Effects of buffer size on associations between the built environment and metro ridership: A machine learning-based sensitive analysis

Xiang Liu[a], Xiaohong Chen[a,b,*], Mingshu Tian[b,*], Jonas De Vos[c]

[a] *Urban Mobility Institute, Tongji University, Shanghai, China*
[b] *Key Laboratory of Road and Traffic Engineering of the Ministry of Education, School of Transportation Engineering, Tongji University, Shanghai, China*
[c] *Bartlett School of Planning, University College London, London, UK*

Abstract

Uncertainty in the relevant buffer size of metro station catchment areas may drive inconsistencies in the findings on the built environment and metro ridership. Although previous studies estimate the effect of this uncertainty, the results are far from definitive. By utilizing finer-grained big data and non-parametric machine learning approaches, this study conducted a sensitivity analysis defining built environment factors within four radial buffer sizes: 300 m, 600 m, 800 m, and 1000 m on associations with metro ridership. The results suggest that: (1) different buffer sizes have little influence on the ordinary least-squares model's predictive power, but significant influence on the machine learning model; (2) the use of a 600 m buffer size around the transit station demonstrates the best model fit and variation explanation compared to others; (3) findings on the relative importance, ranks, and nonlinear associations with metro ridership can be impacted as the choice of geographic delineation of buffer sizes deviate from the true relevant geographic context of built environment variables. The results assist planners in setting a benchmark for metro catchment areas for station-area planning and demand forecasting, more importantly, the findings highlight the importance of meticulously selecting the analytical spatial unit for area-based variables, especially when utilizing non-parametric machine learning approaches in research.

Keywords: buffer size; catchment area; sensitive analysis; built environment; metro ridership; XGBoost

* Corresponding author at: 4800 Cao'an Road, Shanghai 201804, China
E-mail addresses: xliu02@tongji.edu.cn (X. Liu), tongjicxh@163.com (X. Chen), 2133372@tongji.edu.cn (M. Tian), jonas.devos@ucl.ac.uk (J. De Vos).

1. Introduction

Many countries around the world have experienced rapid urbanization and increased motorization over the past decades, while negative externalities such as urban sprawl, traffic congestion, and environmental pollution are pervasive throughout many megacities (Batty et al., 2003). Since the early 1990s, the concept of transit-oriented development (TOD) has garnered widespread attention (Calthorpe, 1993) and has emerged as a promising instrument for mitigating automobile dependence and urban sprawl in numerous countries (Cervero, 1998; Cervero et al., 2002; Bertolini et al., 2012; Nasri and Zhang, 2014; Papa and Bertolini, 2015; Xu et al., 2017).

To understand the determinants of metro ridership and to forecast metro demand, many studies use the direct ridership model (DRM) to investigate the relationship between the station-area built environment and metro ridership (An et al., 2019; Ding et al., 2019; Loo et al., 2010). Compared to the conventional four-step model (i.e., trip generation, trip distribution, mode choice, and traffic assignment), the DRM offers a cost-effective alternative because of ‘‘rapid response, simplicity of use, ease of results interpretation, low information requirements and low cost’’ (Cardozo et al., 2012: 556). Beyond forecasting, the elasticity values in the DRM expose the potential consequences of urban planning strategies on transit ridership, such as the introduction of new urban developments or the densification of existing ones, which are particularly crucial to TOD (Cervero, 2006). However, despite the DRM being widely used in the literature, the findings on the relationship between the built environment over the past three decades have been inconsistent, the discrepancies across studies could be attributed to the heterogeneity in city context, analytical methodology, trip heterogeneity, and notably, the delineation of catchment areas (An et al., 2019; Chen et al., 2022).

Most DRMs typically apply a radius buffer ranging from 300 m to 1000 m, rather

arbitrarily, as the catchment area for a metro station (Guerra et al., 2012; Jun et al., 2015). This distance range is seemingly referred to as walking distances within 5-15 minutes, but it is still unclear what the most appropriate context is for understanding the relationship between the built environment and metro ridership. Although some studies assert that varying catchment areas have minimal influence on a DRM's predictive power for station-level metro ridership (Guerra et al., 2012), others argue that the effects of the station-area built environment vary significantly based on the choice of buffer sizes, as demonstrated in Seoul (Jun et al., 2015). This inconsistency in defining buffers underscores the problem of uncertain geographic context, which is defined as "the spatial uncertainty in the actual areas that exert the contextual influences under study" by Kwan (2012). This uncertainty arises from an incomplete understanding of the "true causally relevant" catchment area where the built environment exerts effects on metro ridership.

However, capturing the "true causally relevant" catchment area where the built environment exerts effects on metro ridership is not easy. Firstly, early studies often defined a station-area built environment based on administrative boundary (e.g., census tracts or ZIP codes) or Transport Analysis Zone (TAZ) (Ding et al., 2019; Guerra et al., 2012; Shao et al., 2020). While administrative boundaries or Traffic Analysis Zones (TAZ) provide convenience in data collection for statistics or travel surveys, they can inadvertently lead to issues due to their coarse-grained nature. These issues include aggregation biases or the modifiable areal unit problem (MAUP), which is a source of statistical bias that can significantly impact the results of statistical hypothesis tests, in other words, the scale and zoning pattern of the catchment area used may lead to different results (James et al., 2014). For example, even though metro station areas typically feature higher employment density than other regions, due to the use of TAZ

as the spatial analytical unit, studies are compelled to assume a uniform distribution of employment. This assumption can inadvertently lead to a misrepresentation of the true spatial distribution of employment, especially in cases where the employment density significantly varies within the TAZ (Shao et al., 2020).

Secondly, previous DRMs often assume that station-based built environment factors have a pre-defined generalized linear relationship with metro ridership. This assumption is based on the general principle that the relationship between the built environment and metro ridership is likely to be proportional, meaning that an increase in one factor leads to a corresponding increase in the other. However, an ordinary least square (OLS) regression approach has been criticized for its ability to account for spatial variation and nonlinear correlations. To deal with the spatial autocorrelation problem, some studies adopted geographically weighted regression (GWR) to examine the relationship between the built environment and metro ridership (Cardozo et al., 2012; Jun et al., 2015). The results confirmed that the effect of the built environment varies across space. However, the GWR approach has its limitations when it comes to investigating the non-linear and threshold effects of the built environment and metro ridership (Shao et al., 2020). In this case, the predictive power and relevance of the DRM might be constrained by the modelling approaches, and the results of the "true causally relevant" catchment area can still be unrobust.

The rise of big data (e.g., mobile phone signal, smart card data, point of interest) and new modelling methods (e.g., machine learning) provide possible means for mitigating the uncertain geographic context problem with regard to the availability of detailed spatial data with advanced models capable of handling intricate nonlinear relationships and interactions between independent and dependent variables. On the one hand, big data offers further promise to expand the representation of the hard-to-reach

population, land use, and spatial features, with great opportunities to overcome the aggregation biases caused by coarse-grained data. On the other hand, machine learning approaches relax the assumption of generalized linearity and illustrate more complex associations between the built environment and metro ridership while accounting for spatial effects (Ding et al., 2019; Li, 2022). It may, therefore, be interesting to see whether new (finer-grained) data and new (machine learning) methods can offer an innovative insight into the effects of buffer size on the relationships between the built environment and metro ridership.

This study aims to use DRM through the application of eXtreme Gradient Boosting (XGBoost) and Shapley Addictive exPlanations (SHAP) to conduct a sensitivity analysis of the uncertain buffer problem in the built environment and metro ridership. It addresses the following research questions: (1) Do different buffer size choices influence the performance of the direct ridership model? (2) If the answer is “yes”, which buffer size has the best model’s predictive power? (3) To what extent do buffer size choices impact the correlations between the built environment and metro ridership?

This study provides a threefold contribution to the existing body of literature. Firstly, it is among the first to use a non-parametric approach and finer-grain data to examine the impact of varying analytical buffer sizes on the relationship between the built environment and metro ridership. Secondly, it provides valuable new empirical evidence regarding the optimal buffer size for maximizing the model's predictive power in the direct ridership model. Thirdly, it uncovers the extent to which buffer effects influence the results of machine learning models.

The structure of this paper unfolds as follows. Section 2 offers a review of existing literature. Section 3 outlines the data employed in this study, delving into the variables and the modelling approaches utilized. Section 4 presents the analytical findings.

Section 5 concludes the paper with the key findings, implications, and limitations.

2. Literature review

2.1 The built environment and metro ridership

Especially over the past three decades, urban sprawl and associated externalities have promised a substantial amount of studies on the relationship between the built environment and travel behaviour (Ewing and Cervero, 2010). The urban and transport strategy of TOD has been generating considerable interest in academics and practices (Bertolini et al., 2012; Calthorpe, 1993; Cervero et al., 2002). Among the TOD studies, the relationship between the built environment and metro ridership is the main focus (Ding et al., 2019).

Prior studies often adopt direct ridership models to study the relationship between the station-area built environment and metro demand (Cervero, 2006). The model has advantages over the traditional four-step model, including simplicity of use, less data requirement, ease of interpretation, quick response and low costs (Cardozo et al., 2012). Although DRMs lack the detailed decision-making process of travellers like the four-step model, they allow for in-depth analysis of the station environment, making them well-suited for studies that aim to understand factors influencing transit use. Additionally, DRMs utilize station catchment areas as the unit of analysis, as opposed to the traffic analysis zone (TAZs) used in the traditional four-step model. This approach aligns more sensibly with the scope of station-level metro ridership modelling, given that the scale of TAZs may be too extensive. DRMs are more adept at capturing the impact of the station-area built environment on metro ridership (Cervero, 2006; Ding et al., 2019). This methodology is also advantageous in harnessing Geographic Information Systems (GIS) as a tool for calculating these variables, alongside using

spatial disaggregation methods that permit a detailed study of the station-area built environment (what Cervero (2006) calls “fine-grained design details”).

DRMs in the literature show that land use variables are key determinants of station-level ridership. Cervero (2006) found that a 10% increase in residential density within 0.5 miles of stations was associated with a 1.9% increase in daily metro ridership. Zhao et al. (2013) conducted an empirical study using large-scale smart card data in Nanjing and found that population density and employment density had a positive influence on intermodal transit trips, while population density had a larger influence on intermodal transit trips than employment density. Similar results were found by Kuby et al. (2004) and Jun et al. (2015).

Land use diversity was found to have a significant effect on metro ridership. Li et al. (2020) discovered a strong correlation between land use diversity and increased ridership in Guangzhou, with the impact being notably higher in outer suburban areas for morning boardings and evening alightings. Shao et al. (2020) reported a significant influence of land use diversity on metro ridership, particularly when the entropy index of the diversity fell between 0.5 and 0.65. However, Chen et al. (2022) found that land use diversity has no significant associations with metro ridership in Wuhan. These conflicting findings underscore the need for further exploration and research to discern the true impact of land use diversity on station-level metro ridership.

Previous studies also investigate the effect of design factors on metro ridership. Chen et al. (2022) found a positive correlation between increased intersection density and higher metro ridership. The rationale behind this association is that higher intersection density enhances accessibility to metro stations and diverse destinations. Specifically, a more pedestrian-oriented street network surrounding metro stations positively impacts metro ridership, as it can effectively reduce the walking distance to

metro stations and offer a wider variety of route choices for commuters (Nasri and Zhang, 2014). Similar results can be found by Ding et al. (2019).

Distance to Central Business District (CBD) is another crucial factor affecting metro ridership. Andersson et al. (2021) and Zhao et al. (2013) found that stations located closer to CBD tend to have higher ridership due to the concentration of job opportunities, commercial activities, and services in the city centre. However, some studies found an insignificant correlation between the distance to CBD and metro ridership (Shao et al., 2020; An et al., 2019). The findings imply that the impact of proximity to CBD on metro usage may vary, particularly in polycentric urban configurations featuring multiple activity hubs or centres.

Station-level metro ridership is affected not only by the built environment but also by station characteristics. Some studies found that terminal or transfer stations were significantly associated with more metro ridership. Sohn and Shim (2010) used the betweenness centrality, closeness centrality, and straightness centrality to represent the network efficiency for each station in Seoul. The results showed that high betweenness centrality has significant positive effects on metro ridership. Similar results were found in Shanghai (An et al., 2019) and Shenzhen (Shao et al., 2020). Previous studies further examined the effect of station age on metro ridership (Pan et al., 2017), demonstrating that stations with longer operating histories tend to have higher ridership. These findings underscore the co-evolution of urban development around metro stations and behavioural adaptation by commuters, and suggesting co-evolution may unfold over extended periods (Deng and Zhao, 2022).

2.2 A benchmark for transit catchment area

While recent years have seen widespread empirical contributions to the behavioural mechanisms underlying how station-area built environment variables affect metro ridership, research on the built environment over the last three decades has been

inconsistent. The inconsistency across studies is attributed to differences in the local context, analytical methods, travel purposes, and specifically, the heterogeneity in the delineation of metro catchment areas across studies, which impedes understanding the totality of the evidence of the built environment's impact on metro ridership (James et al., 2014).

Previous studies often define a radial buffer of 300-1000 m as the most acceptable buffer size as it is seemingly referred to walking distances within 5-15 minutes, but there is still no consensus on the most appropriate catchment radius of metro station. In the United States, the 0.5 miles (roughly equivalent to 800 meters) radius is the most commonly accepted distance for gauging a metro station's catchment area and is the de facto standard for TOD planning (Guerra et al., 2012; Kuby et al., 2004). O'Sullivan and Morrall (1996) reported that in Canada the walking distance guidelines applied to all surface transport modes range from 300 m to 900 m. In South Korea, Seoul Metropolitan Government set an aerial distance of 500 m in radius as a standard for transportation studies (Sohn and Shim, 2010). In Shanghai, the official "Shanghai Master Plan 2017-2035" recommended 600 m as a standard distance for metro planning (An et al., 2019), but the 800 m radius is more frequently used in Beijing (Li and Zhao, 2017), and Shao et al. (2020) used 1000 m as a catchment area radius to capture the correlates of land use on metro ridership in Shenzhen.

This lack of consistency in a benchmark for the buffer sizes connects to the "uncertain geographic context problem" defined by Kwan (2012), which refers to the impact of area-based contextual variables on individual behaviours or outcomes being influenced by the way geographic units (e.g., TAZ) are delineated. The uncertain geographic context problem arises from a limited understanding of the "truly causally relevant" area in which the environment influences individual behaviours (Kwan, 2012).

In other words, inaccurate selection of geographic boundaries for contextual units (e.g., buffer sizes) may lead to flawed or misleading conclusions about the relationship between the built environment and metro ridership. To mitigate the uncertain geographic context problem, Clark and Scott (2014) suggested three principles for selecting the appropriate scale to measure the built environment. First, selecting a scale related to the actual travel distances to/from the metro station; Second, using disaggregate data whenever possible; Third, choosing a scale that is appropriate for the policy-relevant. James et al. (2014) further added that conducting a sensitivity analysis of the uncertain geographic context problem in relation to the built environment and metro ridership is necessary.

Guerra et al. (2012) conducted a sensitivity analysis of different buffer sizes by utilizing station-level built environment variables from 1,449 high-capacity transit stations across 21 American cities. The findings suggested that various catchment areas exerted little influence on a model's predictive power regarding station-level transit ridership. For the six catchment-area radii analyzed, the R^2 for population fluctuated between 0.742 and 0.746, while for jobs, it varied from 0.723 to 0.745. However, it is worth noting that their study used fairly aggregated data (e.g., job data at the zip-code level, and population estimates at block level). This aggregation at the station level (i.e., within the 800 m network distance buffer) may introduce aggregation biases due to the coarse-grained nature of the data (Shao et al., 2020). Additionally, the study utilized OLS to examine the predictive power of DRM among different catchment-area radii, OLS might exhibit biases owing to its inherent restrictions, such as presuming linear relationships and being incapable of adequately addressing spatial autocorrelation (Cardozo et al., 2012).

Jun et al. (2015) utilized a mixed geographically weighted regression model

(MWGR) along with more disaggregated geographic units to investigate the variations in built environment correlates of metro ridership across different pedestrian catchment areas (PCAs). The study discovered that the metro ridership impacts of the built environment vary significantly by PCA. The best-fit value of the MWGR model ranged from 0.773 for core PCAs (with a radius of 300 m) to 0.679 for secondary PCAs (covering 600-900 m). The results suggest that the impact of varying catchment-area radii cannot be disregarded. The selection of buffer size significantly influences the findings concerning the relationship between the built environment and ridership. The contradictory findings call for additional studies to investigate the effect of buffer sizes on the relationship between the built environment and metro ridership, especially utilizing high-resolution spatial data and advanced modelling techniques.

2.3 Analysis with new data and new approach

Over the last decade, the rise of big data, particularly on the built environment side generate a new potential to conduct a finer-grained sensitivity analysis of the uncertain buffer size in relation to the built environment and metro ridership. Big data, such as OpenStreetMap (OSM), Point of Interests (POIs), and smart card data (SCD) have been increasingly utilized in various research fields to derive insights into land use, transportation systems, and travel behaviours (An et al., 2019). Compared to frequently used census tracts, blocks or TAZs data, big data offers some promise to expand the representation of hard-to-reach built environment information (e.g., population density and employment density), with great opportunities to improve the accuracy of sensitivity analysis of buffer size effects on associations between the built environment and metro ridership (Kwan, 2012).

Additionally, advanced analytical approaches, such as machine learning, offer a new tool to reveal complications that are less observable in traditional linear models (Shao et al., 2020). The assumption of a linear relationship between dependent and

independent variables limits the linear model performance when the relationship between these variables is nonlinear or irregular (van Wee and Handy, 2016). This is particularly probable in the context of TOD, given its characteristic high density, mixed-use, and compact nature. In other words, the inherent limitation in traditional linear models could result in suboptimal or inaccurate predictions and may fail to capture important aspects of the sophisticated effects of buffer size. Moreover, traditional linear statistical models often struggle with such unobservable heterogeneity (Tang et al., 2020). They operate under the assumption that relationships among variables are uniform across all observations. However, in reality, this is rarely the case. Unobservable heterogeneity can lead to specification errors in these models, as they do not account for this unseen variation. This issue can result in incorrect or biased estimates of parameters, leading to misleading conclusions.

Machine learning models, on the other hand, do not rely on pre-specified model forms and are capable of capturing complex, non-linear relationships, and interactions among predictors in the relationship between the built environment and travel behaviour (Ding et al., 2019; Shao et al., 2020), which allows them to account for some of the unobservable heterogeneity. The models permit a degree of variation in estimated parameters across different observations. The variation conforms to a continuous distribution, such as a normal distribution, specified by the analyst. This flexibility in the model is designed to accommodate and account for the potential impact of unseen or unrecorded variability within the data (Tang et al., 2020). Besides, machine learning models such as SHAP-based XGBoost have emerged as viable alternatives to traditional spatial statistical models, particularly when dealing with intricate spatial and non-spatial effects that are unknown or not fully understood (Li, 2022). This makes such methods particularly suitable for addressing the “uncertain geographic context

problem" in smaller spatial units, such as metro catchment areas. The capability of these techniques to handle complex spatial relationships provides additional assurance when investigating these intricate contexts.

In summary, numerous studies have investigated the relationship between the built environment and metro ridership, but the findings remain inconsistent. The inconsistency may be due to the heterogeneity in the delineation of metro catchment areas across studies, which have not been thoroughly examined, particularly in studies using finer-grained big data and non-parametric modelling approaches. To address research gaps, this study applied a machine learning approach along with high-resolution data to examine whether and how different analytical buffer sizes affect the correlations between the built environment and metro ridership.

3. Methods

3.1 Study area and variables

This study focuses on Shanghai as its case study. As one of the most populous cities in China, Shanghai has an urbanized area of 6,340.5 km^2 and a total population of 24.9 million in 2021. By the end of 2020, Shanghai's metro network comprised 18 lines, covering approximately 729.2 km, and included 430 metro stations, thereby making it the most extensive metro network worldwide. The metro system, a vital transportation mode for Shanghai's citizens, saw an average daily ridership of 10.2 million in 2019.

The study utilized data from the Shanghai Metro, specifically, the ridership data recorded on September 18th, 2019, a typical Wednesday. This dataset encompasses daily ridership figures for 341 stations across 17 lines in Shanghai, as depicted in Fig. 1. The average daily ridership per station in Shanghai on this day was 40,391. The station with the highest recorded ridership was People's Square station, with a count of

241,742, while Wangyuan Road station observed the lowest ridership, reporting only 1,830 riders.

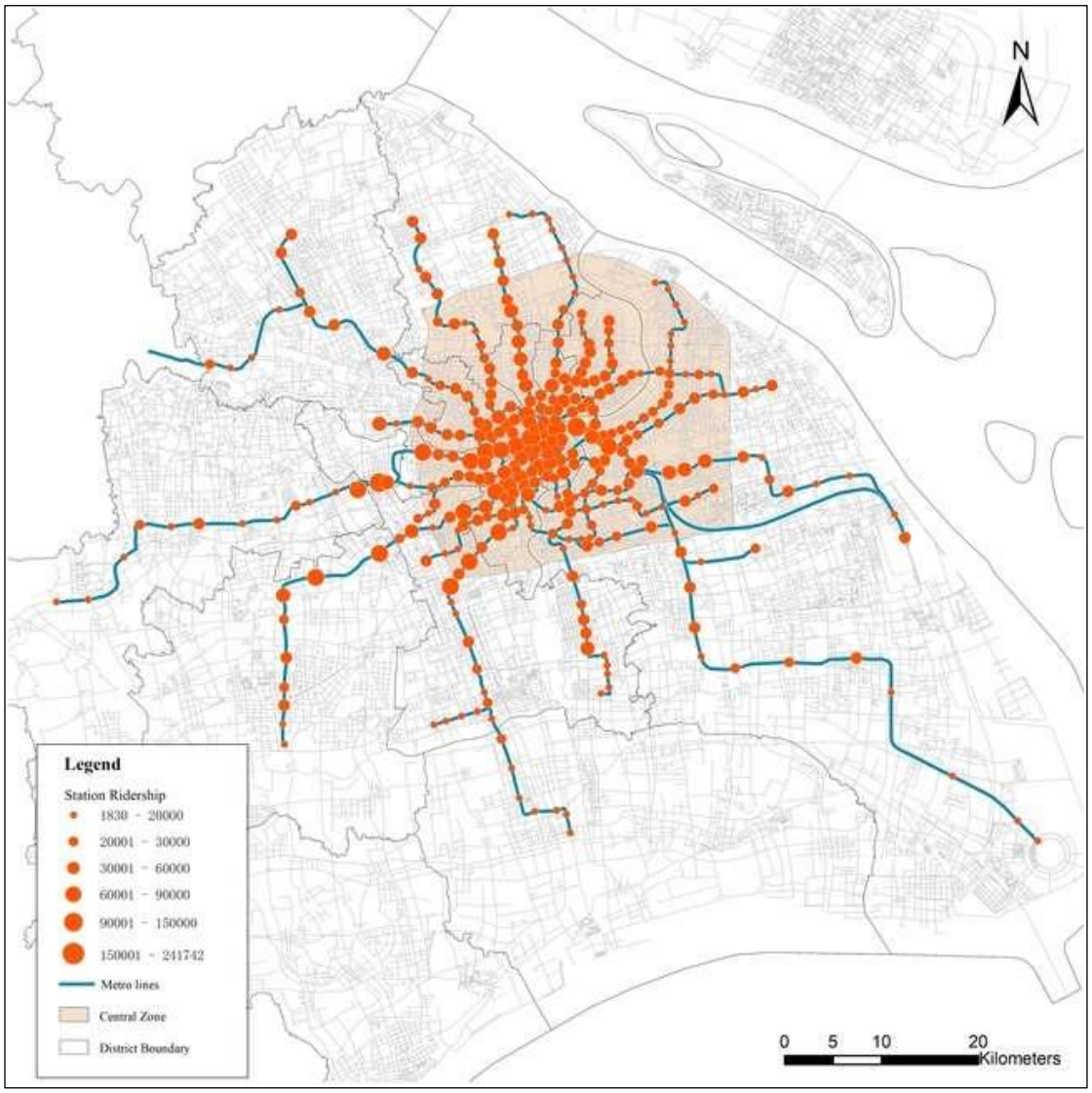

Fig.1. Study area in Shanghai and station ridership (2019).

A total of 15 independent variables were included in DRM and categorized into two groups: (1) built environment variables and (2) station characteristic variables. For built environment variables, population and employment density were measured using mobile phone signal data obtained from China Unicom, one of the largest mobile providers in China. The data provides finer-grained population density information compared to census data and has been validated repeatedly for its high quality (Yang et al., 2023). Commercial land use, business land use, public land use, and residential land

use were measured by counting the POIs within a predefined distance, prior studies have demonstrated a significant impact of these POI types on metro ridership (An et al., 2019), and this variables were included to examine the effects of different land uses on metro ridership. Land use diversity was measured using a land-use entropy approach (Cervero and Kockelman, 1997) based on POIs data from Amap (Yang et al., 2023). The average Floor Area Ratio (FAR) around metro stations was computed using ArcGIS, based on building footprint and floor data obtained from Baidu Map. Street density was measured in ArcGIS using the data from OSM. The number of bus stops was computed in ArcGIS, using data from Amap. The distance to the city centre was calculated as the straight-line distance from the metro station to the People's Square station.

With respect to station characteristic variables, we utilized the number of station entrances and the number of lines, instead of a simple transfer station binary variable, to more comprehensively represent the characteristics of metro stations. We hypothesized that these variables, by expanding service areas both at a micro and macro level, play a pivotal role in influencing metro ridership. The data on the number of entrances and lines per station were collected from shmetro.com. Betweenness centrality was calculated for each station as it provided an important measure of a station's role and importance within the network (Freeman, 1978). This metric quantifies the number of times a station acts as a bridge along the shortest path between two other stations (Sohn and Shim, 2010). Therefore, a station with high betweenness centrality has a large influence on the flow of passengers through the network, as it connects many pairs of stations. By incorporating this metric into the analysis, we can better understand how the strategic positioning of a station within the network can impact metro ridership. Drawing from Sohn and Shim (2010) as well as our preliminary analysis, we have chosen to omit closeness centrality from our study due to its lack of

significant correlation with metro ridership. We assume that closeness centrality tends to reflect the average distance from a node to all other nodes in a network, which can be influenced greatly by the spatial configuration of the transit network. If the network is not designed in a way that proximity (in terms of path lengths) correlates with ridership (e.g., a hub-and-spoke system where ridership is more dependent on transfers than proximity to other stations), the correlation may not be strong. Additionally, to account for the influence of time on metro ridership, we incorporated a variable representing the age of the station. This allows us to investigate the long-term temporal impacts on metro ridership. The rationale behind this is the recognition that the ridership of a metro station is not just dependent on its current attributes, but also on its historical performance and user adoption rate over time. Older stations might have developed established commuter patterns, local dependencies, and more mature surrounding infrastructures, leading to potentially higher ridership. Conversely, newer stations may still be in the phase of growing their user base. By including station age as a variable, we can better account for these temporal dynamics in our analysis.

The independent variables possessing fine-grained spatial configuration were calculated within four different buffer scales: 300 m, 600 m, 800 m, and 1000 m (see Fig. 2). The buffer scales were delineated by drawing straight lines radiating out from the station points, adhering to the radial-based approach typically utilized in DRM and planning standards (Guerra et al., 2012). To address potential issues of multicollinearity, we calculated the variance inflation factor (VIF) for each independent variable. All variables exhibited a VIF value of less than 5, indicating that multicollinearity should not pose a significant problem in the analysis. Table 1 provides a detailed description of all the independent and dependent variables used in this study.

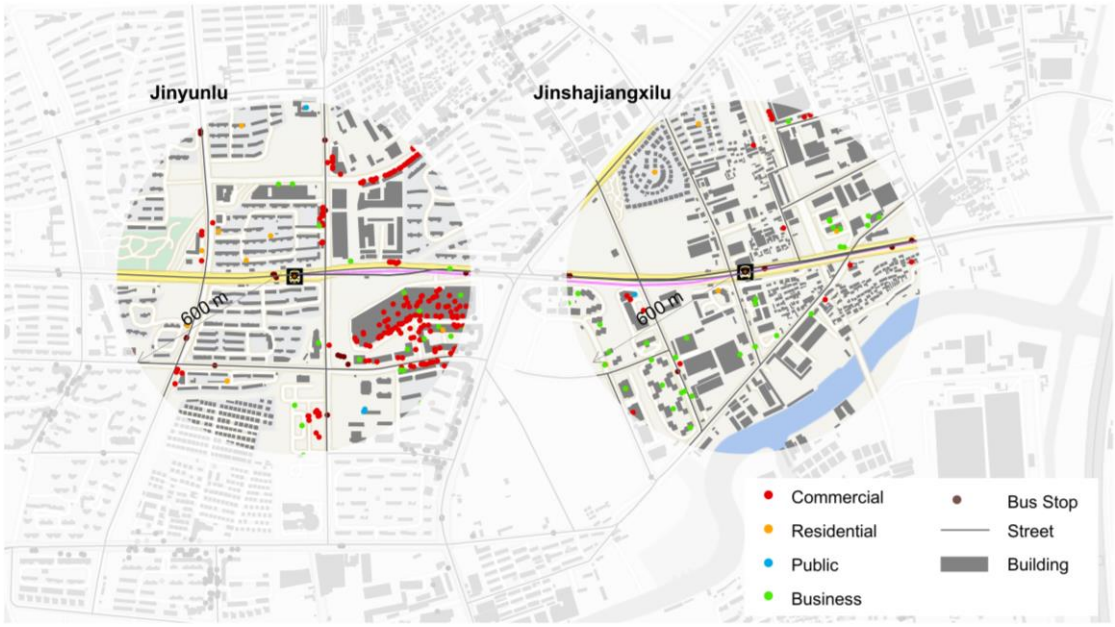

Fig.2. The built environment factors in the station served buffer.

Table 1. Descriptive statistics of variables.

Variable	Description	Mean	St.dev.
Dependent Variable			
Metro ridership	The total daily ridership for the metro station	40390.7	36136.8
Built environment variables			
Population density (300 m)	No. of population per km^2 within a 300 m buffer (thousand people/km^2)	18.09	11.74
Population density (600 m)	No. of population per km^2 within a 600 m buffer (thousand people/km^2)	18.01	11.34
Population density (800 m)	No. of population per km^2 within an 800 m buffer (thousand people/km^2)	17.94	11.05
Population density (1000 m)	No. of population per km^2 within a 1000 m buffer (thousand people/km^2)	17.85	10.77
Employment density (300 m)	No. of employment per km^2 within a 300 m buffer (thousand people/km^2)	8.56	6.63
Employment density (600 m)	No. of employment per km^2 within a 600 m buffer (thousand people/km^2)	8.60	6.63
Employment density (800 m)	No. of employment per km^2 within an 800 m buffer (thousand people/km^2)	8.62	6.59
Employment density (1000 m)	No. of employment per km^2 within a 1000 m buffer (thousand people/km^2)	8.61	6.50
Commercial land use (300 m)	No. of POIs covering macro-level categories of catering, living service, and ping per km^2 within a 300 m buffer (count)	30.2	38.1
Commercial land use (600 m)	No. of POIs covering macro-level categories of catering, living service, and ping per km^2 within a 600 m buffer (count)	88.7	90.6
Commercial land use (800 m)	No. of POIs covering macro-level categories of catering, living service, and ping per km^2 within an 800 m buffer (count)	143.4	137.7
Commercial land use (1000 m)	No. of POIs covering macro-level categories of catering, living service, and ping per km^2 within a 1000 m buffer (count)	251.8	225.6
Business land use (300 m)	No. of companies within a 300 m buffer (count)	11.2	26.9
Business land use (600 m)	No. of companies within a 600 m buffer (count)	41.8	79.8
Business land use (800 m)	No. of companies within an 800 m buffer (count)	73.5	122.1
Business land use (1000 m)	No. of companies within a 1000 m buffer (count)	133.0	190.7

Public land use (300 m)	No. of POIs covering macro-level categories of education and culture within a 300 m buffer (count)	2.2	4.5
Public land use (600 m)	No. of POIs covering first-level categories of education and culture within a 600 m buffer (count)	8.9	11.8
Public land use (800 m)	No. of POIs covering macro-level categories of education and culture within an 800 m buffer (count)	16.8	19.1
Public land use (1000 m)	No. of POIs covering macro-level categories of education and culture within a 1000 m buffer (count)	30.6	33.0
Residential land use (300 m)	No. of POIs covering middle-level categories of residence within a 300 m buffer (count)	6.2	7.8
Residential land use (600 m)	No. of POIs covering middle-level categories of residence within a 600 m buffer (count)	29.1	31.3
Residential land use (800 m)	No. of POIs covering middle-level categories of residence within an 800 m buffer (count)	53.0	55.3
Residential land use (1000 m)	No. of POIs covering middle-level categories of residence within a 1000 m buffer (count)	94.9	96.9
Land use diversity (300 m)	Land use mix (entropy) within a 300 m buffer	0.47	0.25
Land use diversity (600 m)	Land use mix (entropy) within a 600 m buffer	0.62	0.204
Land use diversity (800 m)	Land use mix (entropy) within an 800 m buffer	0.66	0.18
Land use diversity (1000 m)	Land use mix (entropy) within a 1000 m buffer	0.70	0.143
FAR (300 m)	The ratio of floor area over the land area within a 300 m buffer	1.87	1.22
FAR (600 m)	The ratio of floor area over the land area within a 600 m buffer	1.8	1.0
FAR (800 m)	The ratio of floor area over the land area within an 800 m buffer	1.86	1.02
FAR (1000 m)	The ratio of floor area over the land area within a 1000 m buffer	1.27	0.69
Street density (300 m)	The length of street per square kilometre within a 300 m buffer (km/km^2)	8.3	3.8
Street density (600 m)	The length of street per square kilometre within a 600 m buffer (km/km^2)	7.5	3.27

Street density (800 m)	The length of street per square kilometre within an 800 m buffer (km/km^2)	7.7	3.2
Street density (1000 m)	The length of street per square kilometre within a 1000 m buffer (km/km^2)	5.3	2.23
Bus stops (300 m)	No. of bus stops within a 300 m buffer (count)	9.8	6.2
Bus stops (600 m)	No. of bus stops within a 600 m buffer (count)	22.5	13.4
Bus stops (800 m)	No. of bus stops within an 800 m buffer (count)	35.2	19.6
Bus stops (1000 m)	No. of bus stops within a 1000 m buffer (count)	60.1	32.2
Distance to CBD	Straight-line distance to People Square Station (km)	15.25	11.24
Station characteristic variables			
Line	No. of metro lines that the station can ride (count)	1.18	0.48
Entrances	No. of entrances of the metro station (count)	4.5	2.6
Station age	No. of years since the opening of the station (years)	10.4	5.6
Betweenness centrality	The betweenness centrality is expressed as $C_r^B = \sum_{s \neq r \neq t} \frac{\delta_{st}(r)}{\delta_{st}}$, where δ_{st} is the number of shortest paths between node *s* and node *t*, and $\delta_{st}(r)$ is the number of shortest paths from *s* to node *t* that pass through node *r*.	0.045	0.0395

3.2 Analytical approaches

3.2.1 Machine learning approach: eXtreme Gradient Boosting

The eXtreme Gradient Boosting (XGBoost) model was chosen from a range of machine learning algorithms (e.g., random forest and gradient boosting decision tree) to analyze the correlations between the selected explanatory variables and CB ridership. XGBoost, which was introduced by Chen and Guestrin (2016), has become increasingly popular in urban and transportation research (Yang et al., 2021; Liu et al., 2023). It is a robust machine learning algorithm, known for its ability to handle large datasets and model complex, nonlinear relationships. Its design allows for the efficient processing of big data and intricate patterns, surpassing traditional linear models in predictive accuracy. XGBoost also exhibits resilience towards outliers and missing data, and includes regularization parameters to prevent overfitting (Li, 2022).

Similar to the gradient boosting decision tree (GBDT), XGBoost combines decision trees and gradient boosting, but its processing speed is ten times faster than that of GBDT. In addition, XGBoost has been found to outperform alternative methods in terms of predictive accuracy and model fit (Li, 2022). To verify the efficacy of the XGBoost model, we compared the preliminary model results with those obtained with the random forest and gradient-boosting decision tree models. The results confirmed that the XGBoost models outperformed the random forest and GBDT models in terms of predictive accuracy and model fit.

Mathematically, XGBoost uses additive functions to predict the final result as shown in Eq. (2):

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(\mathrm{x}_i), f_k \in \mathcal{F} \quad (2)$$

where K is the number of independent trees used to predict the output $\widehat{y_i}$, and $\mathcal{F}$ represents the space of regression trees. As for the determination of the structure of each

tree, the learning object to be minimized is given as follows:

$$Obj = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (3)$$

where the first term $l(\hat{y}_i, y_i)$ respectively represents the loss function determined by the difference between the predicted value $\hat{y}_i$ and the ground truth y_i, and the later term $\Omega(f_k)$ indicates the regularization parameters as Eq. (4):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega_i\|^2 \quad (4)$$

Here γ penalizes the number of leave nodes of the tree *T* to avoid over-fitting, ω represents the score of the *i*-th leaf, and λ is a regularization parameter of each leaf to keep the variances of leaf weights at a low level. For a fixed structure *q*(x), we can compute the optimal solution ω_j^* of leaf *j* by Eq. (5), and simplify the objective function by Eq. (6).

$$\omega_j^* = -\frac{G_j}{H_i+\lambda}\gamma T \quad (5)$$

$$\widetilde{Obj}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T \quad (6)$$

3.2.2 Interpretation of machine learning: SHapley Addictive exPlanations

To interpret the XGBoost models in this study, we applied the Shapley Additive exPlanations (SHAP) technique, a sophisticated machine learning interpretation method proposed by Lundberg and Lee (2017). SHAP quantifies the contribution of each input variable within the XGBoost model. In contrast to global interpretation methods such as Partial Dependence Plots (PDP), SHAP provides a local interpretation technique that yields explanations at an individual instance level (Li, 2022). This capability facilitates the unveiling of local heterogeneities that may be obscured by issues of data sparsity or outlier effects, while SHAP merges global interpretations with local interpretations by leveraging the additivity attribute of Shapley values (Lundberg and Lee, 2017). This approach becomes particularly advantageous when the research

goal is to probe more deeply into the complex relationships between dependent and independent variables.

SHAP is a consistent feature attribution method using a linear formula combined with the sum of variable effects and an intercept as the approximation of the prediction as Eq. (7):

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \quad (7)$$

SHAP enables us to interpret the resultant models globally and locally using the SHAP feature importance (Li, 2022). Here $f(x)$ represents the original model for a specific input x, and $g(x')$ represents the explanation model matching $f(x)$. Feature attributions are indicated by the Shapley values ϕ_i, defined as a weighted average of all possible differences:

$$\phi_i = \sum_{S \subseteq F \backslash \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right] \quad (8)$$

Take the importance value to feature *i*, which is allocated based on its marginal contribution. Model $f_{S \cup \{i\}}$ is trained with the feature presentation, while the model f_S is trained with the feature withheld. Since the effect of withholding a feature depends on other features in the model, the preceding differences are computed for all possible subsets $S \subseteq F \backslash \{i\}$, as shown in Eq. (8).

The Shapley value, symbolized as φ_i for a specific feature *i* can hold positive, negative, or zero. These indicate whether the feature enhances, reduces, or leaves the prediction unchanged relative to the average prediction. To illustrate, a positive SHAP value ($\varphi_i > 0$) implies that the presence of feature *i* elevates the prediction for the specific instance against the average prediction, suggesting a positive contribution towards the anticipated outcome.

To obtain the optimal parameter settings and avoid the overfitting problem, a five-fold cross-validation procedure was used to train the XGBoost models. This study sets

different numbers of trees (5000, 10000, 15000, 20000, 25000), with depth (4, 5, 6, 8, 9) and the shrinkage parameter (0.001, 0.05, 0.01, 0.1). After iterations, the best performance with the lowest mean absolute error (MAE), root means square error (RMSE), and R^2 were obtained when the number of trees, depth, and shrinkage were set as parameters in Table 2.

Table 2. Hyper-parameters for the XGBoost models.

	300 m	600 m	800 m	1000 m
n_estimators	15000	15000	15000	15000
learning_rate	0.001	0.001	0.001	0.001
max_depth	9	4	9	5
Training set				
training_sample	307 (90%)	307 (90%)	307 (90%)	307 (90%)
R^2	0.99	1.00	0.99	0.99
MAE	4686.0	3541.5	3839.0	4184.5
RMSE	7475.7	5761.8	6504.4	7250.5
Testing set				
testing_sample	34 (10%)	34 (10%)	34 (10%)	34 (10%)
R^2	0.78	0.90	0.85	0.80
MAE	10638.4	8171.7	9084.5	10134.3
RMSE	16172.7	10805.4	12711.0	14975.7

4. Results

4.1 Performance of the XGBoost models

We compared the R^2, MAE, and RMSE between conventional OLS models and the XGBoost model within four different buffers (see Table 3). Firstly, across all buffer sizes, the XGBoost models outperformed the OLS regression models. The XGBoost models improved the R^2 from 0.10 to 0.19, decreased the MAE from 2355.1 to 2723.5, and reduced the RMSE from 319.9 to 2472.2. These results align with our expectations. By easing the linearity assumption, XGBoost can model intricate, non-linear

relationships between the dependent and independent variables. In addition, XGBoost is more robust to outliers and has an innate ability to handle missing values without the need for explicit imputation or removal, thereby enhancing the model's predictive power.

Second, when compared to the XGBoost model, the variance in performance of the OLS models across the four buffer sizes was relatively minor, with the largest R^2 interval being approximately 0.05. This finding aligns with Guerra et al. (2012), which posited that variations in catchment areas have a limited effect on a model's predictive power. Nevertheless, the XGBoost model demonstrated a different trend, where different buffer sizes indeed influence the model's predictive power, with the largest R^2 interval reaching 0.12. The widening difference across various catchment areas may be attributed to the mechanisms intrinsic to machine learning models. The XGBoost models are adept at detecting subtle nuances relationship between the built environment and metro station, resulting in increased disparities among models of varying buffer sizes. This may elucidate the findings of Jun et al. (2015), who discovered that the impact of the built environment on metro ridership significantly varies by PCA by using the MGWR model in conjunction with a finer-grained built environment dataset.

Third, both the OLS and XGBoost models across the four buffer models accentuated that the model incorporating a 600 m buffer size boasts superior predictive power compared to models that use other buffer sizes. This observation makes a compelling case for selecting a 600 m buffer as the recommended choice for predicting station-area metro ridership and station-area surrounding developments. The results align with Jun et al. (2015), which is understandable considering that a 600 m distance typically equates to a walkable distance within 10 minutes from a metro station, and a 600 m span roughly corresponds to the size of one or two street blocks in Shanghai,

which is a typical size that suits a property development. Furthermore, this superiority of the 600 m spatial scale finds corroborative support in the work of Jiang et al. (2020), which revealed pronounced positive external effects on surrounding property prices within a distance of approximately 600 m in Shanghai. This suggests that the enhanced accessibility provided by the proximity to a metro station within this 600 m zone is factored into the housing prices, reflecting its perceived value to residents and investors.

Table 3. Performance of the XGBoost model and comparison.

	300 m		**600 m**		**800 m**		**1000 m**	
	OLS	XGBoost	OLS	XGBoost	OLS	XGBoost	OLS	XGBoost
R^2	0.68	0.78	0.71	0.90	0.68	0.85	0.66	0.80
MAE	13190.3	10638.4	10671.9	8171.7	11439.6	9084.5	12857.8	10134.3
RMSE	16492.6	16172.7	13277.6	10805.4	14079.7	12711.0	16455.4	14975.7

4.2 The relative importance of independent variables

The study further compared the relative importance and rankings of independent variables among four different buffer size models, the results show that buffer size indeed has a significant effect on the relationship between the built environment and metro ridership.

The collective contributions of the built environment and station characteristic factors vary across different buffer sizes (see Table 4). For example, the collective contribution rate of built environment factors stands at 53.1%, 56.9%, 50.9%, and 40.4% for the 300 m, 600 m, 800 m, and 1000 m buffer models, respectively. In contrast, station characteristics demonstrate a continuously increasing contribution rate: 46.9%, 43.1%, 49.1%, and 59.6% for the 300 m, 600 m, 800 m, and 1000 m buffer models, respectively. The results provide two insights. First, station characteristic variables, frequently overlooked in the literature, indeed play a pivotal role in influencing metro ridership, especially for the number of lines and entrances as well as station age, which

construct a basic and direct way to capture metro ridership from surrounding areas. In contrast, the effect of single built environment variables on metro ridership is margin, but "the combined effect of several built environmental variables on travel could be quite large" (Ewing and Cervero, 2010: 275). Second, the results suggest that the influences of the built environment are highly interactive with proximity to the metro station, with the maximum influence exerted within a 600 m catchment area, which is interestingly consistent with the findings in the predictive performance of the model across different buffers. This not only further underscores the reason for the superior model performance when incorporating a 600 m buffer but also advocates for urban designers and planners to prioritize this 600 m catchment area surrounding the station to maximize the effectiveness of built environment-related interventions.

Table 4. The XGBoost model results.

	300 m		**600 m**		**800 m**		**1000 m**	
	Rank	RI	Rank	RI	Rank	RI	Rank	RI
Built environment variables								
Population density	14	2.5%	13	3.0%	15	2.4%	12	2.7%
Employment density	6	6.4%	6	4.7%	8	3.9%	7	5.0%
Commercial land use	3	14.7%	2	19.6%	3	15.5%	6	5.2%
Business land use	12	2.9%	7	3.9%	5	5.5%	4	5.6%
Public land use	15	2.1%	15	2.8%	14	2.5%	13	2.1%
Residential land use	13	2.9%	10	3.4%	10	2.8%	12	3.4%
Land use diversity	8	3.7%	8	3.6%	13	2.6%	14	2.1%
FAR	5	6.7%	5	5.7%	6	4.6%	10	3.1%
Distance to the city centre	11	3.4%	12	3.2%	9	3.9%	8	3.5%
Street density	9	3.7%	11	3.4%	7	4.4%	5	5.6%
Bus stops	7	4.1%	9	3.6%	11	2.8%	15	2.0%

Sum		**53.1%**		**56.9%**		**50.9%**		**40.4%**
Station characteristic variables								
Lines	2	15.2%	3	11.1%	1	22.4%	1	29.8%
Entrances	1	21.1%	1	21.6%	2	17.1%	2	19.5%
Station age	4	7.0%	4	7.4%	4	7.0%	3	7.4%
Betweenness centrality	10	3.6%	14	2.9%	12	2.6%	11	2.9%
Sum		**46.9%**		**43.1%**		**49.1%**		**59.6%**

*RI: Related Importance

The ranking of individual built environment factors display heterogeneous and inconsistent results across different buffer sizes (see Fig. 3). Although the most impactful features generally maintain a consistent rank across various buffer size models, the ranks of less influential variables fluctuate significantly. For example, the number of entrances appears to hold significant importance, ranking first in the 300 m and 600 m models, and second in the 800 m and 1000 m models. Conversely, business land use demonstrates a progressive increase in importance, ranking 12th in the 300 m model, 7th in the 600 m model, 5th in the 800 m model, and 4th in the 1000 m model. A plausible explanation could be that highly important variables account for the majority of variance in metro ridership, thus exerting a larger influence on prediction outcomes. Less influential variables, on the other hand, may owe their reduced importance to weak intrinsic correlations, multicollinearity, or inherent noise, rendering them less critical for accurate predictions and more prone to substitution by other intervening factors (Hu et al., 2023).

Specifically, the findings indicate that commercial land use, FAR, and land use diversity, which are often recommended for enhancing TOD, demonstrate similar effects on metro ridership within a 600 m radius. However, their influence on metro ridership decreases dramatically as the buffer size exceeds 600 m. This suggests the importance of proximity to the metro stations in leveraging these TOD-favorable factors

for TOD, which follows the principles of agglomeration economies. Conversely, the significance of business land use and street density escalates as the buffer size broadens. As Kwan (2012) propounds, different land use variables have their different "truly causally relevant" area where the environment influences travel behaviour. Our findings affirm this proposition and further uncover that the benefits of proximity to the metro station for business land use may be less potent or competitive compared to other land uses (e.g., commercial land use). Meanwhile, the positive implications of spatial connectivity, anchored on the street network, are more evident at larger scales.

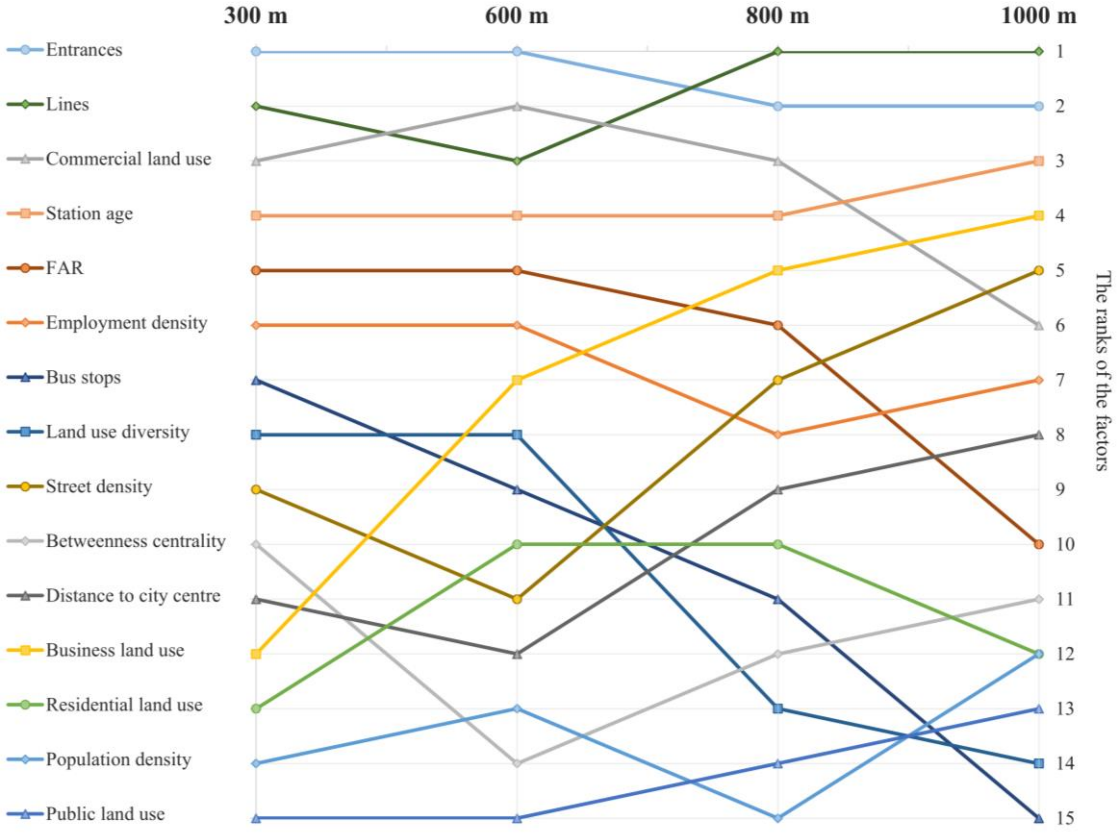

Fig. 3. Buffer effect on the results of the ranks of independent variables.

4.3 Nonlinear effect of independent variables

The influence of different buffer sizes on the SHAP value plots for all variables was also examined to ascertain whether these sizes significantly impact the nonlinear relationship between the built environment and metro ridership.

Fig. 4 illustrates the SHAP plots for population density and employment density across four buffer scales. For population density, the SHAP plots show an inverse U-shaped relationship between population density and metro ridership from 0 to 30 thousand people/km^2. Once it exceeds that threshold, the relationship once again turns positive, albeit somewhat dispersed, and this pattern is consistent across all four buffer size models. Given the minor variations in population across the different buffer zones, this outcome appears reasonable. For employment density, the SHAP plots reveal a less consistent nonlinear relationship across the four buffer sizes. In the 300 m and 600 m models, these plots depict an inverse V-shaped relationship between employment density and metro ridership. However, the adverse influence of employment density on metro ridership weakens considerably in the 800 m and 1000 m models. This discrepancy can potentially be ascribed to the variance in employment density across different buffers (see Table 1), which becomes less pronounced as the spatial scale of analysis expands.

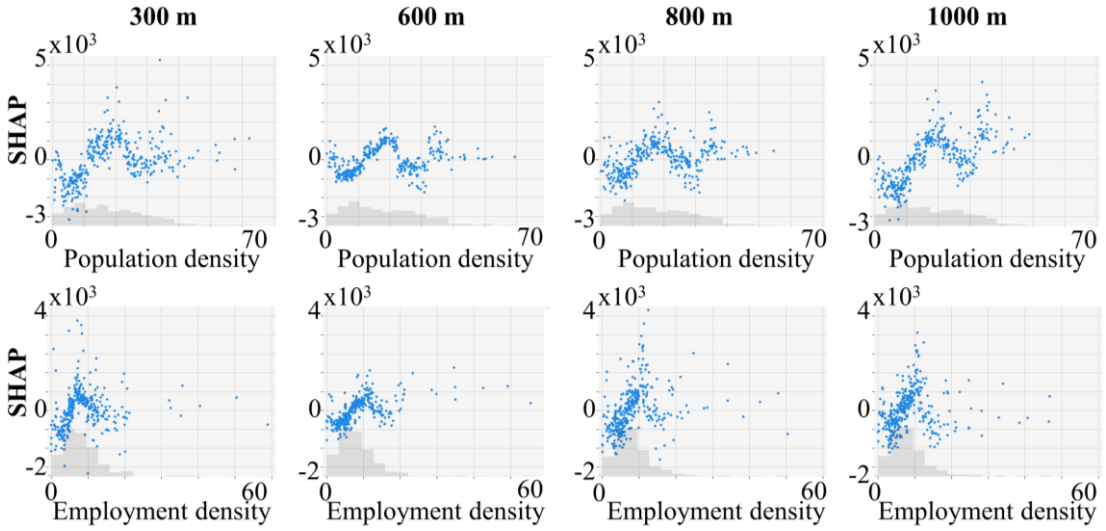

Fig. 4. The SHAP value plots of population and employment density.

Fig. 5 illustrates the SHAP plots for various land uses across four buffer scales, revealing that buffer size substantially impacts the nonlinear relationship between land uses and metro ridership. Despite being measured in counts instead of density, land uses

exhibit significant fluctuations among the four distinct buffer sizes. Take commercial land use as an example. In the 300 m, 600 m, and 800 m models, the curve is relatively steep, but once the value exceeds 100 in the 1000 m model, the curve markedly flattens. This trend suggests that the positive influence of commercial land use on metro ridership diminishes as the distance from the metro station increases. This observation is consistent with the findings detailed in Table 2 and Fig. 3. Similar variations can also be observed for other land use types. These findings further attest to the existence of the "uncertain geographic context problem" within the station catchment area (Kwan, 2012), even this area is on a comparatively smaller spatial scale.

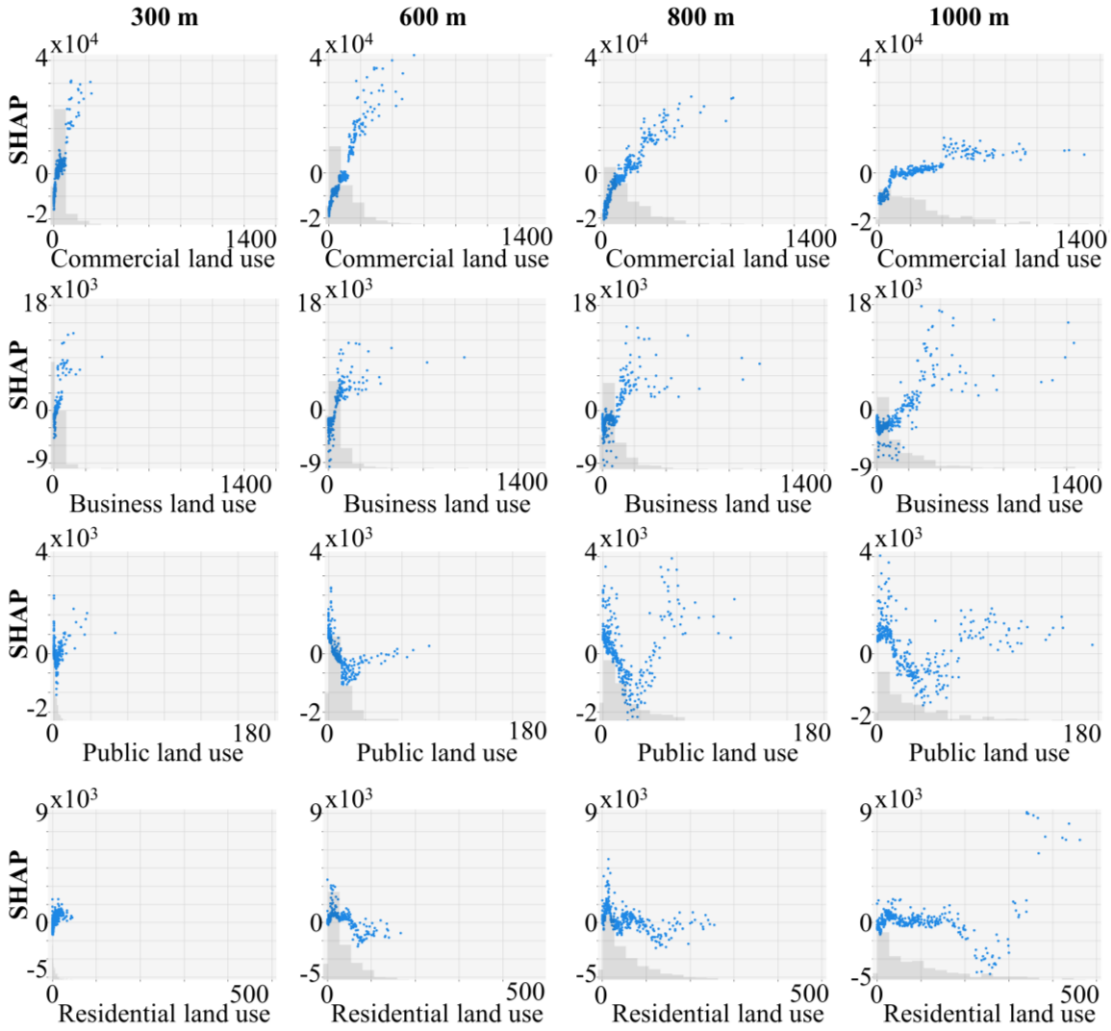

Fig. 5. The SHAP value plots of land use variables.

Fig. 6 illustrates the SHAP plots for diversity, design, and destination accessibility

variables. In contrast to land use variables, metrics such as land use diversity, FAR, and street density are quantified in densities terms rather than counts. Results reveal that the buffer size exerts varying degrees of influence on the relationship between the built environment and metro station ridership. For example, the effect of street density on metro ridership within the 300 m and 600 m buffer models is relatively consistent, as evidenced by flatter curves. However, within the 600 m and 1000 m models, the curves for the street network become fluctuating and ascend upwards. A plausible explanation could be that smaller buffer sizes fail to encapsulate comprehensive information on the street network required to act as a proxy for spatial connectivity. This discrepancy could account for the contrasting findings in the literature, with some studies associating a positive relationship between the street network and metro stations (Shao et al., 2020), while others did not observe such a correlation (An et al., 2019). The SHAP plots for bus stops and distance to CBD also confirm that the buffer size may affect their nuanced correlations with metro ridership.

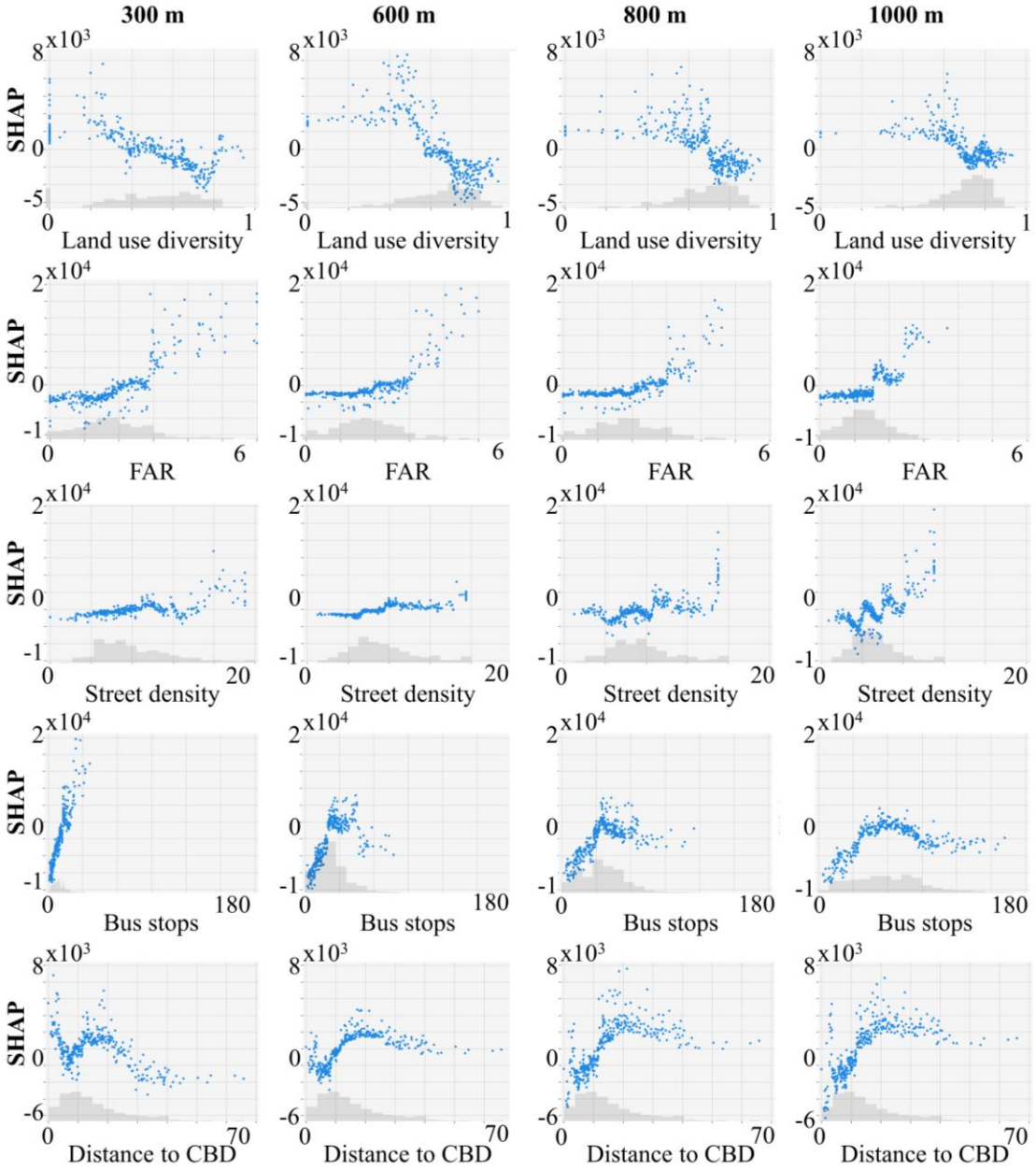

Fig. 6. The SHAP value plots of diversity, design, and destination accessibility variables.

Fig. 7 illustrates the SHAP plots for station characteristic variables. Contrary to built environment variables, these plots indicate that the buffer size does not significantly impact the correlations with metro ridership. This observation is justified for two possible reasons. First, the values of station characteristic variables remain consistent across different buffer sizes. This uniformity prevents the introduction of inconsistent values of independent variables into the analysis, thereby eliminating potential skewness in the results. Second, from a theoretical perspective, the influence

mechanism of station characteristic variables is expected to remain independent of the analytical spatial scale. This is because these variables inherently pertain to the characteristics of the station itself, and are not influenced by the spatial scale of the surrounding environment. Consequently, the scale of analysis does not distort their correlations with metro ridership.

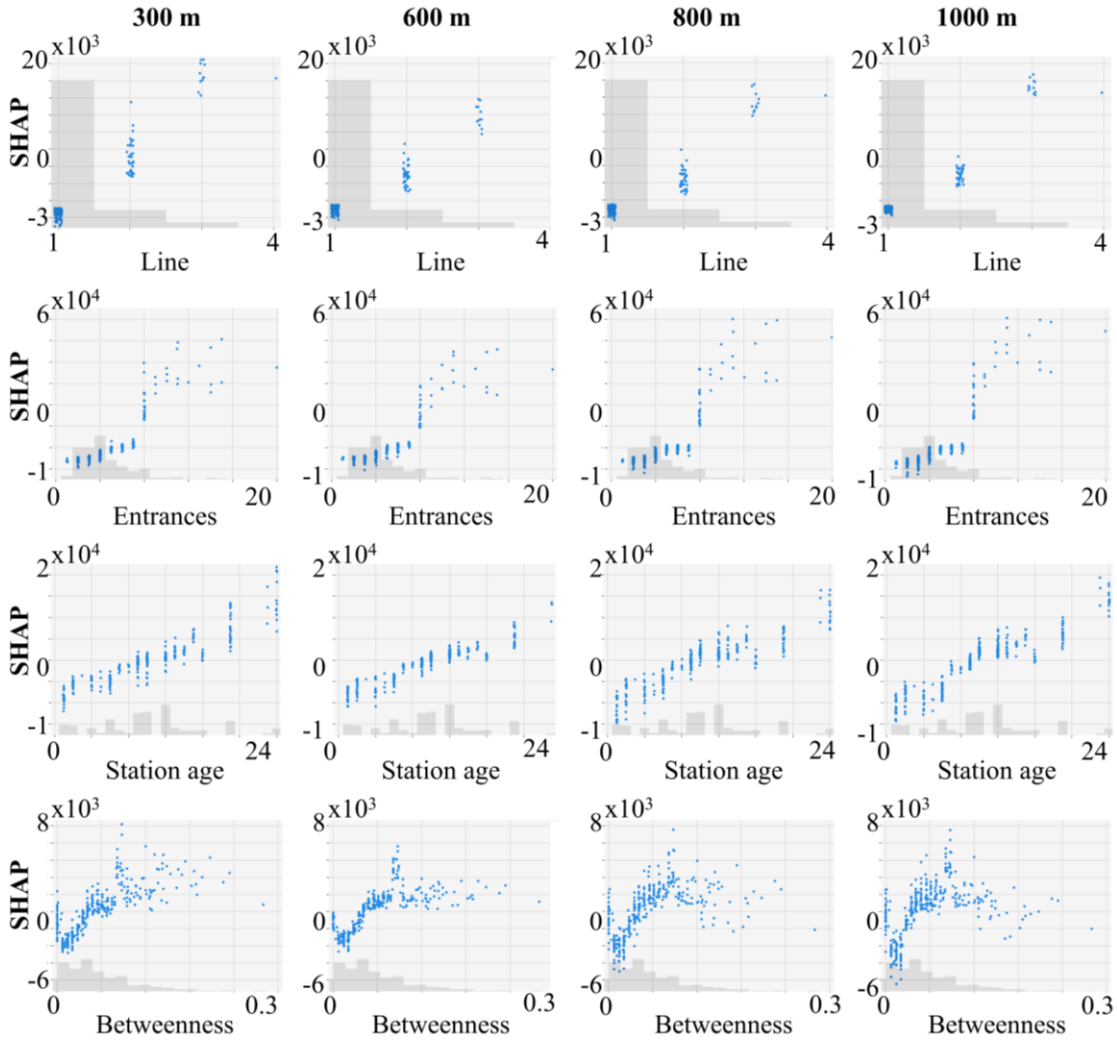

Fig. 7. The SHAP value plots of station characteristics variables.

5. Conclusions

Using smart card data from Shanghai, this study applied the XGBoost and SHAP methods to conduct a sensitivity analysis defining built environment factors within four radial buffer sizes: 300 m, 600 m, 800 m, and 1000 m on associations with metro ridership. The findings contribute to the literature threefold. First, it compares the model

predictive performance of the OLS models and XGBoost models across four different buffers, results show that different buffer sizes have little influence on the OLS model, but significant on the XGBoost model, while the use of a 600 m buffer size around the metro station demonstrates the best model fit and variation explanation for both models compared to other buffers. The results align with Guerra et al. (2012) findings, which indicate that the influence of varying catchment areas on a model's predictive power is minimal. However, we argue that this conclusion might be limited to the OLS model. By incorporating spatial autocorrelation and nonlinear effects into the model, it appears that changing buffer sizes have a negligible impact on the relationship between built environment variables and metro ridership. Besides, results from the OLS and XGBoost models support the use of a 600 m buffer size for forecasting metro ridership or for proposing station-area land use interventions. Despite the variability of these thresholds in case studies across the globe, this study introduces an evidence-based approach for setting the suitable analytical scale in related analyses. This is of particular importance given the prevalent tendency in current research and practice to determine the analytical buffer size somewhat indiscriminately.

Second, the variability in relative importance and rankings of built environment variables and station characteristic variables across the four buffer models demonstrates that different variables possess their own "true causally relevant" areas where the environment has a significant impact on metro ridership (Kwan, 2012). For example, results imply that commercial land use, FAR, and land use diversity, three TOD-favourable factors, exert a stronger influence on metro ridership the closer they are to the metro station. In contrast, the significance of business land use and street density grows as the buffer size expands. These findings could potentially explain the inconsistent empirical results reported in existing literature, while the sensitivity

analyses offer new insight into the ways and degree to which built environment variables influence metro ridership across different geographic delineations of contextual units.

Third, this study further examined how the choice of buffer influences nonlinear correlations between variables and metro ridership. The findings reveal a notable heterogeneity among the selected independent variables. The nonlinear effects of built environment variables, most notably employment density, commercial land use, FAR, street density, and bus stops, show significant sensitivity to the choice of analytical buffers. This sensitivity could potentially be attributed to two factors. One is the different "true causally relevant" areas wherein each variable exerts its influence on metro ridership. The other is the variation in the spatial distribution of variables within different buffers. These results corroborate the results of different rankings of independent variables among the different buffers. In sum, the findings underscore the need for researchers to be meticulous when defining the analytical spatial area in their studies. This caution is particularly essential when employing more sophisticated machine learning methodologies.

Some elements need to be further investigated. First, the buffers in this study were calculated by a radical buffer method. The advantage of this method is that it is easy to calculate and understand, but it may not be the best representation of pedestrian catchment areas in reality. Other methods such as line-based buffer or time-based buffer might gain more interesting results. Second, the study was undertaken in Shanghai, a megacity with a complex and distinctive urban context and travel behaviour, therefore, the results of this study may be specific to Shanghai, and future research should explore additional cases and compare variations among cities to draw more generalized conclusions.

Acknowledgements

This research was sponsored by the National Natural Science Foundation of China (No.71734004) and the International Exchange Program for Graduate Students, Tongji University (No. 2023020023). The authors also like to thank the reviewers for their comments and the editors for their editorial help and patience.

References

An, D., Tong, X., Liu, K., Chan, E. H., 2019. Understanding the impact of built environment on metro ridership using open source in Shanghai. *Cities*, 93, 177-187.

Andersson, D.E., Shyr, O.F., Yang, J., 2021. Neighbourhood effects on station-level transit use: Evidence from the Taipei metro. *Journal of Transport Geography,* 94, 103127.

Batty, M., Besussi, E., Chin, N., 2003. *Traffic, urban growth and suburban sprawl*. CASA Working Papers. London: Centre for Advanced Spatial Analysis (UCL).

Bertolini, L., Curtis, C., Renne, J., 2012. Station area projects in Europe and beyond: Towards transit oriented development? *Built Environment*, 38(1), 31-50.

Calthorpe, P., 1993. *The next American metropolis: Ecology, community, and the American dream*. Princeton Architectural Press.

Cardozo, O.D., García-Palomares, J.C., Gutiérrez, J., 2012. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography,* 34, 548-558.

Cervero, R., 1998. *The transit metropolis: a global inquiry*. Island Press.

Cervero, R., 2006. Alternative approaches to modeling the travel-demand impacts of smart growth. *Journal of the American Planning Association*, 72(3), 285-295.

Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: Density, diversity, and

design. *Transportation Research Part D: Transport and Environment*, 2(3), 199-219.

Cervero, R., Ferrell, C., Murphy, S., 2002. Transit-oriented development and joint development in the United States: A literature review. *TCRP research results digest*, 52.

Chen, L., Lu, Y., Liu, Y., Yang, L., Peng, M., Liu, Y., 2022. Association between built environment characteristics and metro usage at station level with a big data approach. *Travel Behaviour and Society*, 28, 38-49.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, 785–794.

Clark, A., Scott, D., 2014. Understanding the impact of the modifiable areal unit problem on the relationship between active travel and the built environment. *Urban Studies*, 51(2), 284-299.

Deng, Y., Zhao, P., 2022. The impact of new metro on travel behavior: Panel analysis using mobile phone data. *Transportation Research Part A: Policy and Practice*, 162, 46-57.

Ding, C., Cao, X., Liu, C., 2019. How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds. *Journal of Transport Geography*, 77, 70-78.

Ewing, R., Cervero, R., 2010. Travel and the built environment: A meta-analysis. *Journal of the American Planning Association*, 76(3), 265-294.

Freeman, L. C., 1978. Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215-239.

Guerra, E., Cervero, R., Tischler, D., 2012. Half-mile circle: Does it best represent

transit station catchments? *Transportation Research Record*, 2276(1), 101-109.

Hu, S., Xiong, C., Chen, P., Schonfeld, P., 2023. Examining nonlinearity in population inflow estimation using big data: An empirical comparison of explainable machine learning models. *Transportation Research Part A: Policy and Practice*, 174, 103743.

James, P., Berrigan, D., Hart, J.E., Hipp, J.A., Hoehner, C.M., Kerr, J., Major, J.M., Oka, M. and Laden, F., 2014. Effects of buffer size and shape on associations between the built environment and energy balance. *Health & Place*, 27, 162-170.

Jiang, Y., Gu, P., Cao, Z., Chen, Y., 2020. Impact of transit-oriented development on residential property values around urban rail stations. *Transportation Research Record*, 2674(4), 362-372.

Jun, M. J., Choi, K., Jeong, J. E., Kwon, K. H., Kim, H. J., 2015. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *Journal of Transport Geography*, 48, 30-40.

Kuby, M., Barranda, A., Upchurch, C., 2004., Factors influencing light-rail station boardings in the United States. *Transportation Research Part A: Policy and Practice*, 38(3), 223-247.

Kwan, M.P., 2012. The uncertain geographic context problem. *Annals of the Association of American Geographers*, 102(5), 958-968.

Li, S., Lyu, D., Huang, G., Zhang, X., Gao, F., Chen, Y., Liu, X., 2020. Spatially varying impacts of built environment factors on rail transit ridership at station level: a case study in Guangzhou, China. *Journal of Transport Geography*, 82, 102631.

Li, S., Zhao, P., 2017. Exploring car ownership and car use in neighborhoods near metro stations in Beijing: Does the neighborhood built environment matter? *Transportation Research Part D: Transport and Environment*, 56, 1-17.

Li, Z., 2022. Extracting spatial effects from machine learning model using local

interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, 96, 101845.

Liu, X., Chen, X., Potoglou, D., Tian, M., Fu, Y., 2023. Travel impedance, the built environment, and customized-bus ridership: a stop-to-stop level analysis. Transp. Res. Part D: Transp. Environ. 122, 103889.

Loo, B. P., Chen, C., Chan, E.T., 2010. Rail-based transit-oriented development: lessons from New York City and Hong Kong. *Landscape and Urban Planning*, 97(3), 202-212.

Lundberg, S. M., Lee, S. I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Nasri, A., Zhang, L., 2014. The analysis of transit-oriented development (TOD) in Washington, DC and Baltimore metropolitan areas. *Transport Policy*, 32, 172-179.

O'Sullivan, S., Morrall, J., 1996. Walking distances to and from light-rail transit stations. *Transportation Research Record,* 1538(1), 19-26.

Pan, H., Li, J., Shen, Q., Shi, C., 2017. What determines rail transit passenger volume? Implications for transit oriented development planning. *Transportation Research Part D: Transport and Environment*, 57, 52-63.

Papa, E., Bertolini, L., 2015. Accessibility and transit-oriented development in European metropolitan areas. *Journal of Transport Geography*, 47, 70-83.

Shao, Q., Zhang, W., Cao, X., Yang, J., Yin, J., 2020. Threshold and moderating effects of land use on metro ridership in Shenzhen: Implications for TOD planning. *Journal of Transport Geography*, 89, 102878.

Sohn, K., Shim, H., 2010. Factors generating boardings at metro stations in the Seoul metropolitan area. *Cities*, 27(5), 358-368.

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y. and Huang, H., 2020. Statistical

and machine-learning methods for clearance time prediction of road incidents: A methodology review. *Analytic Methods in Accident Research*, 27, 100123.

van Wee, B., Handy, S., 2016. Key research themes on urban space, scale, and sustainable urban mobility. *International Journal of Sustainable Transportation*, 10(1), 18-24.

Xu, W., Guthrie, A., Fan, Y., Li, Y., 2017. Transit-oriented development in China: Literature review and evaluation of TOD potential across 50 Chinese cities. *Journal of Transport and Land Use*, 10(1), 743-762.

Yang, C., Chen, M., Yuan, Q., 2021. The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis. *Accident Analysis & Prevention*, 158, 106153.

Yang, L., Yu, B., Liang, Y., Lu, Y., Li, W., 2023. Time-varying and non-linear associations between metro ridership and the built environment. *Tunnelling and Underground Space Technology*, 132, 104931.

Zhao, J., Deng, W., Song, Y., Zhu, Y., 2013. What influences Metro station ridership in China? Insights from Nanjing. *Cities*, 35, 114-124.