series B: Human Geography*, vol. 86, no. 4, pp. 267–280, 2004.

[25] G. McKenzie, K. Janowicz, S. Gao, and L. Gong, "How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest," *Computers, Environment and Urban Systems*, vol. 54, pp. 336–346, 2015.

[26] R. A. Becker, R. Caceres, K. Hanson et al., "A tale of one city: using cellular network data for urban planning," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 18–26, 2011.

[27] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui, "GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 831–840, New York, NY, USA, August 2014.

[28] Y. Yue, Y. Zhuang, A. G. Yeh, J. Y. Xie, C. L. Ma, and Q. Q. Li, "Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy," *International Journal of Geographical Information Science*, vol. 31, no. 4, pp. 658–675, 2017.

[29] C. Qian, Y. Zhou, Z. Ji, and Q. Feng, "The influence of the built environment of neighborhoods on residents' Low-Carbon travel mode," *Sustainability*, vol. 10, no. 3, p. 823, 2018.

[30] L. Yang, Q. Shen, and Z. Li, "Comparing travel mode and trip chain choices between holidays and weekdays," *Transportation Research Part A: Policy and Practice*, vol. 91, pp. 273–285, 2016.

[31] B. Sun, A. Ermagun, and B. Dan, "Built environmental impacts on commuting mode choice and distance: evidence from Shanghai," *Transportation Research Part D: Transport and Environment*, vol. 52, pp. 441–453, 2017.

[32] B. P. Loo, C. Chen, and E. T. Chan, "Rail-based transit-oriented development: lessons from New York City and Hong Kong," *Landscape and Urban Planning*, vol. 97, no. 3, pp. 202–212, 2010.

[33] S. S. Wells and B. G. Hutchinson, "Impact of commuter-rail services in Toronto region," *Journal of Transportation Engineering*, vol. 122, no. 4, pp. 270–275, 1996.

[34] K. Choi and M. Zhang, "The net effects of the built environment on household vehicle emissions: a case study of Austin, TX," *Transportation Research Part D: Transport and Environment*, vol. 50, pp. 254–268, 2017.

[35] H. Bi, Z. Ye, and H. Zhu, "Data-driven analysis of weather impacts on urban traffic conditions at the city level," *Urban Climate*, vol. 41, Article ID 101065, 2022.

[36] L. Böcker, J. Prillwitz, and M. Dijst, "Climate change impacts on mode choices and travelled distances: a comparison of present with 2050 weather conditions for the Randstad Holland," *Journal of Transport Geography*, vol. 28, pp. 176–185, 2013.

[37] L. Ma, H. Xiong, Z. Wang, and K. Xie, "Impact of weather conditions on middle school students' commute mode choices: empirical findings from Beijing, China," *Transportation Research Part D: Transport and Environment*, vol. 68, pp. 39–51, 2019.

[38] Z. Chen and Y. Wang, "Impacts of severe weather events on high-speed rail and aviation delays," *Transportation Research Part D: Transport and Environment*, vol. 69, pp. 168–183, 2019.

[39] C. Kamga and M. A. Yazıcı, "Temporal and weather related variation patterns of urban travel time: considerations and caveats for value of travel time, value of variability, and mode choice studies," *Transportation Research Part C: Emerging Technologies*, vol. 45, pp. 4–16, 2014.

[40] I. Tsapakis, T. Cheng, and A. Bolbol, "Impact of weather conditions on macroscopic urban travel times," *Journal of Transport Geography*, vol. 28, pp. 204–211, 2013.

[41] H. Li, Q. Wang, and W. Xiong, "New model of travel-time prediction considering weather conditions: case study of urban expressway," *Journal of Transportation Engineering Part A: Systems*, vol. 147, no. 3, Article ID 04020161, 2021.

[42] H. A. Aaheim and K. E. Hauge, "Impacts of climate change on travel habits: a national assessment based on individual choices," *CICERO Report*, 2005.

[43] A. F. Abidin and M. Kolberg, "Towards improved vehicle arrival time prediction in public transportation: integrating SUMO and Kalman filter models," in *Proceedings of the 2015 17th UKSim-AMSS International Conference on Modelling and Simulation (UKSim)*, pp. 147–152, IEEE, Cambridge, United Kingdom, March 2015.

[44] Z. He, H. Yu, Y. Du, and J. Wang, "SVM based multi-index evaluation for bus arrival time prediction," in *Proceedings of the 2013 International Conference on ICT Convergence (ICTC)*, pp. 86–90, IEEE, Jeju Island, South Korea, October 2013.

[45] M. Chen, J. Yaw, S. I. Chien, and X. Liu, "Using automatic passenger counter data in bus arrival time prediction," *Journal of Advanced Transportation*, vol. 41, no. 3, pp. 267–283, 2007.

[46] Y. Zhang and H. Ge, "Freeway travel time prediction using Takagi–Sugeno–Kang fuzzy neural network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 28, no. 8, pp. 594–603, 2013.

[47] X. Ran, Z. Shan, Y. Shi, and C. Lin, "Short-term travel time prediction: a spatiotemporal deep learning approach," *International Journal of Information Technology and Decision Making*, vol. 18, no. 04, pp. 1087–1111, 2019.

[48] B. Gupta, S. Awasthi, R. Gupta et al., "Taxi travel time prediction using ensemble-based random forest and gradient boosting model," in *Advances in Big Data and Cloud Computing*, pp. 63–78, Springer, Singapore, 2018.

[49] P. He, G. Jiang, S. K. Lam, and D. Tang, "Travel-time prediction of bus journey with multiple bus trips," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4192–4205, 2019.

[50] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, CA, USA, August 2016.

[51] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser et al., "Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[52] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021.

[53] X. Wen, Y. Xie, L. Wu, and L. Jiang, "Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP," *Accident Analysis & Prevention*, vol. 159, Article ID 106261, 2021.

[54] W. Zhang, D. Lu, Y. Zhao, X. Luo, and J. Yin, "Incorporating polycentric development and neighborhood life-circle planning for reducing driving in Beijing: Nonlinear and threshold analysis," *Cities*, vol. 121, Article ID 103488, 2022.

[55] H. Yang, G. Zhai, L. Yang, and K. Xie, "How does the suspension of ride-sourcing affect the transportation system and environment?" *Transportation Research Part D: Transport and Environment*, vol. 102, Article ID 103131, 2022.

[56] C. Jun and Y. Dongyuan, "Estimating smart card commuters origin-destination distribution based on APTS data," *Journal of Transportation Systems Engineering and Information Technology*, vol. 13, no. 4, pp. 47–53, 2013.

[57] X. Chen, X. Dai, and Q. Chen, "Approach on the information collection, analysis and application of bus intelligent card," *China Civil Engineering Journal*, vol. 2, pp. 105–110, 2004.

[58] J. Chen and D. Yang, "A method for determining bus IC card passenger boarding station based on intelligent scheduling data," *Journal of Transportation Systems Engineering and Information Technology*, vol. 13, pp. 76–80, 2013.

[59] J. Zuo, Q. Wang, and J. Chen, "Spatio-temporal characteristics fusion algorithm of urban bus GPS data and IC card data," *Journal of Transport Information and Safety*, vol. 39, pp. 101–108, 2021.

[60] J. Chen, Y. Lv, and M. Cui, "A method to judge the stop of bus IC card passengers based on travel mode," *Journal of Xi'an University of Architecture and Technology*, vol. 50, pp. 23–29, 2018.

[61] G. Ke, Q. Meng, T. Finley et al., "Lightgbm: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[62] Z. Pan, S. Fang, and H. Wang, "LightGBM technique and differential evolution algorithm-based multi-objective optimization design of DS-APMM," *IEEE Transactions on Energy Conversion*, vol. 36, no. 1, pp. 441–455, 2021.

[63] J. Ren, Z. P. Yu, G. L. Gao, G. K. Yu, and J. Yu, "A CNN-LSTM-LightGBM based short-term wind power prediction method based on attention mechanism," *Energy Reports*, vol. 8, pp. 437–443, 2022.

[64] L. S. Shapley, "A value for n-person games," *Classics in Game Theory*, vol. 69, 1997.

[65] S. Ji, X. Wang, T. Lyu et al., "Understanding cycling distance according to the prediction of the XGBoost and the interpretation of SHAP: a non-linear and interaction effect analysis," *Journal of Transport Geography*, vol. 103, Article ID 103414, 2022.

[66] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[67] K. Li, H. Xu, and X. Liu, "Analysis and visualization of accidents severity based on LightGBM-TPE," *Chaos, Solitons & Fractals*, vol. 157, Article ID 111987, 2022.

[68] S. M. Lundberg, G. G. Erion, and S. I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, https://arxiv.org/abs/1802.03888.

[69] X. Li, L. Shi, J. Tang et al., "Determinants of passengers' ticketing channel choice in rail transit systems: new evidence of e-payment behaviors from Xi'an, China," *Transport Policy*, vol. 140, pp. 30–41, 2023.

[70] D. H. Vu, K. M. Muttaqi, and A. P. Agalgaonkar, "A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables," *Applied Energy*, vol. 140, pp. 385–394, 2015.

[71] J. Feng, M. Dijst, J. Prillwitz, and B. Wissink, "Travel time and distance in international perspective: a comparison between Nanjing (China) and the Randstad (The Netherlands)," *Urban Studies*, vol. 50, no. 14, pp. 2993–3010, 2013.

[72] X. Hu, Y. Cao, T. Peng, R. Gao, and G. Dai, "Nonlinear influence model of built environment of residential area on electric vehicle miles traveled," *World Electric Vehicle Journal*, vol. 12, no. 4, p. 247, 2021.

[73] S. Mathew and S. S. Pulugurtha, "Quantifying the effect of rainfall and visibility conditions on road traffic travel time reliability," *Weather, Climate, and Society*, vol. 14, no. 2, pp. 507–519, 2022.