frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 04 November 2024
DOI 10.3389/fpubh.2024.1469578

OPEN ACCESS

EDITED BY
Chen Zhong,
University College London, United Kingdom

REVIEWED BY
Jingxian Wu,
University of Shanghai for Science and Technology, China
Wei Wei,
Chongqing Jiaotong University, China

*CORRESPONDENCE
Yi Zhang
✉ darrenzhy@sjtu.edu.cn

RECEIVED 24 July 2024
ACCEPTED 17 October 2024
PUBLISHED 04 November 2024

CITATION
Peng B, Wang T, Zhang Y and Li C (2024) How to improve public environmental health by facilitating metro usage on weekend: exploring the non-linear and threshold impacts of the built environment.
Front. Public Health 12:1469578.
doi: 10.3389/fpubh.2024.1469578

How to improve public environmental health by facilitating metro usage on weekend: exploring the non-linear and threshold impacts of the built environment

Bozhezi Peng, Tao Wang, Yi Zhang* and Chaoyang Li

State Key Laboratory of Ocean Engineering, School of Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, China

Introduction: The accelerated motorization has brought a series of environmental concerns and damaged public environmental health by causing severe air and noise pollution. The advocate of urban rail transit system such as metro is effective to reduce the private car dependence and alleviate associated environmental outcomes. Meanwhile, the increased metro usage can also benefit public and individual health by facilitating physical activities such as walking or cycling to the metro station. Therefore, promoting metro usage by discovering the nonlinear associations between the built environment and metro ridership is critical for the government to benefit public health, while most studies ignored the non-linear and threshold effects of built environment on weekend metro usage.

Method: Using multi-source datasets in Shanghai, this study applies Gradient Boosting Decision Trees (GBDT), a nonlinear machine learning approach to estimate the non-linear and threshold effects of the built environment on weekend metro ridership.

Results: Results show that land use mixture, distance to CBD, number of bus line, employment density and rooftop density are top five most important variables by both relative importance analysis and Shapley additive explanations (SHAP) values. Employment density and distance to city center are top five important variables by feature importance. According to the Partial Dependence Plots (PDPs), every built environment variable shows non-linear impacts on weekend metro ridership, while most of them have certain effective ranges to facilitate the metro usage. Maximum weekend ridership occurs when land use mixture entropy index is less than 0.7, number of bus lines reaches 35, rooftop density reaches 0.25, and number of bus stops reaches 10.

Implication: Research findings can not only help government the non-linear and threshold effects of the built environment in planning practice, but also benefit public health by providing practical guidance for policymakers to increase weekend metro usage with station-level built environment optimization.

KEYWORDS

built environment, metro ridership, machine learning, nonlinearity, public environmental health

1 Introduction

The accelerated urbanization and motorization have brought severe environmental challenges, including air pollution, traffic congestion and climate change (1). The air and noise pollution brought by car dependence are the critical threats to public health, which can damage the physical and mental health of people (2). Under this circumstance, transit-oriented development (TOD) has been advocated among many countries to promote urban rail transit system (3). The large-scale construction of metro system has shifted from developed countries to developing contexts (4). In China, the metro system has been constructed in 59 cities and the total length has reached 11232.65 km by the end of 2023 (5). Since the urban traffic carbon emission is a critical reason of climate change, it is imperative for urban planners to improve public environmental health by facilitating metro usage.

Due to the large capacity, low cost and travel time reliability, metro system has become an effective transport alternative for not only the commuting trips on weekdays but also the leisure trips on weekends (1). The trip purpose and travel behavior of metro users can be significantly different between weekdays and weekends (6). For example, there are more commuting trips on weekdays with obvious rush hours, while more entertaining trips with no obvious peak hours on weekends (7). Since the travel modes for commuting people are relatively fixed, promoting metro usage on weekend can not only mitigate traffic congestion and reduce carbon emissions, but also benefit public health from multiple perspectives.

Compared to other factors which may influence the metro ridership (e.g., weather, fare, etc.), the built environment is more suitable to optimize at different metro stations (8). With the development of geographic information systems (GIS) and availability of big data, direct ridership models (DRMs) become more popular in recent metro ridership literature (9). Many of DRMs derived from ordinary least squares (OLS) regression (10), multilevel regression (11), or geographically weighted regression (12), by assuming a linear or loglinear relationship. However, the nonlinearity between the built environment and metro usage has been recently investigated by different DRMs based on several machine learning algorithms, such as Gradient Boosting Decision Tree (GBDT), Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) (4, 13, 14). Moreover, the non-linear influences of built environment have been discovered on different travel behavior, including shared bikes (15), shared e-scooters (16), ride-splitting (17), driving distance (18) and ride-sourcing (19). Among these emerging non-linear studies on metro ridership, most of them focused on non-linear effects of the built environment on weekday metro ridership (4, 13, 14), ignoring the temporal heterogeneity on weekend metro usage. The non-linear associations between built environment and metro ridership can be quite diverse between weekdays and weekends, while previous studies failed to address this issue.

To fill the gap, this study aims to promote the metro usage and improve the public environmental health by discovering the non-linear impacts of built environment on weekend metro usage. By utilizing various datasets and GBDT approach, this study attempts to address two research questions: (1) What is the relative importance of each built environment variable in affecting weekend metro ridership? (2) Does the built environment show non-linear impacts on weekend metro usage? What are the threshold and effective ranges?

The remaining part of this paper is structured as follows. Next section reviews the studies on associations between the built environment and metro ridership. Section three introduces the data, variables and methodology. Section four concludes the results. Section five discusses the research findings. The last section summarizes the paper and points out the limitation.

2 Literature review

Due to the popularity of urban rail transit system and transit-oriented development, studies on impacts of built environment on metro usage have brought increasing attentions in the past few decades (4, 13, 14, 20). In the past few years, DRMs have become popular than traditional ridership prediction model because of the convenience of data collection (13).

Although scholars have evaluated the built environment from different aspects, most measured the built environment by "5Ds," including density, diversity, design, destination accessibility and distance to transit (21). Higher activity density can increase the possibility of using metro system. For example, population density or employment density have significant and positive impacts on metro usage (22), both in developed countries (23) and developing countries (24–26). However, recent studies employed GBDT model and proposed that non-linear impacts of density on metro usage may appear negligible if it beyond certain threshold (4, 13, 14, 20).

Diversity, such as mixed land use, can improve metro usage by making the metro station surroundings more appealing. For example, land use mixture has been explored to have positive impacts on metro ridership in Spain, South Korea and China (12, 27, 28), but studies in other countries show insignificant effects of land use mix (23, 29–31). Recently, the non-linear impacts of diversity on metro ridership has been found to be non-trivial only when they are within certain ranges (13, 14).

Design measures road network within the station area. Design features, including street or intersection density can show positive impacts (14, 23, 24, 32–34) or negative effects on metro ridership (26, 28, 35). Recently, studies on DRMs have pointed out that design features have positive impacts on metro usage only if they are in certain ranges (13, 14).

Destination accessibility measures the accessibility to certain areas (city center) or facilities (shopping center). For example, some studies examined the non-linear associations between distance to city center and metro ridership (4, 13). However, the impacts of distance to city center are found to be insignificant in other studies (24, 34). Distance to transit, including bus stop and bus route, have also been explored by studies in different contexts (14, 20, 36, 37).

To sum up, some studies have used DRMs to analyze the impacts of built environment on metro usage, but almost neglect the metro usage on weekends. For these gaps, this study tries to improve the public environmental health by discovering the non-linear impacts of built environment on weekend metro usage.

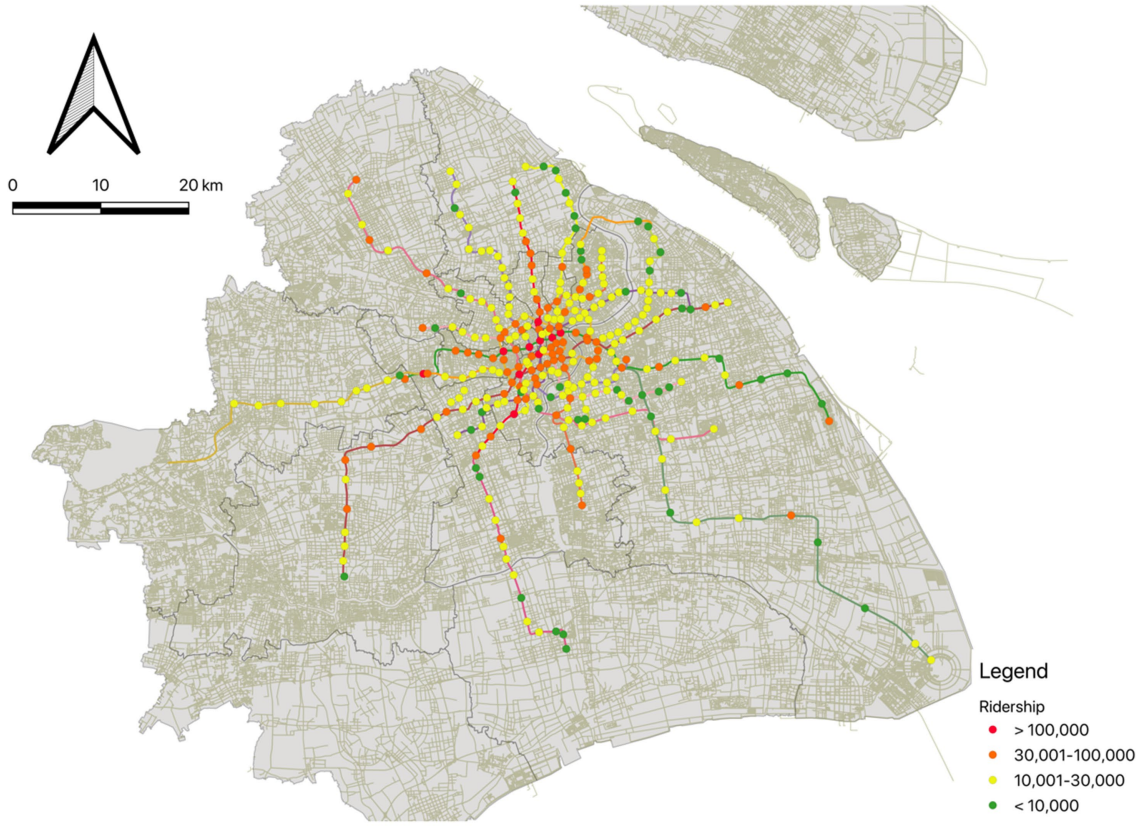

FIGURE 1
Weekend metro ridership of 328 metro stations in Shanghai.

3 Materials and methods

3.1 Study area

In this study, we utilized 1 month smartcard data on May 2023, including 17 lines and 328 stations in Shanghai (Figure 1). The smartcard data was provided by the Shanghai Government Data Portal[1] with hourly passengers. The raw smartcard data included number of hourly inbound and outbound passengers for each metro station. Daily ridership on weekends of each station was then aggregated by adding up hourly inbound and outbound passengers. The average station ridership on weekends is 27,259 riders per day. People's Square Station, the interchange station of three lines which located at the city center, has the highest ridership of 243,938 passengers. Thirty-two stations have more than 50,000 riders per day, while 55 stations have fewer than 10,000 daily ridership.

3.2 Variables

Twelve built environment variables were included to measure 5Ds built environment in this study. Multiple sources and platforms are used to collected the data of the built environment characteristics, including OpenStreetMap and AMAP API.

3.2.1 Density

Population density around metro station is calculated within 500 m buffer, based on the WorldPop population data with 100 m resolution. Using POI data from AMAP API, employment density is determined by the percentage of employment-related point of interests within 500 m buffer. Rooftop density, the measure of land development, was also calculated within 500 m buffer based on the rooftop area datasets (38).

3.2.2 Diversity

Land use mix, the entropy index for different land use, was utilized to measure station level land use diversity. Since the land use data was not open access to the public, 23 categories of point of interests are used as the alternative to calculate the land use mix entropy within 500 m buffer, including catering, shopping, education, employment, entertainment, tourism, public service, sports, green space, etc.

3.2.3 Design

Intersection and road density were used to measure street design, based on the data from OpenStreetMap. Road density was measured by removing highways and sidewalks from the OpenStreetMap street

1 https://data.sh.gov.cn/

network, while number of intersections is measured by counting 3-way or more intersections.

3.2.4 Destination accessibility

Network distance to CBD and straight distance to the nearest Sub-CBD were involved to measure the effects of accessibility. The CBD and several city Sub-CBDs were chosen according to the official document by Shanghai government (39). Network distance to the nearest highway entrance is was also used to assess the destination accessibility.

3.2.5 Distance to transit

Bus stop and bus line are counted within the station service area, while straight distance to the nearest bus stop is selected to measure the distance to transit.

All the built environment characteristics are measured within 500 m buffer by QGIS (Figure 2). Table 1 summarizes the statistics of all built environment variables.

3.3 Methodology

We employ Gradient Boosting Decision Tree (GBDT) approach to analyze the non-linear effects of the built environment on weekend metro usage. GBDT has several merits for this study. GBDT do not pre-assume linear association between different variables (40). It can also visualize the non-linear relationship by depicting partial dependent plot, which shows the marginal effect on the predictions (41). Meanwhile, GBDT helps to evaluate the contribution of each feature by automatically calculating feature importance (14). Moreover, GBDT is not sensitive to multicollinearity problems, which makes it possible to examine non-linear impacts of different features on weekend metro usage, even they are highly correlated. The effectiveness of GBDT has been recently proved by several studies to evaluate the non-linear effects of built environment on different kinds of travel behavior.

Mathematically, GBDT sets the approximation function $F_T(x)$ by combined several decision trees, and aims to minimize the loss function $L[y,F(x)]=[y-F(x)]^2$. The approximation function $F_T(x)$ is given by Equation 1:

$$F_T(x)=\sum_{t=1}^{T} f_t(x)=\sum_{t=1}^{T}\theta_t h(x;\eta_t) \quad (1)$$

where T is number of trees, η_t is the parameter of the t^{th} tree $h(x;\eta_t)$, θ_t is the weight of $h(x;\eta_t)$ which can be calculated by minimizing the loss function. The optimization process includes several iterative steps. First, the initialization function is determined as Equation 2:

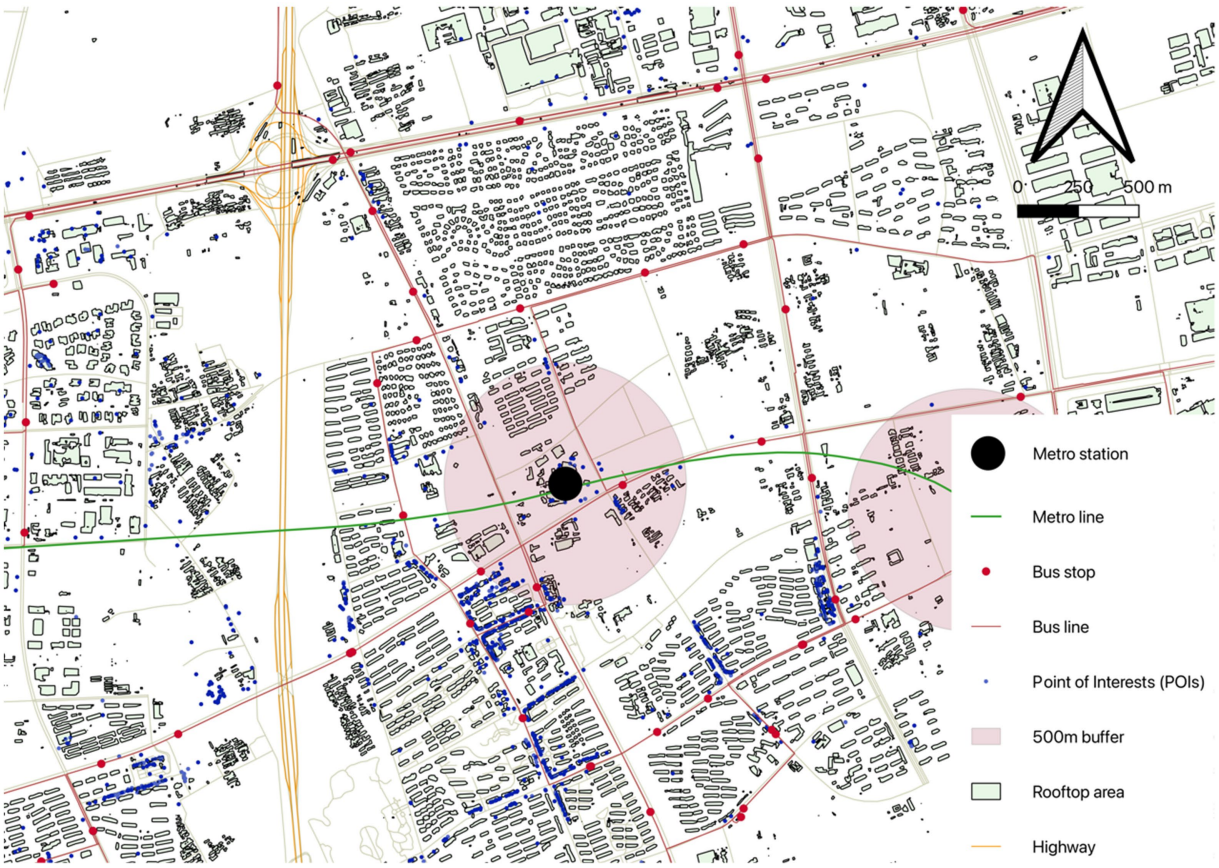

FIGURE 2
An example of built environment variables within station area.

$$f_0(x) = \text{argmin}_\theta \sum_{i=1}^{N} L(y_i, \theta) \tag{2}$$

Second, the residual error $r_{t,i}$ is derived for each sample i in t^{th} iteration as Equation 3:

$$r_{t,i} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)} \tag{3}$$

Third, $(x_i, r_{t,i})$ are utilized to fit the $t^{th}(t = 1,2,\ldots,T)$ tree $h(x;\eta_t)$ by getting the $R_{t,j}, (j = 1,2,\ldots,J_t)$, while J_t is the tree size. After that, we can use tree traversal to determine the optimal gradient as Equation 4:

$$\theta_t = \text{argmin}_\theta \sum_{i=1}^{N} L(y_i, F_{t-1}(x_i) + \theta h(x;\eta_t)) \tag{4}$$

Thus, we can rewrite the iterative equation as Equation 5:

$$f_t(x) = f_{t-1}(x) + \theta_t h(x;\eta_t) \tag{5}$$

To moderate overfitting, learning rate is proposed as the shrinkage parameter $\varepsilon\,(0 < \varepsilon \le 1)$ (41). Therefore, the final function could be written as Equation 6:

$$F(x) = f_t(x) = f_{t-1}(x) + \varepsilon\,\theta_t h(x;\eta_t) \tag{6}$$

For each feature, the feature importance can be calculated by the final model. The importance of feature x_i can be determined as Equation 7 (42):

$$I_{x_i} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} d_j} \tag{7}$$

where j denotes tree nodes, and d_j refers the differences of loss function when make j^{th} tree splitting.

Mathematically, the partial dependence of an independent variable x_s can be calculated as Equation 8 (43):

$$F_s(x_s) = E_{x_c}\left[F(x_s, x_c)\right] \tag{8}$$

where x_c represents other variables. Then, the partial function $F_s(x_s)$ can be determined by averaging over all samples as Equation 9:

$$\overline{F_s}(x_s) = \frac{1}{N}\sum_{i=1}^{N} F(x_s, x_c) \tag{9}$$

Shapley additive explanations (SHAP) can also interpret the model outputs by machine learning models (44). Shapley value (45) are used in SHAP to evaluate the effects of each variable as Equation 10 (44):

$$\varphi_v = \sum_{z' \subseteq x'} \frac{|z'|!(V - |z'| - 1)!}{V!}\left[f_x(z') - f_x(z'/V)\right] \tag{10}$$

where V denotes number of variables, φ_v represents the contribution of variable v, $f(x)$ refers model outputs, $|z'|$ counts non-zero entries in z'.

However, GBDT does have certain restrictions. For example, it cannot perform significance tests and produce coefficient of variables, while feature importance can be used as the substitution. It is also easy to overfit, while using cross-validation and suitable shrinkage parameter can solve this problem (41). In this study, we conduct 5-fold cross-validation and selected the learning rate as 0.001. We get the optimal GBDT model with the lowest RMSE after 2,893 iterations, and the pseudo-R^2 is 0.83.

4 Results

4.1 Feature importance of the built environment

Table 2 presents the relative feature importance and ranking in determining metro usage on weekends. Land use mixture has the largest predictive power, with the relative importance of 16.26%. As the measurement of diversity, it has been observed as a critical factor on metro ridership prediction by many previous studies in different contexts (4, 13, 28, 33). Distance to CBD, a measure of regional accessibility, has the second large relative importance, with a contribution of 13.61%. Other destination accessibility variables, including distance to highway (7.21%) and distance to Sub-CBD (5.21%), also have non-trivial impacts on weekend metro usage. The importance of bus line is also substantial, accounting for 12.17% and ranking 3rd over all independent variables. This corresponds to the existing findings that distance to transit can notably affect the metro usage (4, 13). Among five categories of built environment, density features have the largest relative importance of 29.29% on ridership prediction, collectively contributed by three density variables. By contrast, design variables (e.g., intersection and street density) only shown trivial impacts on weekend metro usage, with relative importance of only 2.78 and 2.61%, respectively.

4.2 SHAP beeswarm plot of the built environment

To discover the contribution of each variable and analyze how variables of stations influence the metro usage, SHAP beeswarm plots (also called the SHAP summary plots) are employed in this study.

The SHAP beeswarm plot sorts variables by mean absolute value of SHAP values, while uses SHAP value to show the effect distribution of variables (Figure 3). Each station is displayed by one point for each variable, while the horizontal axis presents SHAP values. Value of each feature is shown in different colors.

As shown in Figure 3, land use mixture ranks first by SHAP values, which is similar to the feature importance results. Meanwhile, Number of bus line, distance to CBD, employment and rooftop area are other top five significant variables, which is same

10.3389/fpubh.2024.1469578

TABLE 1 Statistics of all variables.

Variables	Description	Mean	S.D.	Min	Max	Data source
Dependent variable						
Weekend ridership	Daily metro ridership on weekends (count)	27,259	27,839	1,011	243,938	Metro smartcard data of Shanghai on May 2023
Built environment variables						
Density						
Population density	Population density within 500 m buffer (1,000 people/km^2)	16.83	10.52	0.04	41.58	WorldPop population data 2023
Employment density	Ratio of employment POI within 500 m buffer	0.14	0.15	0.03	0.95	Point-of-interest (POI) data 2023
Rooftop density	Rooftop area ratio within 500 m buffer	0.18	0.06	0.02	0.45	Vectorized rooftop area data 2020
Diversity						
Land use mixture	The entropy index $\frac{-\sum_{i=1}^{m}(p_i)\ln(p_i)}{\ln(m)}$ where m denotes different POI and p_i represents the ratio.	0.72	0.13	0.14	0.87	Point-of-interest (POI) data 2023
Design						
Road density	Road centerline length per km^2 (km/km^2)	5.83	1.94	1.22	13.84	OpenStreetMap data 2023
Intersection	Number of intersections within 500 m buffer (count)	9.24	6.12	0.00	42.00	OpenStreetMap data 2023
Destination accessibility						
CBD	Network distance to CBD (km)	14.14	10.29	0.16	65.90	OpenStreetMap data 2023
Sub-CBD	Straight distance to the nearest Sub-CBD (km)	6.75	5.22	0.00	37.85	OpenStreetMap data 2023
Highway	Network distance to the nearest highway (km)	1.25	1.46	0.04	6.62	OpenStreetMap data 2023
Distance to transit						
Bus stop	Number of bus stops within 500 m buffer (count)	6.42	3.50	1.00	26.00	Point-of-interest (POI) data 2023
Bus line	Number of bus routes within 500 m buffer (count)	17.15	10.27	0.00	62.00	Point-of-interest (POI) data 2023
Nearest bus stop	Straight distance to the nearest bus stop (km)	0.12	0.07	0.01	0.42	Point-of-interest (POI) data 2023

TABLE 2 Relative importance and ranking of variables.

Category	Features	Ranking	Relative importance
Density (29.29%)	Population density	6	9.18%
	Employment density	4	10.06%
	Rooftop density	5	10.05%
Diversity (16.26%)	Land use mixture	1	16.26%
Design (5.39%)	Road density	12	2.61%
	Intersection	11	2.78%
Destination accessibility (26.03%)	CBD	2	13.61%
	Sub-CBD	9	5.21%
	Highway	7	7.21%
Distance to transit (23.03%)	Bus stop	8	7.06%
	Bus line	3	12.16%
	Nearest bus stop	10	3.81%

with the feature importance results but with little difference with ranking. Moreover, number of bus stops, which ranked only 8th by relative importance, are the 6th most significant variable by SHAP values.

Bus line, rooftop density, bus stop and population density are positively related with SHAP value, while land use mixture, CBD and employment density show negative associations. It means that large number of bus lines and bus stops, high rooftop density and population density (in red color) can increase more metro ridership on weekends, while high land use diversity, long distance to CBD and high employment density (in red color) lower the weekend metro ridership.

4.3 Non-linear impacts of built environment on weekend metro usage

To explore the relationship between the built environment and weekend metro ridership, partial dependence plots (PDPs) are employed in this study. Overall, all independent variables shown non-linear associations with weekend metro usage. Figure 4 presents the non-linear impacts of built environment variables on weekend metro usage.

As shown in Figure 4, the weekend metro ridership remains (at about 40,000) when land use mixture entropy is smaller than 0.65. However, the weekend metro ridership drops substantially to less than 25,000 when the entropy moves from 0.65 to 0.75, and no further decrease occurs.

Bus line is positively associated with weekend metro usage. The metro usage keeps stable at less than 25,000 when bus route is

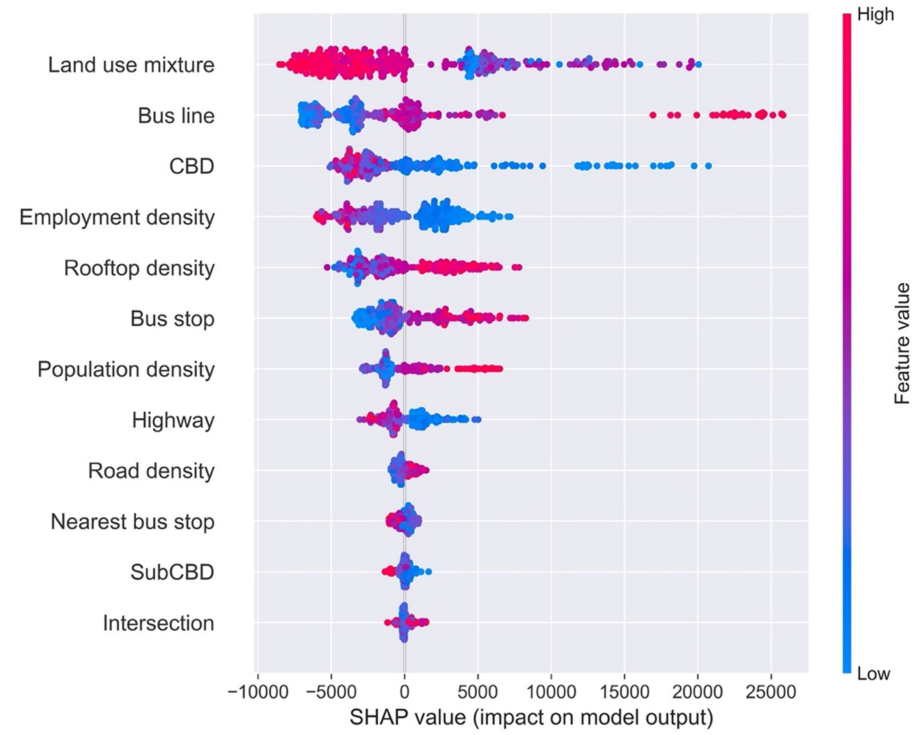

FIGURE 3
SHAP beeswarm plot of built environment.

TABLE 3 Effective ranges of variables.

Variables	Effective range/ Threshold	Association
Land use mixture	0.65–0.75 (scale)	Negative
Bus line	15–20, 30–35 (count)	Positive
CBD	2–10 (km)	Negative
Employment density	0–0.2 (scale)	Negative
Rooftop density	0.2–0.25 (scale)	Positive
Bus stop	3–10 (count)	Positive
Road density	5–7 (km/km^2)	Positive
Highway	0.5–2 (km)	Negative

less than 15. After that, the ridership suddenly increases to 30,000 as bus route moves from 15 to 20, and no increase in metro ridership has been found when bus line is between 20 and 30. However, the weekend metro ridership sharply increases from 30,000 to 50,000 when number of bus line reaches 35, and then remain constant.

The association between distance to CBD and weekend metro ridership is negative. The weekend ridership drops dramatically from 45,000 to 25,000 when the distance to CBD grows from 0 to 10 km. However, no further decrease of metro ridership has been found when the distance to CBD exceeds 10 km. Similar pattern has been found for distance to Sub-CBD.

Rooftop density has positive effects on weekend metro ridership. When the rooftop density is less than 0.2, metro ridership keeps 25,000. After that, the weekend ridership rises substantially from 25,000 to 31,000 when rooftop density between 0.2 and 0.25. As shown in the PDPs, as bus stop increases from 0 to 10, weekend metro ridership rises by 6,000. However, this effect looks negligible when there are more than 10 bus stops.

Meanwhile, the distance between metro station and nearest transit station has negative impacts to weekend metro usage, with an effective interval of 100–200 m. However, this effect is limited and the difference in metro ridership is only about 1,500, echoing the small relative importance of this variable in metro ridership prediction.

Overall, PDPs show the average effect of the built environment variables without specific instances. To visualize the partial dependence of one variable on weekend metro ridership for each station, we also combined Individual Conditional Expectation (ICE) curves with PDPs as shown in Figure 5. In Figure 5, the ICE curves are presented in light blue lines, while the PDP is shown in dark blue line as the average.

As shown in Figure 5, for each independent variable, all the ICE curves seem to follow the similar pattern with the partial dependence plot. It means that there is no obvious heterogeneous relationship created by interactions. Under this circumstance, employing PDPs in this study can provide good summary of the impacts of built environment on predicted metro usage on weekends.

5 Discussion

Promoting metro usage on weekends by optimizing station-level built environment is a critical way to address a series of

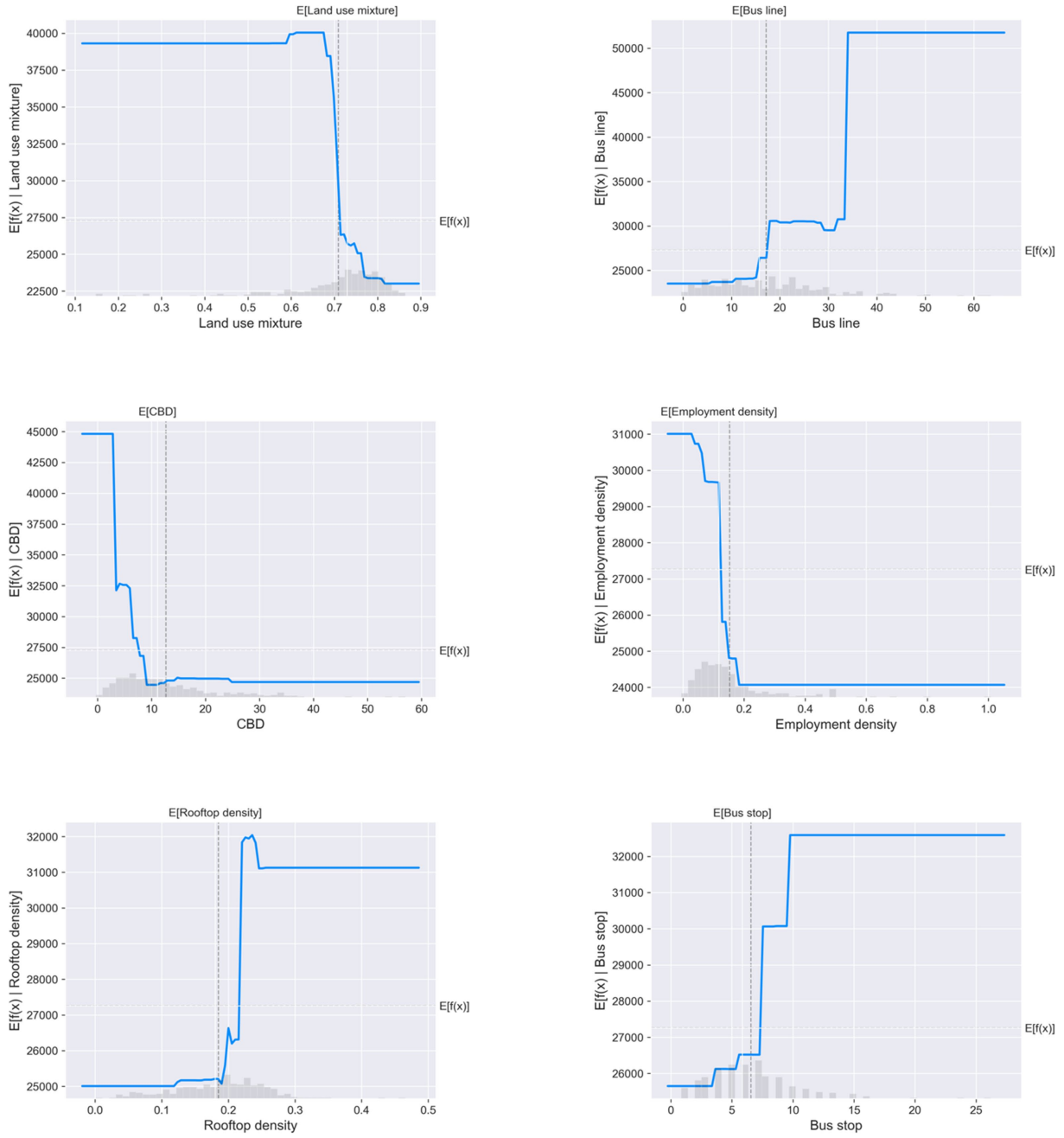

FIGURE 4 (Continued)

environmental challenges from accelerated urbanization and mobilization. This study employed GBDT approach to evaluate the non-linear associations between the built environment and weekend metro ridership in Shanghai. Several model interpretation methods are utilized to unravel the non-linear impacts of factors on weekend metro usage.

Based on the results of relative importance and SHAP values, we recognized that land use mixture, the distance to CBD and number of bus lines are three most important factors on affecting weekend metro usage.

Among these three factors, land use mixture and distance to CBD are found to be negative associated with weekend metro ridership, while number of bus lines is found to be positive related with weekend metro usage. Higher land use mixture usually represents more average land use types within the station catchment area. However, many metro users take metro on weekend for a specific purpose (e.g., shopping, food, or tourism), and stations with relatively lower land use mixture may thus have more metro riders. It is intuitive that distance to CBD is negative related with weekend metro usage. Due to the traffic jam and shortage of parking spaces within CBD area in megacities like Shanghai, driving to the CBD on weekends may not as convenient as taking the metro. Therefore, metro stations which are close to the CBD can attract more metro users during the weekend. More bus lines near the metro station can provide sufficient first/last-mile services for

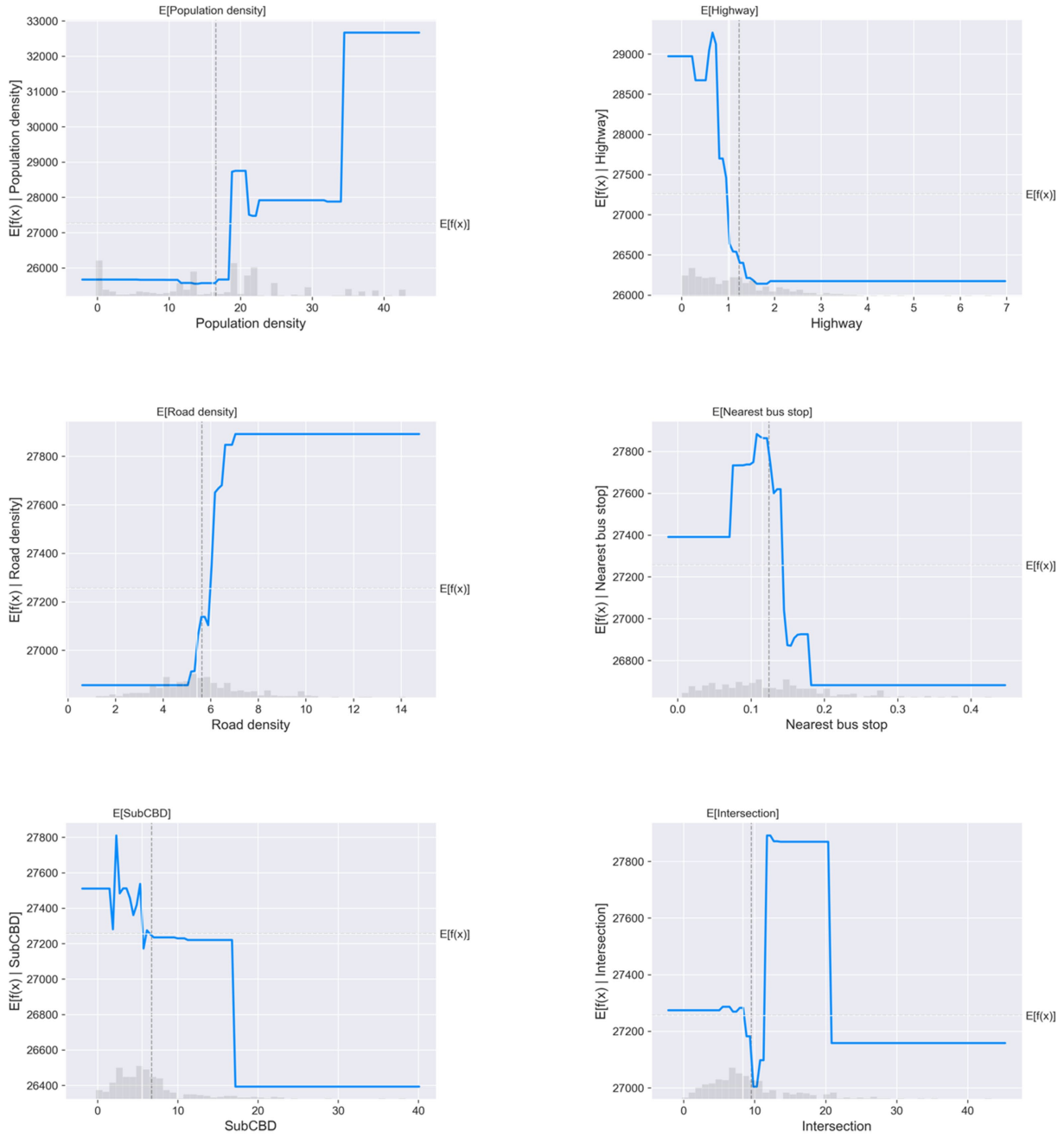

FIGURE 4
Partial dependence plots on weekend metro usage.

metro users to access the metro station on weekend, which enlarge the station catchment area and facilitate weekend metro usage.

Based on the results of partial dependence plots, all the built environment variables show non-linear impacts on weekend metro usage with certain threshold and effective ranges (Table 3). Weekend metro ridership shows a significant decrease when land use mixture moves from 0.65 to 0.75. The complex relationship between land-use diversity and weekday or weekend metro usage is also found by many literature in different contexts (4, 14, 20).

The distance to CBD is negatively related with weekend metro usage between 2 to 10 km, which seems to be reasonable that metro stations near city center may have densified population and thus more metro passengers. The distance to CBD has no significant impact on weekend metro usage when it is beyond this range. The sharp rise of weekend metro usage has been found when number of bus lines increases from 15 to 20 and 30 to 35, while the ridership remains nearly constant when number of bus lines is within other ranges. Existing literature has also suggested the positive impacts of bus lines on metro usage, while the impacts can be mediated if bus route is more than 40 (4). Rooftop area has a positive association with weekend metro ridership, with a dramatic rise between 0.20 and 0.25. This indicates that high level of land use development can facilitate the weekend metro usage, but excessive development may have trivial effects on further increase. Number of bus stops has positive effects on weekend metro ridership, with an effective

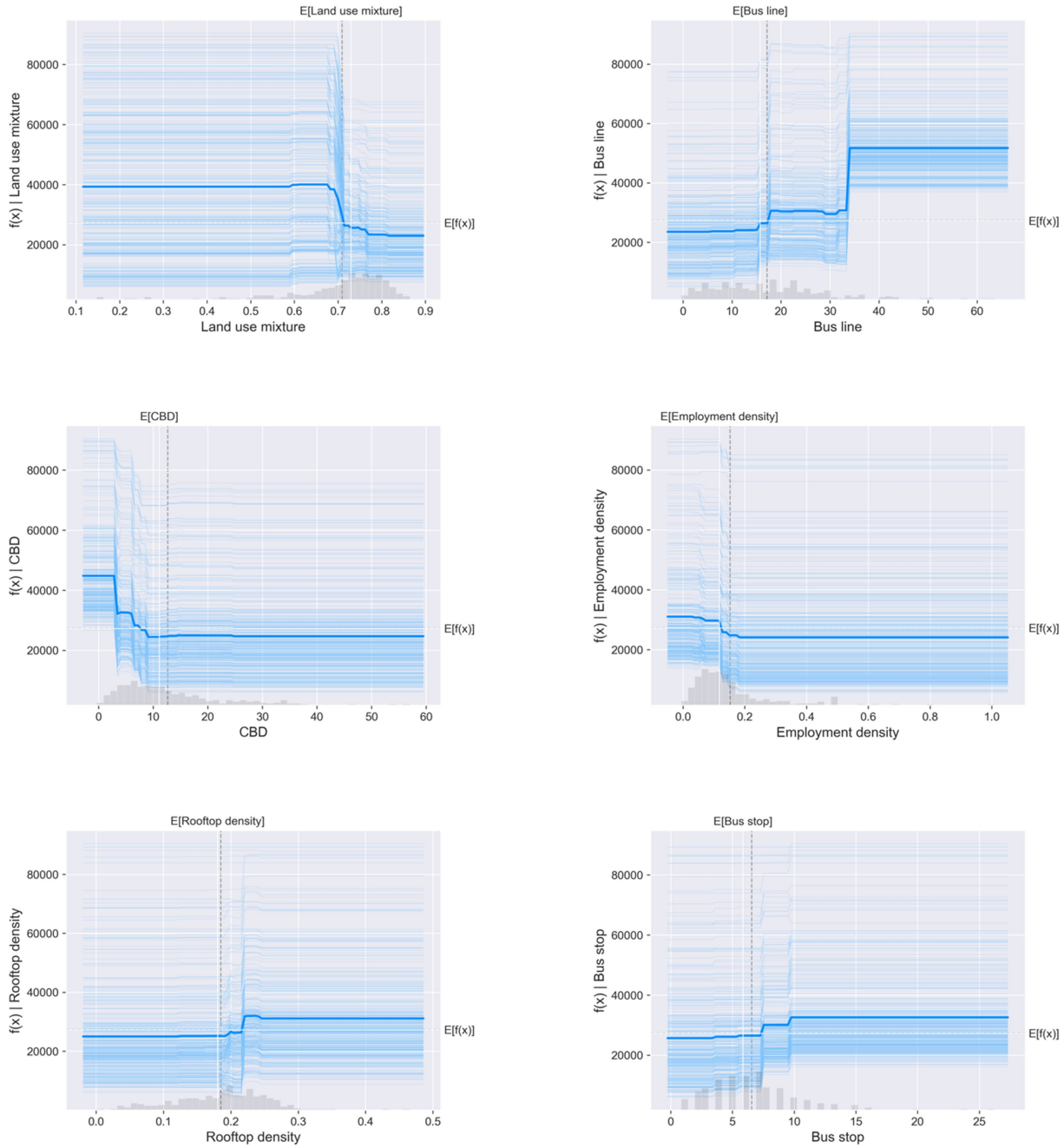

FIGURE 5 (Continued)

range between 3 and 10. Similar threshold impacts of bus stops are found in different cities but with different thresholds (13, 14).

6 Conclusion

To improve the public environmental health by facilitating metro usage on weekend, this study employed GBDT approach to evaluate the non-linear and threshold impacts of the built environment on weekend metro ridership. Compared to conventional models with linear presumption, investigating the non-linear effects help policymakers and urban planners recognize the thresholds and effective ranges of the built environment characteristics, which can benefit public health by making customized strategies and policy interventions. The empirical finding may contribute threefold to the existing studies.

First, this study estimates the feature importance of built environment characteristics in predicting weekend metro ridership. According to the results, the top-five variables with highest importance are land use mixture (16.26%), distance to CBD (13.61%), bus line (12.17%), employment density (10.06%) and rooftop density (10.05%). Results can help urban planners identify the role of different built environment characteristic and issue differentiation strategies.

Second, it depicts SHAP beeswarm plot to show the impact of each variable on the prediction. The top-5 important variables by

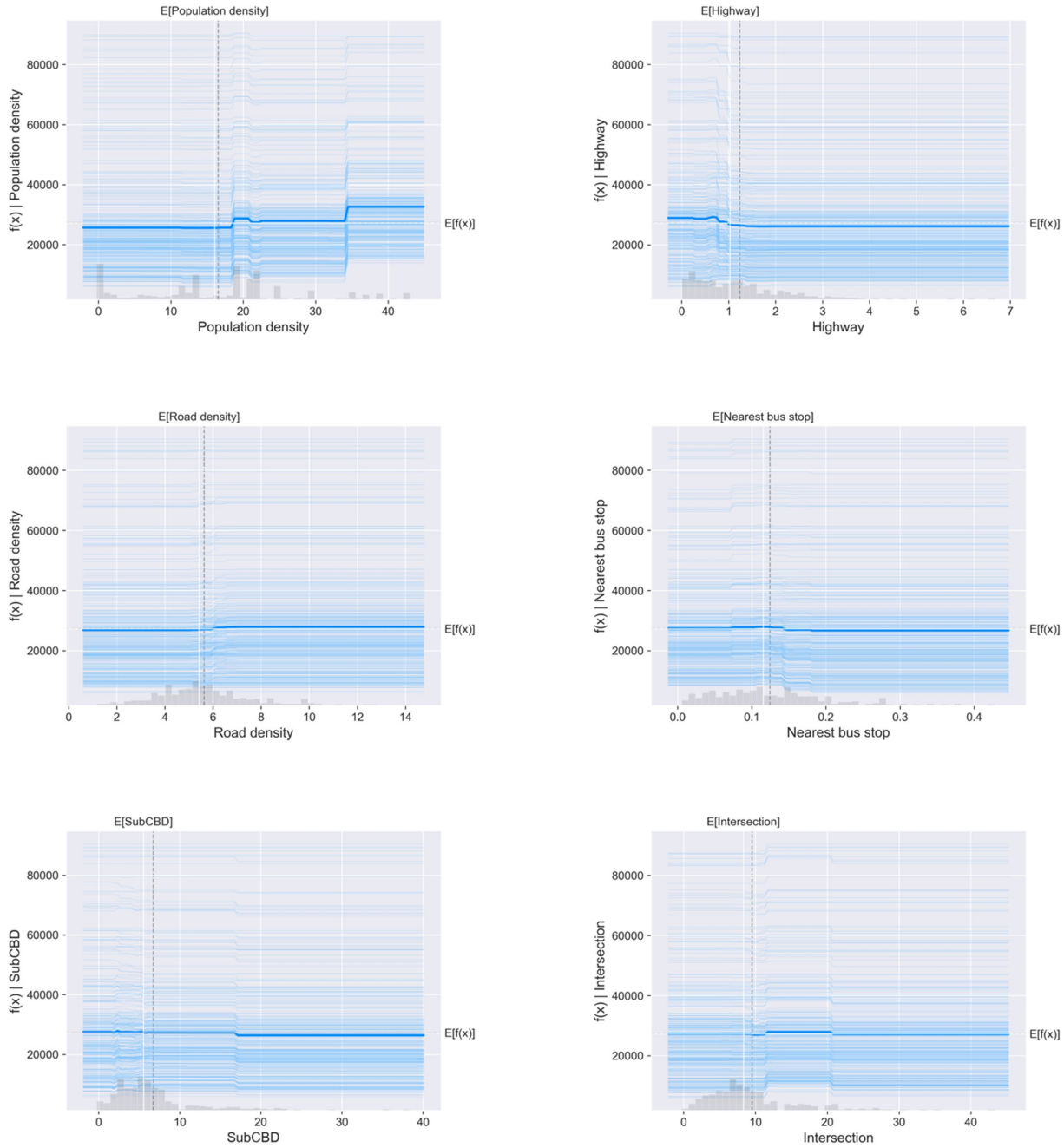

FIGURE 5
Combination of PDPs and ICEs of built environment on metro usage.

SHAP beeswarm plot are same to relative importance. Bus line, rooftop density, bus stop and population density are positively related with SHAP value, while land use mixture, distance to CBD and employment density are negativity associated with SHAP value. Therefore, urban designers should pay different attention to the built environment characteristics to promote metro usage.

Third, we depict the non-linear impacts of the built environment by combining PDPs with ICEs. Most variables have obvious thresholds on determining weekend metro ridership. Results show that maximum weekend ridership occurs when land use mixture entropy is smaller than 0.7, number of bus lines reaches 35, rooftop density reaches 0.25, and number of bus stops reaches 10. The non-linear relationship and their effective ranges help policymakers increase metro ridership on weekends by optimizing station-level land use.

Several limitations merit further study. First, the influences of built environment features on weekend metro usage may vary in different contexts. Therefore, relevant studies are encouraged to explore or validate the non-linear associations between the built environment and weekend metro ridership. Second, this study uses the 500 m buffer for most independent variables, while 400 m buffer (13) and 800 m (20) are used by different station-level built environment studies. Because the real service area of metro stations

may vary in different cities and stations, future studies are welcome to testify the results with different buffer zones. Third, we only explore the effects of a limited number of built environment characteristics. With the development of big data and GIS, more comprehensive built environment attributes (e.g., number of parking spaces, demographics, sidewalk density) with finer data are welcomed for further exploration in different contexts. Fourth, PDPs may be misguided when independent variables are correlated with each other (e.g., bus stop and bus line), while accumulated local effects (ALE) plots can be used as an unbiased alternative to address the multicollinearity issue in further studies. Fifth, most data used in this study are before the pandemic, while the comparison between pre-pandemic and post-pandemic need further exploration in the future.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

BP: Conceptualization, Data curation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. TW: Data curation, Resources, Supervision, Writing – review & editing. YZ: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. CL: Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing.

Funding

This study is supported by National Social Science Foundation (No. 22AZD082), Shanghai Social Science Foundation (Nos. 2023BSH003, 22Z350204369, and 2022BSH005), Shanghai Scientific Research Foundation (Nos. 23DZ1202900, 23DZ1203200, 23DZ1202400, 22DZ1203200, 21Z510203259, and 21DZ1200800), Special Project of Healthy Shanghai Action (No. JKSHZX_2022–13), and the Scientific Research Fund (No. K2015K017).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Liu B, Xu Y, Guo S, Yu M, Lin Z, Yang H. Examining the nonlinear impacts of origin-destination built environment on metro ridership at Station-To-Station level. *ISPRS Int J Geo Inf*. (2023) 12:59. doi: 10.3390/ijgi12020059

2. Wu J, Li C, Zhu L, Liu X, Peng B, Wang T, et al. Nonlinear and threshold effects of built environment on older adults' walking duration: do age and retirement status matter? *Front Public Health*. (2024) 12:1418733. doi: 10.3389/fpubh.2024.1418733

3. Calthorpe P. The next American metropolis: Ecology, community, and the American dream. NYC, USA: Princeton Architectural Press (1993).

4. Gan Z, Yang M, Feng T, Timmermans HJP. Examining the relationship between built environment and metro ridership at station-to-station level. *Transp Res Part D: Transp Environ*. (2020) 82:102332. doi: 10.1016/j.trd.2020.102332

5. CAMET. Overview of urban rail transit lines in Mainland China in 2022. Beijing, China: China Association of Metros (2023).

6. Kuby M, Barranda A, Upchurch C. Factors influencing light-rail station boardings in the United States. *Transp Res A Policy Pract*. (2004) 38:223–47. doi: 10.1016/j.tra.2003.10.006

7. Ma X, Liu C, Wen H, Wang Y, Wu YJ. Understanding commuting patterns using transit smart card data. *J Transp Geogr*. (2017) 58:135–45. doi: 10.1016/j.jtrangeo.2016.12.001

8. He Y, Zhao Y, Tsui K-L. Modeling and analyzing impact factors of metro station ridership: An approach based on a general estimating equation. *IEEE Intell Transp Syst Mag*. (2020) 12:195–207. doi: 10.1109/MITS.2020.3014438

9. Cao J, Tao T. Using machine-learning models to understand nonlinear relationships between land use and travel. *Transp Res Part D: Transp Environ*. (2023) 123:103930. doi: 10.1016/j.trd.2023.103930

10. Zhao J, Deng W, Song Y, Zhu Y. What influences metro station ridership in China? Insights from Nanjing. *Cities*. (2013) 35:114–24. doi: 10.1016/j.cities.2013.07.002

11. 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 D.C. Metrorail system. *Transp Res A Policy Pract*. (2018) 116:635–49. doi: 10.1016/j.tra.2018.06.011

12. Li S, Lyu D, Huang G, Zhang X, Gao F, Chen Y, et al. Spatially varying impacts of built environment factors on rail transit ridership at station level: a case study in Guangzhou, China. *J Transp Geogr*. (2020) 82:102631. doi: 10.1016/j.jtrangeo.2019.102631

13. Ding C, Cao X, Liu C. How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds. *J Transp Geogr*. (2019) 77:70–8. doi: 10.1016/j.jtrangeo.2019.04.011

14. Shao Q, Zhang W, Cao X, Yang J, Yin J. Threshold and moderating effects of land use on metro ridership in Shenzhen: implications for TOD planning. *J Transp Geogr*. (2020) 89:102878. doi: 10.1016/j.jtrangeo.2020.102878

15. Caigang Z, Shaoying L, Zhangzhi T, Feng G, Zhifeng W. Nonlinear and threshold effects of traffic condition and built environment on dockless bike sharing at street level. *J Transp Geogr*. (2022) 102:103375. doi: 10.1016/j.jtrangeo.2022.103375

16. Yang H, Zheng R, Li X, Huo J, Yang L, Zhu T. Nonlinear and threshold effects of the built environment on e-scooter sharing ridership. *J Transp Geogr*. (2022) 104:103453. doi: 10.1016/j.jtrangeo.2022.103453

17. Yang H, Luo P, Li C, Zhai G, Yeh AGO. Nonlinear effects of fare discounts and built environment on ridesplitting adoption rates. *Transp Res A Policy Pract*. (2023) 169:103577. doi: 10.1016/j.tra.2022.103577

18. Tao T, Næss P. Exploring nonlinear built environment effects on driving with a mixed-methods approach. *Transp Res Part D: Transp Environ*. (2022) 111:103443. doi: 10.1016/j.trd.2022.103443

19. Jin T, Cheng L, Zhang X, Cao J, Qian X, Witlox F. Nonlinear effects of the built environment on metro-integrated ridesourcing usage. *Transp Res Part D: Transp Environ*. (2022) 110:103426. doi: 10.1016/j.trd.2022.103426

20. 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. doi: 10.1016/j.tust.2022.104931

21. Ewing R, Cervero R. Travel and the built environment: a meta-analysis. *J Am Plan Assoc*. (2010) 76:265–94. doi: 10.1080/01944361003766766

22. Cervero R. Alternative approaches to modeling the travel-demand impacts of smart growth. *J Am Plan Assoc*. (2006) 72:285–95. doi: 10.1080/01944360608976751

23. Durning M, Townsend C. Direct ridership model of rail rapid transit systems in Canada. *Transp Res Rec*. (2015) 2537:96–102. doi: 10.3141/2537-11

24. Zhao J, Deng W, Song Y, Zhu Y. Analysis of metro ridership at station level and station-to-station level in Nanjing: an approach based on direct demand models. *Transportation*. (2014) 41:133–55. doi: 10.1007/s11116-013-9492-3

25. Loo BPY, Chen C, Chan ETH. Rail-based transit-oriented development: lessons from New York City and Hong Kong. *Landsc Urban Plan*. (2010) 97:202–12. doi: 10.1016/j.landurbplan.2010.06.002

26. Huang J, Chen S, Xu Q, Chen Y, Hu J. Relationship between built environment characteristics of TOD and subway ridership: a causal inference and regression analysis of the Beijing subway. *J Rail Transport Plan Manag*. (2022) 24:100341. doi: 10.1016/j.jrtpm.2022.100341

27. Gutiérrez J, Cardozo OD, García-Palomares JC. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *J Transp Geogr*. (2011) 19:1081–92. doi: 10.1016/j.jtrangeo.2011.05.004

28. Jun M-J, Choi K, Jeong JE, Kwon KH, Kim HJ. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *J Transp Geogr*. (2015) 48:30–40. doi: 10.1016/j.jtrangeo.2015.08.002

29. Ryan S, Frank LF. Pedestrian environments and transit ridership. *J Public Transp*. (2009) 12:39–57. doi: 10.5038/2375-0901.12.1.3

30. Cardozo OD, García-Palomares JC, Gutiérrez J. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Appl Geogr*. (2012) 34:548–58. doi: 10.1016/j.apgeog.2012.01.005

31. Liu C, Erdogan S, Ma T, Ducca FW. How to increase rail ridership in Maryland: direct ridership models for policy guidance. *J Urban Plan Develop*. (2016) 142:04016017. doi: 10.1061/(ASCE)UP.1943-5444.0000340

32. Ewing R, Hamidi S, Gallivan F, Nelson AC, Grace JB. Combined effects of compact development, transportation investments, and road user pricing on vehicle miles traveled in urbanized areas. *Transp Res Rec*. (2013) 2397:117–24. doi: 10.3141/2397-14

33. Tu W, Cao R, Yue Y, Zhou B, Li Q, Li Q. Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *J Transp Geogr*. (2018) 69:45–57. doi: 10.1016/j.jtrangeo.2018.04.013

34. Gan Z, Feng T, Yang M, Timmermans H, Luo J. Analysis of metro station ridership considering spatial heterogeneity. *Chin Geogr Sci*. (2019) 29:1065–77. doi: 10.1007/s11769-019-1065-8

35. An D, Tong X, Liu K, Chan EHW. Understanding the impact of built environment on metro ridership using open source in Shanghai. *Cities*. (2019) 93:177–87. doi: 10.1016/j.cities.2019.05.013

36. Li S, Lyu D, Liu X, Tan Z, Gao F, Huang G, et al. The varying patterns of rail transit ridership and their relationships with fine-scale built environment factors: big data analytics from Guangzhou. *Cities*. (2020) 99:102580. doi: 10.1016/j.cities.2019.102580

37. Gao F, Yang L, Han C, Tang J, Li Z. A network-distance-based geographically weighted regression model to examine spatiotemporal effects of station-level built environments on metro ridership. *J Transp Geogr*. (2022) 105:103472. doi: 10.1016/j.jtrangeo.2022.103472

38. Zhang Z, Qian Z, Zhong T, Chen M, Zhang K, Yang Y, et al. Vectorized rooftop area data for 90 cities in China. *Scientific Data*. (2022) 9:66. doi: 10.1038/s41597-022-01168-x

39. SUPLRAB, Shanghai Master Plan (2017–2035) (2018) Shanghai, China: Shanghai Urban Planning and Land Resource Administration Bureau.

40. Ding C, Cao X, Wang Y. Synergistic effects of the built environment and commuting programs on commute mode choice. *Transp Res A Policy Pract*. (2018) 118:104–18. doi: 10.1016/j.tra.2018.08.041

41. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. (2001) 29:1189–232. doi: 10.1214/aos/1013203451

42. Ding C, Wang D, Ma X, Li H. Predicting short-term Subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability*. (2016) 8. doi: 10.3390/su8111100

43. Hastie T, et al. The elements of statistical learning: data mining, inference, and prediction, vol. *2* NYC, USA: Springer (2009).

44. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Proces Syst*. (2017) 30

45. Shapley L.S., A value for n-person games (1953). doi: 10.7249/P0295