land

MDPI

Article

Nonlinear Relationship of Multi-Source Land Use Features with Temporal Travel Distances at Subway Station Level: Empirical Study from Xi'an City

Peikun Li [1], Quantao Yang [2,*] and Wenbo Lu [3]

[1] Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Ministry of Transport, Beijing Jiaotong University, Beijing 100044, China; lipeikun@bjtu.edu.cn
[2] Department of Public Security, Shaanxi Police College, Xi'an 710021, China
[3] School of Transportation, Southeast University, Nanjing 214135, China; wenbo1996@seu.edu.cn
* Correspondence: yqt@chd.edu.cn

Abstract: The operation of the subway system necessitates a comprehensive understanding of passenger flow characteristics at station locations, as well as a keen awareness of the average travel distances at these stations. Moreover, the travel distances at the station level bear a direct relationship with the built environment composed of land use characteristics within the station's catchment area. To this end, we selected the land use features within an 800 m radius of the station (land use area, distribution of points of interest, and the surrounding living environment) as the influencing factors, with the travel distances at peak hours on the subway network in Xi'an as the research subject. An improved SSA-XGBOOST-SHAP interpretable machine learning framework was established. The research findings demonstrate that the proposed enhanced model outperforms traditional machine learning or linear regression methods in terms of R-squared, MAE, and RMSE. Furthermore, the distance from the city center, road network density, the number of public transit routes, and the land use mix have a pronounced influence on travel distances, reflecting the significant impact that mature built environments can have on passenger attraction. Additionally, the analysis reveals a notable nonlinear relationship and threshold effect between the built environment variables comprising land use and the travel distances during peak hours. The research results provide data-driven support for operational strategy management and line capacity optimization, as well as theoretical underpinnings for enhancing the efficiency and sustainability of the entire subway system.

Keywords: land use area; travel distance; built environment; nonlinear relationship; interpretable machine learning; SHAP value

check for updates

Citation: Li, P.; Yang, Q.; Lu, W. Nonlinear Relationship of Multi-Source Land Use Features with Temporal Travel Distances at Subway Station Level: Empirical Study from Xi'an City. *Land* **2024**, *13*, 1021. https://doi.org/10.3390/land13071021

Academic Editors: Tiziana Campisi, Georgios Georgiadis and Elena Cocuzza

Received: 10 June 2024
Revised: 3 July 2024
Accepted: 4 July 2024
Published: 8 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As a form of public transportation that harmoniously integrates environmental sustainability, reliability, and safety, subway travel has garnered widespread acclaim in large and medium-sized cities globally. In recent years, China has particularly witnessed a rapid proliferation of newly inaugurated subway lines and stations. Transit-oriented development (TOD) is gradually emerging as a pivotal trend in sustainable and green urban planning [1]. With the expansion of operational lines and stations, a pressing issue confronting cities and service operators is how to effectively enhance the efficient operation of subway lines while maximizing passenger satisfaction.

The built environment in the context of transportation refers to the overarching state of surrounding buildings, points of interest (POIs), and living conditions at a specific location. These built environments should cater to the needs of individuals during their travels, providing a convenient, safe, and comfortable setting. This encompasses various elements such as the ease of transportation, sidewalks, greenery, and public service facilities around subway stations [2,3]. The quality of the built environment around subway stations directly

influences the city's image and residents' quality of life. Therefore, analyzing the intricate relationship between the built environment and station passenger flow characteristics is crucial for advancing the sustainable development of TOD. The built environment around subways impacts travel characteristics, and understanding this complex relationship can assist in precisely managing passenger flow characteristics during subway construction planning [4].

Currently, research on the built environment and passenger flow characteristics primarily begins with exploring the relationship between subway passenger flow and the built environment. This includes investigating the nonlinear relationships between station-to-station passenger flow and factors such as the built environment, station types, and demographic and economic characteristics [5]. The effects of density and accessibility indicators in residential built environments on origin and destination (OD) passenger flow are particularly significant [6,7]. Employment density and public transportation accessibility indicators around station areas also significantly contribute to passenger flow [8]. Additionally, the characteristics of land use around stations and their influence on inbound passenger flow during different time periods have been extensively analyzed. The travel patterns of special groups (the older adults, disabled, etc.) on subways have also garnered attention, revealing that land use mix has a smaller impact on the travel of these groups [9].

However, for actual subway operations, it is essential not only to control passenger flow but also to understand the distance between passengers' travel destinations and origins at specific times. This understanding can help optimize strategies such as adjusting departure frequencies and routing plans to reduce transportation capacity waste and achieve high operational efficiency [10,11]. Specifically, studying the relationship between travel distance and the built environment can elucidate passengers' motivations and considerations for choosing subways and the OD distribution characteristics of passenger flow on subway lines. Different built environments may have varying attractions for passengers, influencing their choice of travel distance by subway. This insight aids urban planners and transportation management departments in making informed adjustments to schedules, facilities, and services based on station passenger flow [12]. Furthermore, based on the characteristics of travel distances and built environments, subway operators can increase trips in high-demand and long-distance areas to meet passenger needs and optimize capacity allocation. Finally, during the planning stages of lines and stations, analyzing the relationship between subway travel distances and the built environment can provide valuable insights for optimizing and enhancing subway line planning [13].

In existing research on travel distances, the travel distance of motor vehicles is correlated to varying degrees with the density of the surrounding environment at origin and destination points, the mix of land use, and the distance from the city center. Non-motorized travel distances are associated with street structure and accessibility indices in the built environment [14,15]. Research on subway travel has found a ring-like aggregation pattern where travel distances gradually increase from the city center outward [16,17]. However, research on the differences in travel distances during different times of the day, considering the varying functional characteristics of station areas, remains underexplored.

Methodologically, there are various approaches to analyzing the relationship between the built environment and travel characteristics. These include quantile regression models based on linear regression assumptions [18], logit models [14], geographically weighted regression (GWR) models, and their variants [17]. With the advantage of not requiring adherence to linear relationships between variables, machine learning methods have been increasingly applied in recent years to such problems. These methods include Gradient Boosting Decision Trees (GBDT) [5], eXtreme Gradient Boosting (XGBOOST) [1], and Random Forests [19]. Although machine learning has achieved better fitting results, the selection of optimal hyperparameters for tree-based models significantly affects the characteristics of the fitting results. Therefore, reasonable hyperparameter selection is a key issue that requires attention. Additionally, the interpretability of machine learning fitting results needs further in-depth analysis to serve real-world operational production.

Addressing the issues described above, this study focuses on the Xi'an subway system. Based on station-level smart card data, it analyzes the relationship and threshold effects between subway travel distances and selected built environment variables during weekday morning and evening peak periods. The innovations and contributions of this study are as follows:

(1) Proposing a machine learning method that combines the Salp Swarm Algorithm (SSA) with XGBOOST to fit the complex relationship between the built environment and travel distances. This method addresses the common issue of relying on experience for hyperparameter selection in machine learning methods and captures the nonlinear relationship and threshold effects between built environment variables and dependent variables more effectively.
(2) Investigating the changes in travel distances during weekday morning and evening peak periods due to the different functional attributes of stations at different times. Using SHAP (SHapley Additive exPlanations) attribution analysis, this study explains the relationship between built environment variables and travel distances, analyzing the contribution of different types of built environment factors to station travel distances.
(3) Using Xi'an as a case study, this research applies the above methods to explain the distribution of travel distance differences caused by spatiotemporal heterogeneity at the station level, providing an analytical framework for similar issues in other urban subway networks.

2. Related Work

TOD has significantly increased residents' reliance on subways. However, there is still limited research on the spatiotemporal distribution characteristics of subway travel distance differences based on massive smart card data. Smart card data enable precise analysis of passenger flow origins and destinations. Researchers can utilize rich spatiotemporal travel records to uncover the dynamic behavior of public transit passengers. Compared to changes in passenger flow, the evolution of travel distances provides a unique perspective for understanding the impact of pandemics on subway networks and can guide adjustments in subway network operation strategies [5].

In fact, urban morphology is closely related to residents' travel behavior, including travel time, peak travel formation time, and travel distance [6,20]. Generally, the built environment determines residents' travel demand, leading to significant differences in travel distances between TOD communities and non-TOD communities [21,22]. Previous descriptive survey analyses have found that increasing land use mix can enhance the proximity of potential destinations within specific areas, thereby shortening travel distances [23–25]. However, the spatial distribution of average subway travel distances is often uneven. For example, in Nanjing, China, travel distances exhibit concentric diffusion from the central city to the surrounding areas, which is closely related to the distance of stations from the city center, land use intensity, and employment/entertainment ratio based on POI [18]. Residents in areas with well-developed transit environments have more alternative travel routes, thus effectively shortening travel distances [14].

Research on temporal distribution differences has found that station-to-station passenger volume varies across different time periods due to the built environment characteristics of origins and destinations. Specifically, during the morning peak period, origin stations significantly contribute to attracting passenger flow across the network, while during the evening peak period, destination stations contribute more to passenger volume [26,27]. A study on commuting distances and housing in Beijing shows that the average commuting distance is 7.2 km, and the commuting distance in Beijing is positively correlated with the distance from the city center, exhibiting a ring-shaped aggregation pattern [16,17].

Furthermore, there is still insufficient research on changes in subway travel distances under the impact of the COVID-19 pandemic. In the long run, the pandemic has indeed altered residents' travel habits, including travel times, choice of transportation modes, and

travel duration [28]. For instance, in the post-pandemic period, cross-border travel in China is predominantly short-distance, and medium- to long-distance travel has recovered more slowly, which is closely related to the increased risk of infection associated with longer travel times [29]. During the pandemic, teenagers significantly reduced long-distance travel and showed a marked decline in their willingness to use public transportation [30]. Meanwhile, the proportions of walking and cycling in short-distance travel have significantly increased in the post-pandemic period [31].

In terms of research methods, current approaches based on linear regression assume a linear relationship between built environment factors and passenger flow. However, studies have shown that the relationship between built environment variables and passenger flow is complex and nonlinear, making simple linear regression assumptions unsuitable for fitting systematic multi-station passenger flow relationships [32–34]. With the superior ability of machine learning methods to handle nonlinearity and threshold determination between variables, tree-based supervised machine learning methods have been applied to address these issues [35–38].

In this regard, this study focuses on Xi'an, using smart card big data and multiple built environment characteristics to construct a nonlinear relationship between multi-source built environment variables and subway travel distances during weekday morning and evening peak periods using an improved SSA-XGBOOST method. SHAP attribution analysis, which uses Shapley values from game theory as explanatory metrics, compensates for the poor interpretability of tree-based models (such as decision trees and random forests) by offering strong global and local interpretability, fair distribution of variable contributions, and excellent visualization effects. Therefore, SHAP attribution analysis is introduced to explain the feedback mechanisms of different variables. The research results can provide important theoretical and data support for macro-level subway construction planning in subway operations.

3. Data

3.1. Research Area

The study area selected for this research is Xi'an, Shaanxi Province, China. Xi'an is located in the central-western region of China and is a city with a permanent population exceeding ten million. It is also one of the most popular tourist cities in China. The urban development of Xi'an is concentrically distributed, with areas closer to the city center being more developed, while those farther from the center are less developed. Figure 1 shows the geographical distribution of Xi'an as well as the distribution of its subway network, which consists of 4 lines and 87 stations. According to statistics, the passenger intensity of Xi'an's subway system has long ranked among the top ten cities in the country.

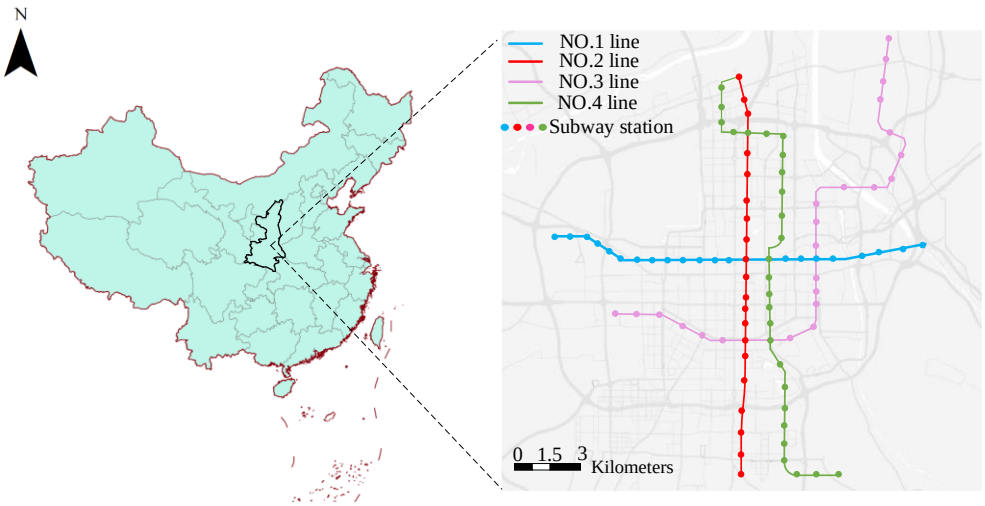


Figure 1. Study area and subway station distribution.

3.2. Built Environment Variables

When selecting the built environment around subway stations, it is first necessary to determine the spatial scope of the study. Based on extensive research, it is generally considered that the effective attraction distance around a station is within a 15 min walking distance. The radius of influence is often used to measure the range of passenger attraction, with most studies focusing on catchment area coverage between 500 m and 1000 m. Considering previous research results on the attraction range of Xi'an's subway, an 800 m catchment area was deemed appropriate for this study [39,40]. In selecting built environment factors, the 5D principles are usually followed, which include density, design, diversity, distance, and destination accessibility. Therefore, the environmental variables chosen for this study also fully reflect these characteristics. Therefore, the built environment around stations can often be summarized into the following three basic attributes.

First, land use directly determines the activity attributes of travelers within a certain range around the station. The area size influences the proportion of the population engaged in activities such as living, consuming, working, studying, and leisure in the region. Therefore, land use intensity has a significant impact on travel activities.

Second, the distribution of POIs complements the distribution of land use properties. POIs represent specific locations in geographic space that people are interested in or need to visit, such as shops, restaurants, and entertainment venues. Different types of POIs attract different types of passengers, thus influencing the demand for subway travel and destination choices.

Third, station-specific attributes include the distance from the city center and the number of station entrances and exits. Nearby attributes include the number of residential houses, average housing prices, bus line density, and road density. These attributes affect the functional properties of the station and the convenience and efficiency of subway travel. For example, stations closer to the city center are typically more attractive to passengers because they offer faster, direct travel options, reducing transfer or walking time. The number of station entrances and exits affects passenger entry and exit efficiency and convenience. The number of residential houses around the station reflects the density and scale of the local population; a higher number of residential houses implies more potential passengers, as nearby residents are more likely to use the subway for commuting and travel. Average housing prices reflect the economic status and social characteristics of the area around the station. Higher average housing prices indicate that residents may have more purchasing power and higher income levels, which might lead to a preference for non-public transport modes. Bus line density and road density reflect the availability of more bus routes and other connecting tools for passengers, potentially impacting subway travel. The definitions and numerical ranges of various variables are detailed in Table 1. When iterating models on built environment variables, it is imperative to scrutinize the multicollinearity among the variables. This is typically assessed using the Variance Inflation Factor (VIF), which measures the ratio of the variance of the regression coefficient estimates when assuming no linear correlation among the independent variables, compared to the actual variance. A commonly utilized threshold for interpretation is 10. A VIF value below 10 indicates an absence of multicollinearity among the variables [41]. In this study, calculations revealed VIF values for the selected variables to be below 5, signifying the absence of significant multicollinearity among them.

Table 1. Built environment variable information.

Variable	Description	Mean	STD	Unit
LU_{comm}	Area of commercial land around station	120.018	107.658	Acre
LU_{recr}	Area of recreational land around station	19.805	35.065	Acre
LU_{resi}	Area of residential land around station	626.096	288.492	Acre
LU_{work}	Area of working land around station	61.441	56.512	Acre
LU_{seco}	Area of secondary school land around station	6.634	6.508	Acre

Table 1. *Cont.*

Variable	Description	Mean	STD	Unit
LU_{univ}	Area of university land around station	14.640	29.219	Acre
LU_{mixt}	Degree of land use mix around station	0.343	0.232	
POI_{cond}	Number of condominiums around station	56.287	55.451	
$POI_{spo\&lei}$	Number of sports and leisure centers around station	32.333	33.693	
POI_{scen}	Number of scenic spots around station	6.345	10.959	
POI_{cult}	Number of cultural services around station	69.459	76.169	
POI_{hote}	Number of hotel residences around station	51.414	61.134	
POI_{shop}	Number of shopping points around station	8.667	8.537	
POI_{cate}	Number of catering services around station	179.655	154.154	
POI_{comp}	Number of companies around station	103.436	132.414	
$SURR_{road}$	Density of road network around station	3.368	1.504	km/km^2
$SURR_{bus}$	Number of bus lines around station	89.977	47.165	
$SURR_{price}$	House prices around station	11,149.7	2697.441	CNY
$SURR_{hous}$	Number of houses around station	12,720.334	11,057.133	
$STAT_{dist}$	Distance from station to city center	7.494	4.035	km

3.3. Smart Card Data

The urban rail transit smart card data for this study were sourced from the stored records of Xi'an Metro's operating company. Given the global impact of the COVID-19 pandemic in 2020, passenger travel patterns were also disrupted to varying degrees. This study selected a typical week in October 2020, after Xi'an had experienced the effects of the pandemic, with residents' travel habits now restored to a new normal. This research aims to investigate passenger travel distances under the new normal conditions post-pandemic. Focusing on the entire day, this study specifically examines peak travel times during morning and evening rush hours. Based on the passenger flow characteristics in Xi'an, the morning peak period was set from 8:00 to 9:00, and the evening peak period from 17:30 to 18:30 [40]. As shown in Table 2, the backend records of a single card swipe include information such as the trip ID (used to match entry and exit stations to form a single trip chain), swipe time, entry and exit types, service gates, and lines. This information needs to be cleaned according to certain rules to facilitate passenger flow feature statistics at specific time granularities.

Table 2. Example of AFC system swipe record.

ID	Swipe Time	Type of Entry/Exit Station	Line	Station
1F022702	20201123062152	1	2	9860
1F022711	20201123062217	1	3	1600
1F024108	20201123083509	2	1	3300
1F02531A	20171123062105	2	3	258

By combining the Python programming language and the MongoDB database, we organized the AFC (Automatic Fare Collection) card swipe data and completed the statistical analysis of passenger flow data for various rail stations at different times. During the statistical process, we excluded data that could not form complete travel chains or had obviously unreasonable entry and exit records. Therefore, to ensure the accuracy and validity of the data, the raw AFC system data were filtered and cleaned, and an effective OD matrix of passenger trips was extracted for subsequent research analysis.

4. Methodology

4.1. XGBOOST Algorithm

Tree-based XGBOOST is highly effective in nonlinear fitting due to its flexible parameter tuning and strong noise resistance. XGBOOST is a Boosting iterative method but includes an L2 regularization term. The objective function of XGBOOST is as follows:

$$obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

$$\sum_{i=1}^{n} l(y_i, \hat{y}_i) = \sum_{i=1}^{n} \left\{ y_i - \left[\hat{y}_i{}^{(t-1)} + f_t(x_i) \right] \right\}^2 \quad (2)$$

The regularization term is calculated as follows:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (3)$$

In the above equations, y_i is the actual travel distance for station i, $\hat{y}_i$ is the fitted travel distance for station i, and $l(y_i, \hat{y}_i)$ is the loss function used to measure the difference between the observed and fitted values. $\Omega(f_k)$ is the regularization term used to control the model complexity. T is the number of leaf nodes. ω is the average target variable of the leaf nodes in the training sample. γ is the parameter for adjusting the number of leaf nodes. λ is the parameter for preventing the average value of the target variable in the leaf nodes from being too large. $\hat{y}_i{}^{(t)}$ represents the fitted value of sample i after t iterations.

4.2. SSA Algorithm

Although XGBOOST has high accuracy in handling multi-class samples and can address overfitting issues, it is a complex algorithm that is very sensitive to parameter selection and requires careful tuning to achieve optimal performance. When solving the XGBOOST model, the setting of model hyperparameters can significantly affect the model's fitting effectiveness. The SSA is a recently proposed optimization algorithm inspired by the biological behavior of sea cucumbers. Its principle involves simulating the foraging and predator avoidance behaviors of sea cucumbers to search for the optimal solution.

SSA is typically used for global optimization problems and can be employed to optimize the parameters of functions. It features global search capabilities and rapid convergence, allowing it to find optimal solutions within the search space. For hyperparameter tuning of the XGBOOST model, the SSA algorithm can be used to search for the optimal combination of hyperparameters. The specific steps are as follows:

- Define the problem's fitness function: Train the XGBOOST model based on the hyperparameter combinations and evaluate the model performance using methods such as cross-validation. Use the evaluation metrics as the value of the fitness function.
- Initialize the population: generate an initial set of hyperparameter combinations as the population.
- Iterative search: in each generation, evaluate the population based on the fitness function and select individuals with higher fitness.
- Generate new individuals: based on the selected individuals, use the operations of the Sea Cucumber Optimization Algorithm (such as foraging and predator avoidance) to generate new individuals.
- Update the population: update the population based on the newly generated individuals.
- Termination condition: stop the search when the predetermined number of iterations is reached or the stopping condition is met, and return the hyperparameter combination with the highest fitness.

4.3. SHAP Attribution Analysis

The foundational theory behind the SHAP interpreter is game theory. The core idea is to calculate the marginal contribution of features to the model's output, providing both local and global interpretability. SHAP values can measure the impact of features on the prediction target by calculating their marginal contribution to the model's output across all feature combinations. The mean absolute SHAP value of a feature reflects its importance. The relationship between SHAP values and feature values can indicate the direction of the feature's impact. When the SHAP value is positively correlated with the feature value, it suggests that the feature has a positive impact on the prediction target. Conversely, a negative correlation indicates a negative impact. The formula for SHAP values is as follows:

$$\phi_i(x) = \sum_{S \subseteq \{1,2,\cdots,M\} \setminus \{i\}} \frac{P(S \cup \{i\})}{P(S)} [f(x/S \cup \{i\} - f(x/S)] \quad (4)$$

In the formula, $\phi_i(x)$ represents the SHAP value of the i-th built environment feature. x is the input sample's feature vector. M is the total number of features. S is a subset of features. $P(S)$ represents the number of all possible permutations of the feature subset S, $f(x/S)$ represents the model's fitted regression result given the feature subset S.

5. Results

5.1. Descriptive Analysis

Figure 2 shows the distribution characteristics of the average travel distance across stations during two peak periods. Most stations have travel distances ranging from 9 to 11 km. An interesting phenomenon is that the number of stations with longer travel distances is higher during the morning peak period compared to the evening peak period. This indicates that there are certain differences in travel distances between the two time periods, warranting detailed analysis.

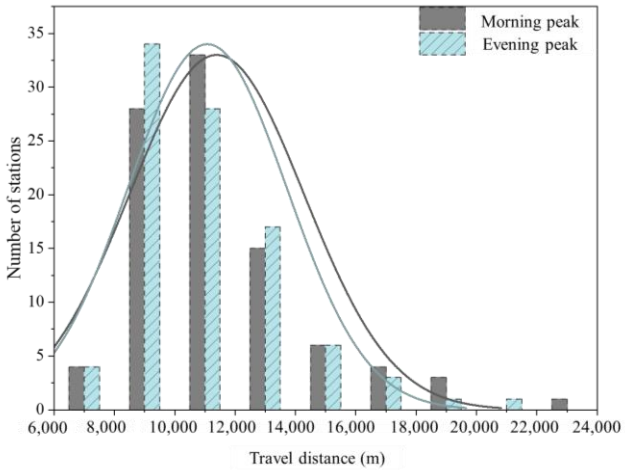


Figure 2. Distribution of station-level average travel distances.

5.2. Hyperparametric Results of SSA Method

Through the iterative calculation process of hyperparameter selection using the SSA algorithm, the optimal combination of hyperparameters for solving the XGBOOST model was determined. The root mean square error (RMSE) was chosen as the training objective of the model to identify the best hyperparameter combination. Table 3 provides the selection range and the optimal values for the XGBOOST hyperparameters.

Table 3. Hyperparameter selection results.

Hyperparameterization	Explanation	Range of Values	Preferred Value
max_depth	Maximum depth of the tree	[1, 10]	4
learning_rate	Learning rate	[0.01, 0.1]	0.09
subsample	Subsampling rate	[0.5, 0.9]	0.75
colsample_bytree	Column sampling rate	[0.5, 0.9]	0.62
n_estimators	Number of regression trees	[100, 200]	154
Gamma	Leaf node splitting threshold	[0, 5]	0

5.3. Significance Contribution of Built Environment Factors

Table 4 lists the top ten built environment variables in terms of relative importance during the morning and evening peak periods, with cumulative contributions exceeding 96%. The information in Table 4 indicates that the $STAT_{dist}$ has a very high contribution

to travel distance in both periods, especially during the evening peak period, where the contribution reaches 66.339%. In fact, according to relevant studies, the travel distance of stations throughout the day also gradually increases with the $STAT_{dist}$, showing a strong correlation, which aligns with common understanding.

Secondly, during the morning peak period, the contribution of POI_{cond} reaches 27.997%, highlighting the significant impact of such stations on travel distance during this time. Additionally, it can be observed that $SURR_{road}$, LU_{mixt}, and $SURR_{bus}$ also have a relatively noticeable impact on travel distance. This reflects that areas with a more mature built environment indeed have a significant influence on passenger travel. LU_{seco} contributes to travel distance in both time periods, indicating that school commute times, which are mainly concentrated in these periods, have a certain impact on travel distance. LU_{recr}, POI_{cult}, and LU_{resi} also have some impact on both periods, but their influence is relatively weaker.

Table 4. Built environment variable contribution ranking.

Morning Peak			Evening Peak		
Variable	**Relative Importance (%)**	**Ranking**	**Variable**	**Relative Importance (%)**	**Ranking**
$STAT_{dist}$	43.210	1	$STAT_{dist}$	66.339	1
POI_{cond}	27.997	2	$SURR_{road}$	10.259	2
$SURR_{bus}$	6.215	3	$SURR_{bus}$	6.191	3
LU_{mixt}	5.349	4	$SURR_{hous}$	6.134	4
LU_{resi}	3.996	5	LU_{mixt}	2.779	5
$SURR_{road}$	3.843	6	LU_{recr}	1.516	6
LU_{seco}	2.009	7	POI_{cult}	1.159	7
POI_{scen}	1.463	8	LU_{resi}	1.057	8
POI_{cult}	1.099	9	LU_{seco}	0.983	9
LU_{recr}	1.011	10	POI_{comp}	0.901	10

5.4. Comparison of Model Results

To validate the advantages of the proposed model in fitting results, a comparison was made with the results of XGBOOST, GBDT, and ordinary least squares (OLS) regression methods, using three classic indicators: R-squared, mean absolute error (MAE), and RMSE. Table 5 presents the comparison results.

Table 5. Comparison of model results.

Peak Period	Model	R-Squared	MAE	RMSE
Morning peak	SSA-XGBOOST	0.633	1114.660	1420.421
	XGBOOST	0.611	1145.251	1468.254
	GBDT	0.577	1215.802	1515.450
	OLS	0.415	1315.454	1592.725
Evening peak	SSA-XGBOOST	0.583	1277.125	1572.279
	XGBOOST	0.544	1312.252	1624.201
	GBDT	0.511	1342.251	1671.007
	OLS	0.408	1411.052	1715.445

As shown in Table 5, the SSA-XGBOOST method proposed in this study performs the best across all the proposed indicators. The fitting effect for the morning peak period is better than that for the evening peak period. Additionally, it can be seen that the fitting effect of machine learning methods is superior to that of linear regression methods, indicating that the relationship between the built environment and travel distance is relatively complex and cannot be effectively fitted by linear methods.

5.5. SHAP Summary Chart

To specifically analyze the positive and negative feedback characteristics of various built environment variables on travel distance, SHAP attribution analysis needs to be used for an in-depth analysis. The feedback characteristics of the two peak periods' built environment variables on travel distance are given in Figure 3. The sum of absolute SHAP values corresponding to the constructed environmental variables decreases progressively from top to bottom in Figure 3.

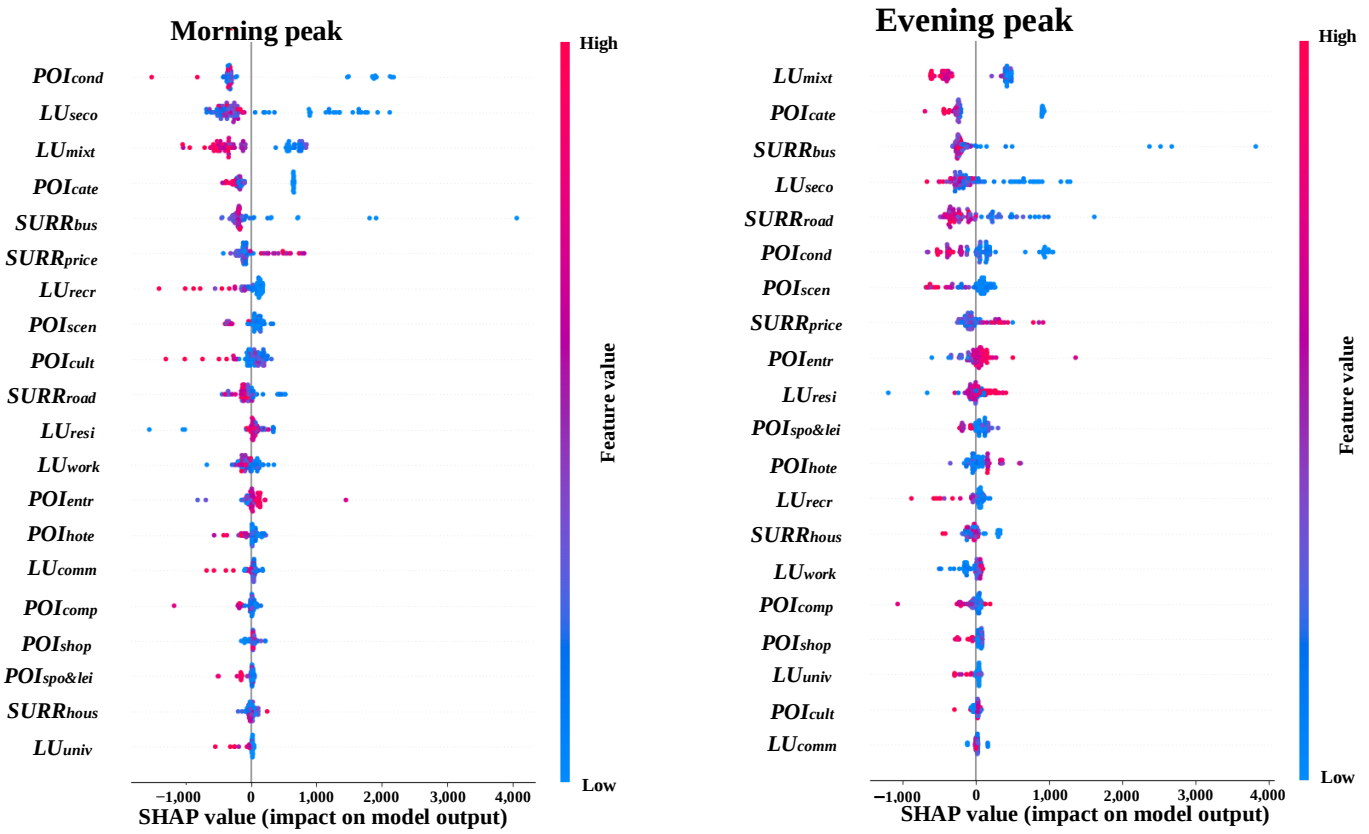


Figure 3. SHAP summary chart of travel distance.

In Figure 3, each dot represents an actual value of the built environment variable, with the color gradually changing from blue to red, representing the sample values from small to large. The horizontal axis represents the SHAP values corresponding to each variable. From the plot, the following information can be obtained: during the morning peak period, higher values of POI_{cond} show negative SHAP values, indicating that POI_{cond} have negative feedback on travel distance during the morning peak period. Similarly, higher values of LU_{seco} and LU_{mixt} also lead to a decrease in travel distance. Conversely, $SURR_{price}$ show positive feedback. For the evening peak period, as the LU_{mixt} increases, the corresponding SHAP values are negative, indicating a strong negative correlation. The positive and negative feedback of the remaining variables will not be elaborated upon here. To more clearly express the complex nonlinear relationship between built environment variables and travel variables, SHAP partial dependence plots should be used.

5.6. SHAP Partial Dependence

For the morning peak travel distance at stations in Figure 4, many variables exhibit a combination of positive and negative feedback mechanisms with trends and threshold limits. From the plot, the following information can be derived. As the number increases in POI_{cond}, the SHAP value remains negative, with only small values showing a slight positive value. This indicates that the greater the number of commercial residential areas

around most stations, the shorter the morning peak travel distance. Similarly, the number of $SURR_{bus}$ shows the same characteristic. The threshold is 50, meaning that after the number of bus stops exceeds 50, the travel distance gradually decreases as the number of $SURR_{bus}$ increases. LU_{mixt} has positive SHAP values, which turn negative as the value increases, indicating a threshold around 0.3. This reflects a positive-to-negative feedback feature. $SURR_{road}$, LU_{seco}, and POI_{scen} also exhibit a similar positive-to-negative feedback pattern. The SHAP value for LU_{resi} remains around zero but has some points with positive SHAP values, indicating a weak positive feedback characteristic. POI_{cond} shows a bimodal feedback pattern, with initial negative feedback turning positive and then negative again. This illustrates the complex impact of such built environment variables on morning peak travel distance.

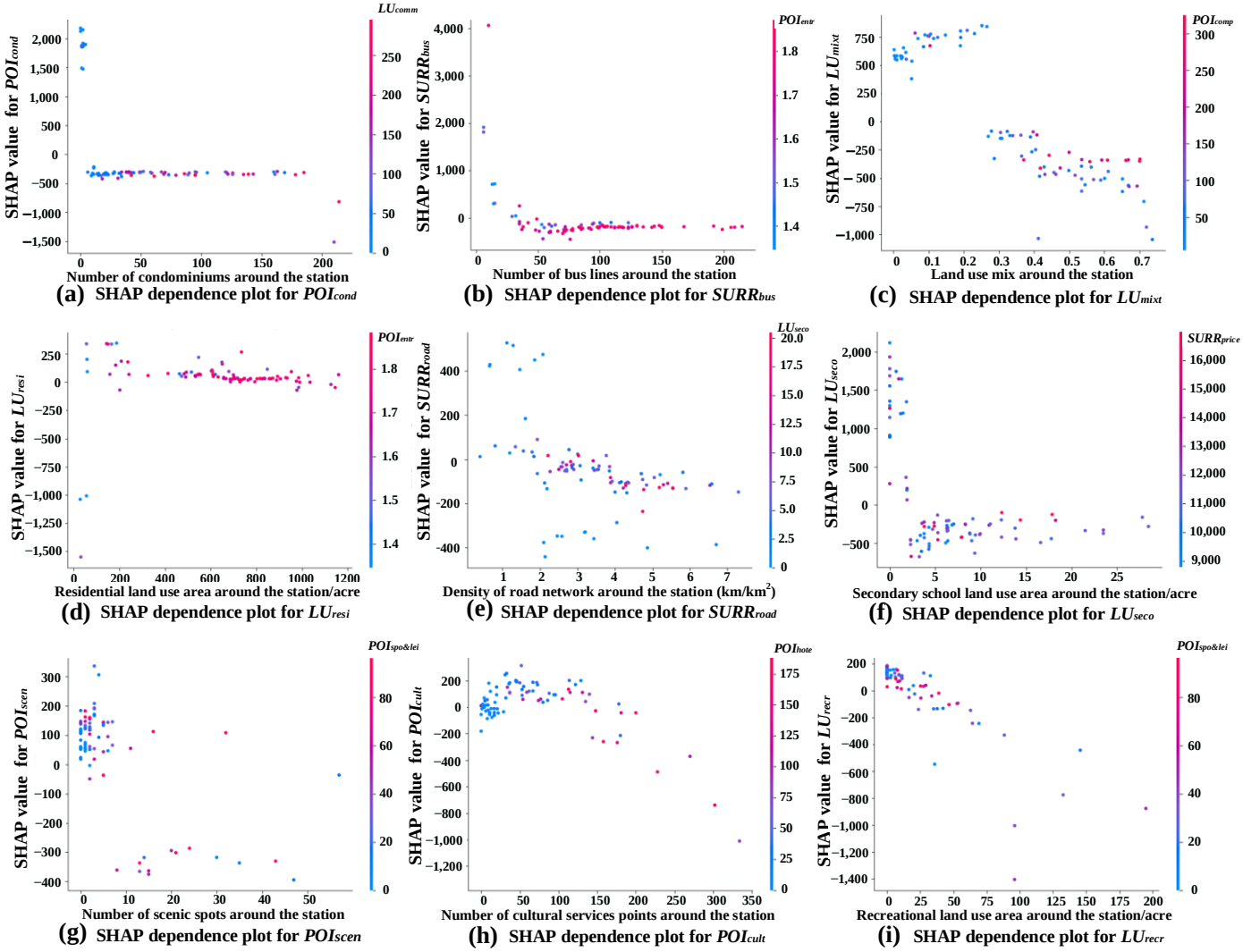


Figure 4. SHAP partial dependency plots for built environment in the morning peak period.

Figure 5 shows that $SURR_{road}$, $SURR_{bus}$, LU_{recr}, LU_{seco}, and POI_{comp} exhibit a typical positive-to-negative feedback mechanism, with their respective thresholds being 3 km/km^2, 30, 25 acres, 5 acres, and 80. $SURR_{hous}$ shows SHAP values fluctuating around zero as the number of $SURR_{hous}$ increases, indicating a weak correlation for this variable. LU_{mixt} exhibits a strong segmentation characteristic around a threshold of 0.35. Before this threshold, SHAP values are relatively high, but they drop sharply beyond this point, which is a very significant finding. POI_{cult} and LU_{resi} both exhibit a negative-to-positive feedback mechanism, with corresponding thresholds around 50 and 600, respectively.

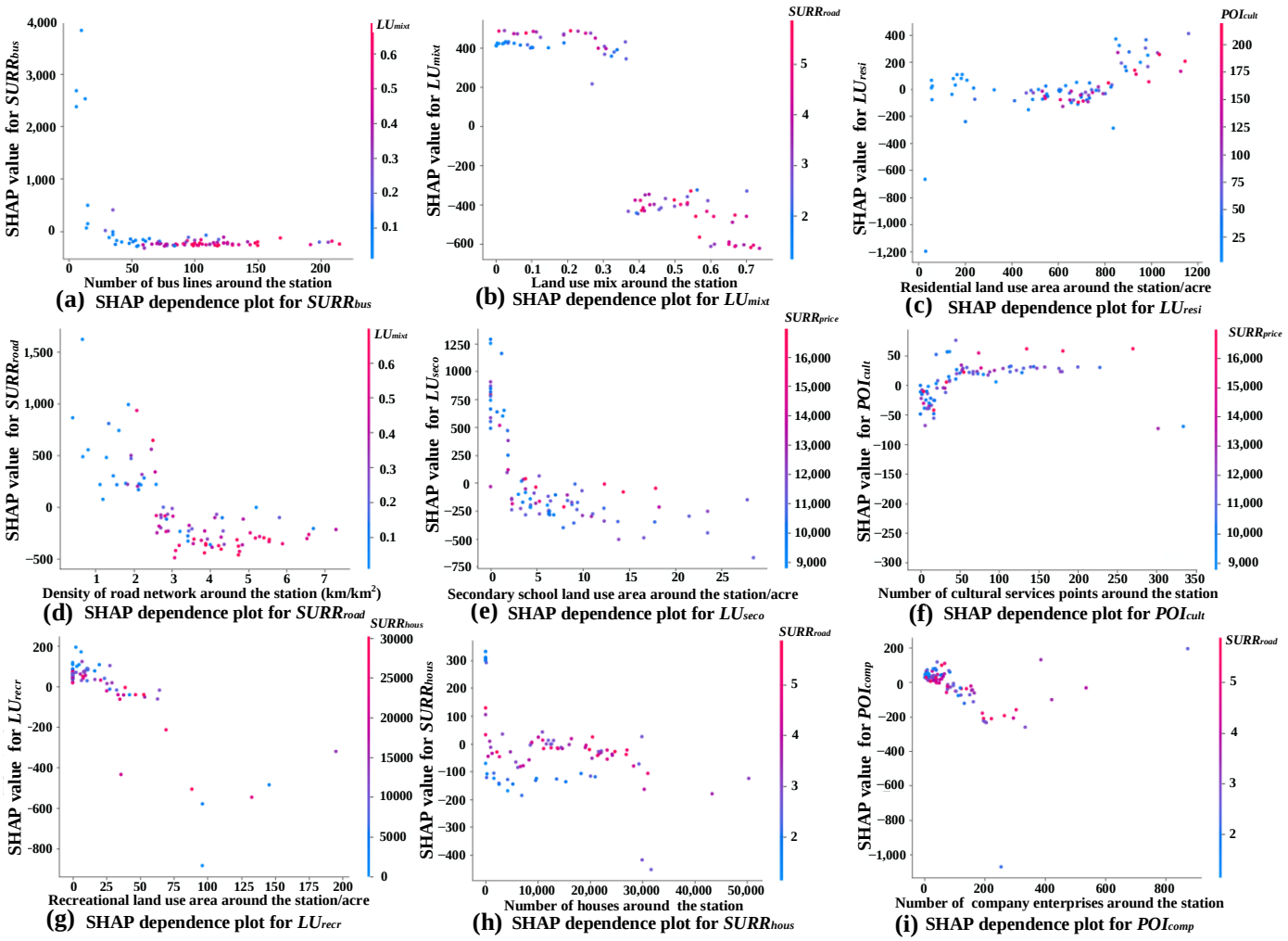


Figure 5. SHAP partial dependency plots for built environment in the evening peak period.

6. Discussion and Conclusions

6.1. Key Findings

To systematically analyze the spatiotemporal characteristics of passenger travel distances and their relationship with the built environment at the station level during peak operational periods under normalized conditions, this paper uses Xi'an as a case study. Multi-source built environment data are incorporated into the improved SSA-XGBOOST-SHAP analytical framework. We elucidate the nonlinear relationships and threshold effects between travel distance and the built environment, and the fitting results are well-explained. This is the first systematic study of the relationship between travel distance and the built environment across different time periods, providing better insights into the resilience of public transportation, especially subway travel.

Specifically, it reveals that during both morning and evening peak periods, the distance to the city center is the most important feature affecting travel distance. The relationship between station characteristics and the built environment includes the following: land use mix, road network density, and the number of bus stops all significantly contribute during different periods. The relationship between different built environments and travel distance is not simply a nonlinear correlation but shows characteristics of initially negative, then positive, or initially positive, then negative, effects at different times. Some built environment factors may not have a significant impact on travel distance or may show more complex bimodal characteristics.

More specifically, the subway network passenger flow characteristics based on subway travel distance can be summarized as follows: different built environment factors influence subway passengers' travel choices and routes. For example, an increase in the number of shopping service points and companies may attract more people to take the subway, thereby reducing travel distance. There is a linear relationship between the distance to the city center and travel distance. The closer to the city center, the shorter the travel distance. This is likely because the city center is a hub of commercial and social activities, making people more inclined to live nearby. This has important implications for the planning and layout of the subway network. The number of bus routes has a positive feedback relationship with travel distance. Considering the connection between buses and subways, operators can optimize bus route planning to improve travel efficiency and convenience for passengers. The number of entrances and exits and road network density have a negative correlation with travel distance. Enhancing the diversity of exit locations and increasing pathways connecting stations can improve the travel experience for subway users.

Our research findings are relatively consistent with and further complement the existing framework of knowledge on the influence of the built environment on travel behavior. Early studies have consistently emphasized the importance of proximity to the city center in determining travel distances. For example, previous research [8,14] has found that central urban areas attract shorter trips due to higher densities of commercial and social activities, supporting the strong correlation we observed between travel distance and proximity to the city center. Moreover, the significance of land use mix, road network density, and the number of bus stops as influential factors aligns with the emphasis on the role of diverse land uses and connectivity in promoting public transit usage [15,23,37]. However, our study provides a more nuanced understanding by revealing nonlinear and threshold effects, which may have been overlooked by previous models.

In terms of complex feedback mechanisms, we have identified instances where initially negative effects turn positive or initially positive effects turn negative during different time periods, adding depth to the existing literature. For example, while previous research [18] recognized the complex interplay between built environment characteristics and travel behavior, our use of the SSA-XGBOOST-SHAP framework allows for a more detailed exploration of these dynamics, highlighting specific times when each factor's influence changes.

Additionally, our findings on the impact of the number of shopping service points and companies on reducing travel distances corroborate the work, which demonstrated that mixed-use developments can shorten travel distances by concentrating daily needs within closer proximity [26,27]. However, our study extends this understanding by quantifying the direct effects on subway travel specifically, thereby offering targeted insights for urban planners and transit authorities.

In summary, while our study confirms several established principles within the field, it also introduces novel insights into the temporal and nonlinear aspects of the built environment's impact on travel behavior. These contributions not only enhance our understanding of the resilience of public transportation systems but also offer practical implications for improving urban transit infrastructure.

6.2. Policy Implications

Considering the rapid expansion of developing countries, especially in major urban areas like those in China, and the accelerated progress of urban rail construction, the analytical results of this study can offer policy guidance and recommendations for urban rail layout planning, route trends, and station area development.

Firstly, the analysis demonstrates the policy impacts brought about by land use characteristics. Policies related to residential areas exhibit significant correlations with changes in travel distances up to a certain threshold, with the correlation between the rate of change in travel distance and the positive or negative variation not being constant. When variables of this nature reach a certain value, sites with a substantial aggregation of residential land

uses and spatial distribution far from the city center exhibit significant positive correlations. This underscores the importance during the planning phase of controlling the number of residential units and balancing the job–residence ratio [18]. Likewise, establishing non-residential facilities near metro stations to increase the selectivity of travel purposes in the vicinity can help reduce travel distances. Thus, encouraging the transfer of some job opportunities and consumption facilities toward non-central areas is beneficial for the sustainable development of cities [42].

Secondly, the connectivity between travel distances and other infrastructure is also significant. In research focusing on passenger flow, it is suggested that stations with longer travel distances can be considered stations with higher demand as passenger flow increases. These stations exhibit a higher dependence on subway travel. This difference is believed to be related to the accessibility of travel corresponding to the transportation infrastructure around the station area [43]. This study also confirms a significant correlation between subway travel and bus route density, consistent with previous research in other cities in China [44]. Bus and subway services have cooperative and competitive influences on travel distances. Providing free transfers between subway and bus stations in Seoul, South Korea, greatly encourages public transportation travel [45]. Therefore, in the process of urban rail transit construction and subsequent TOD development, accelerating the connection with other transportation modes and densities is beneficial for increasing the demand for subway services and improving the efficiency of intermodal connections.

Furthermore, several studies have confirmed that the distance of a station from the city center is the most influential factor in travel distances. To better align with urban TOD development, more emphasis should be placed on constructing diverse living facilities at stations farther from the city center, particularly focusing on providing more job opportunities and shopping facilities to balance the uneven travel distances caused by characteristics such as employment and education. Another finding of this study indicates that educational facilities have a significant impact on travel distances during peak periods, showing a negative correlation. This underscores that the distribution of educational infrastructures is also a key factor in balancing travel distances. Accessibility to education affects community residents' travel patterns, representing a crucial measure to enhance residents' travel happiness [10].

6.3. Limitations and Future Work

Although this paper presents many analyses of the spatiotemporal characteristics and nonlinear factors of travel distance at the station level, not all conclusions are infallible. Firstly, we believe that such a research strategy should be expanded to other cities, especially those outside mainland China, to determine whether the model's applicability is influenced by urban differences. Secondly, regarding the choice of buffer zones, although many studies have chosen an 800 m range, other walking-scale buffer zones (e.g., 400 m, 600 m, and 1000 m) are also feasible. Moreover, the same-scale buffer zone may not be entirely suitable for different types of subway stations within a single city. Therefore, it is recommended that future studies conduct more comprehensive and detailed multi-city research within feasible ranges.

Author Contributions: Conceptualization, methodology, software, validation by P.L.; investigation, writing—original draft preparation, by W.L.; resources, writing—review and editing, supervision, by Q.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 71871013; 71871027).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors express their gratitude to Hao Wang and Yuqing Wang for their valuable feedback and technical support in enhancing this research work. Equally appreciated are the three anonymous reviewers for their valuable editing suggestions on the manuscript.

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

References

1. Liu, L.; Wang, Y.; Hickman, R. How Rail Transit Makes a Difference in People's Multimodal Travel Behaviours: An Analysis with the XGBoost Method. *Land* **2023**, *12*, 675. [CrossRef]
2. Kwon, J.-H.; Cho, G.-H. The Long-Lasting Impact of Past Mobility Dependence on Travel Mode Share in a New Neighborhood: The Case of the Seoul Metropolitan Area, South Korea. *Land* **2023**, *12*, 1922. [CrossRef]
3. Kim, M.; Cho, G.-H. Analysis on bike-share ridership for origin-destination pairs: Effects of public transit route characteristics and land-use patterns. *J. Transp. Geogr.* **2021**, *93*, 103047. [CrossRef]
4. Yin, J.; Cao, X.J.; Huang, X. Association between subway and life satisfaction: Evidence from Xi'an, China. *Transp. Res. Part D Transp. Environ.* **2021**, *96*, 102869. [CrossRef]
5. Gan, Z.; Yang, M.; Feng, T.; Timmermans, H.J. Examining the relationship between built environment and metro ridership at station-to-station level. *Transp. Res. Part D Transp. Environ.* **2020**, *82*, 102332. [CrossRef]
6. Chen, E.; Ye, Z.; Bi, H. Incorporating smart card data in spatio-temporal analysis of metro travel distances. *Sustainability* **2019**, *11*, 7069. [CrossRef]
7. Tao, T.; Cao, J. Exploring nonlinear and collective influences of regional and local built environment characteristics on travel distances by mode. *J. Transp. Geogr.* **2023**, *109*, 103599. [CrossRef]
8. Peng, H.; Wen-xi, L.; Yan, L.; Qi, X. Spatial Patterns of Nonlinear Effects of Built Environment on Beijing Subway Ridership. *J. Transp. Syst. Eng. Inf. Technol.* **2023**, *23*, 187. [CrossRef]
9. Wang, A.; Peng, J.; Ren, P.; Yang, H.; Dai, Q. Impact of the built environment of rail transit stations on the travel behavior of persons with disabilities: Taking 189 rail transit stations in Wuhan City as an example. *Prog Geogr* **2021**, *40*, 1127–1140. [CrossRef]
10. Lin, Y.; Fu, H.; Zhong, Q.; Zuo, Z.; Chen, S.; He, Z.; Zhang, H. The influencing mechanism of the communities' built environment on residents' subjective well-being: A case study of Beijing. *Land* **2024**, *13*, 793. [CrossRef]
11. Guo, H.; Bai, Y.; Hu, Q.; Zhuang, H.; Feng, X. Optimization on metro timetable considering train capacity and passenger demand from intercity railways. *Smart Resilient Transp.* **2021**, *3*, 66–77. [CrossRef]
12. Shang, P.; Li, R.; Yang, L. Optimization of urban single-line metro timetable for total passenger travel time under dynamic passenger demand. *Procedia Eng.* **2016**, *137*, 151–160. [CrossRef]
13. Ta, N.; Zhao, Y.; Chai, Y. Built environment, peak hours and route choice efficiency: An investigation of commuting efficiency using GPS data. *J. Transp. Geogr.* **2016**, *57*, 161–170. [CrossRef]
14. Khan, M.; Kockelman, K.M.; Xiong, X. Models for anticipating non-motorized travel choices, and the role of the built environment. *Transp. Policy* **2014**, *35*, 117–126. [CrossRef]
15. Guzman, L.A.; Peña, J.; Carrasco, J.A. Assessing the role of the built environment and sociodemographic characteristics on walking travel distances in Bogotá. *J. Transp. Geogr.* **2020**, *88*, 102844. [CrossRef]
16. Long, Y.; Thill, J.-C. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems* **2015**, *53*, 19–35. [CrossRef]
17. Wang, Z.; Hu, Y.; Zhu, P.; Qin, Y.; Jia, L. Ring aggregation pattern of metro passenger trips: A study using smart card data. *Phys. A Stat. Mech. Appl.* **2018**, *491*, 471–479. [CrossRef]
18. Gan, Z.; Feng, T.; Wu, Y.; Yang, M.; Timmermans, H. Station-based average travel distance and its relationship with urban form and land use: An analysis of smart card data in Nanjing City, China. *Transp. Policy* **2019**, *79*, 137–154. [CrossRef]
19. Xu, X.-Y.; Kong, Q.-X.; Ji, J.-M.; Liu, J.; Sun, Q. Analysis of spatio-temporal heterogeneity impact of built environment on rail transit passenger flow. *J. Transp. Syst. Eng. Inf. Technol.* **2023**, *23*, 194. [CrossRef]
20. Zhou, M.; Wang, D.; Guan, X. Co-evolution of the built environment and travel behaviour in Shenzhen, China. *Transp. Res. Part D: Transp. Environ.* **2022**, *107*, 103291. [CrossRef]
21. Chang, T.; Yang, D.; Yang, Y.; Huo, J.; Wang, G.; Xiong, C. Impact of urban rail transit on business districts based on time distance: Urumqi Light Rail. *J. Urban Plan. Dev.* **2018**, *144*, 04018024. [CrossRef]
22. Chen, F.; Wu, J.; Chen, X.; Wang, J. Vehicle kilometers traveled reduction impacts of transit-oriented development: Evidence from Shanghai City. *Transp. Res. Part D Transp. Environ.* **2017**, *55*, 227–245. [CrossRef]
23. Choi, K. The influence of the built environment on household vehicle travel by the urban typology in Calgary, Canada. *Cities* **2018**, *75*, 101–110. [CrossRef]
24. Kim, J.; Brownstone, D. The impact of residential density on vehicle usage and fuel consumption: Evidence from national samples. *Energy Econ.* **2013**, *40*, 196–206. [CrossRef]
25. Zhang, L.; Hong, J.; Nasri, A.; Shen, Q. How built environment affects travel behavior: A comparative analysis of the connections between land use and vehicle miles traveled in US cities. *J. Transp. Land Use* **2012**, *5*, 40–52. [CrossRef]
26. Choi, J.; Lee, Y.J.; Kim, T.; Sohn, K. An analysis of Metro ridership at the station-to-station level in Seoul. *Transportation* **2012**, *39*, 705–722. [CrossRef]
27. Iseki, H.; Liu, C.; Knaap, G. The determinants of travel demand between rail stations: A direct transit demand model using multilevel analysis for the Washington DC Metrorail system. *Transp. Res. Part A Policy Pract.* **2018**, *116*, 635–649. [CrossRef]
28. Eliasson, J. Will we travel less after the pandemic? *Transp. Res. Interdiscip. Perspect.* **2022**, *13*, 100509. [CrossRef]

29. Wang, X.; Pei, T.; Li, K.; Cen, Y.; Shi, M.; Zhuo, X.; Mao, T. Analysis of changes in population's cross-city travel patterns in the pre-and post-pandemic era: A case study of China. *Cities* **2022**, *122*, 103472. [CrossRef]
30. Liu, J.; Cao, Q.; Pei, M. Impact of COVID-19 on adolescent travel behavior. *J. Transp. Health* **2022**, *24*, 101326. [CrossRef]
31. Chen, K.; Steiner, R. Longitudinal and spatial analysis of Americans' travel distances following COVID-19. *Transp. Res. Part D Transp. Environ.* **2022**, *110*, 103414. [CrossRef]
32. An, D.; Tong, X.; Liu, K.; Chan, E.H. Understanding the impact of built environment on metro ridership using open source in Shanghai. *Cities* **2019**, *93*, 177–187. [CrossRef]
33. Caset, F.; Blainey, S.; Derudder, B.; Boussauw, K.; Witlox, F. Integrating node-place and trip end models to explore drivers of rail ridership in Flanders, Belgium. *J. Transp. Geogr.* **2020**, *87*, 102796. [CrossRef]
34. Chen, E.; Ye, Z.; Wang, C.; Zhang, W. Discovering the spatio-temporal impacts of built environment on metro ridership using smart card data. *Cities* **2019**, *95*, 102359. [CrossRef]
35. Liu, M.; Liu, Y.; Ye, Y. Nonlinear effects of built environment features on metro ridership: An integrated exploration with machine learning considering spatial heterogeneity. *Sustain. Cities Soc.* **2023**, *95*, 104613. [CrossRef]
36. Chen, E.; Ye, Z.; Wu, H. Nonlinear effects of built environment on intermodal transit trips considering spatial heterogeneity. *Transp. Res. Part D Transp. Environ.* **2021**, *90*, 102677. [CrossRef]
37. Su, S.; Wang, Z.; Li, B.; Kang, M. Deciphering the influence of TOD on metro ridership: An integrated approach of extended node-place model and interpretable machine learning with planning implications. *J. Transp. Geogr.* **2022**, *104*, 103455. [CrossRef]
38. Yan, X.; Liu, X.; Zhao, X. Using machine learning for direct demand modeling of ridesourcing services in Chicago. *J. Transp. Geogr.* **2020**, *83*, 102661. [CrossRef]
39. Teixeira, J.F.; Lopes, M. The link between bike sharing and subway use during the COVID-19 pandemic: The case-study of New York's Citi Bike. *Transp. Res. Interdiscip. Perspect.* **2020**, *6*, 100166. [CrossRef]
40. Yu, L.; Cong, Y.; Chen, K. Determination of the peak hour ridership of metro stations in Xi'an, China using geographically-weighted regression. *Sustainability* **2020**, *12*, 2255. [CrossRef]
41. De Gruyter, C.; Butt, A.; Davies, L. Exploring the potential for unbundling off-street car parking in residential apartment buildings. *Transport Policy* 2024. [CrossRef]
42. Yang, L.; Yu, B.; Liang, Y.; Lu, Y.; Li, W. Time-varying and non-linear associations between metro ridership and the built environment. *Tunn. Undergr. Space Technol.* **2023**, *132*, 104931. [CrossRef]
43. Yang, H.; Zhao, Z.; Jiang, C.; Wen, Y.; Muneeb Abid, M. Spatially Varying Relation between Built Environment and Station-Level Subway Passenger-Distance. *J. Adv. Transp.* **2022**, *2022*, 7542560. [CrossRef]
44. Zhao, J.; Deng, W.; Song, Y.; Zhu, Y. What influences Metro station ridership in China? Insights from Nanjing. *Cities* **2013**, *35*, 114–124. [CrossRef]
45. Sohn, K.; Shim, H. Factors generating boardings at Metro stations in the Seoul metropolitan area. *Cities* **2010**, *27*, 358–368. [CrossRef]

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