Travel impedance, the built environment, and customized-bus ridership: A stop-by-stop level analysis

Xiang Liu[a,*], Xiaohong Chen[a,b], Dimitris Potoglou[c], Mingshu Tian[b], Ying Fu[a]

[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] *School of Geography and Planning, Cardiff University, Cardiff, Wales, UK*

[*] Corresponding author. Address: 4800 Cao'an Road, Shanghai 201804, China. Email: xliu02@tongji.edu.cn

Travel impedance, the built environment, and customized-bus ridership: A stop-to-stop level analysis

ABSTRACT

Customized buses (CBs) are a complementary but essential component of the public transit system and have gained increasing popularity. However, knowledge remains limited regarding how CB service performs and what factors significantly influence the performance, particularly at a stop-to-stop level. By utilizing over two years of CB subscription data from Shanghai, we applied both the multiplicative model and the XGBoost model to identify key determinants of CB ridership and to examine nonlinear associations at a stop-to-stop level. The results suggest that: (1) travel impedance and the built environment, especially distance to the nearest metro station and distance to the city center, are significant predictors of CB ridership; (2) the effects of the built environment are nonlinear and vary by side of stop and time of day. The findings assist CB providers and policymakers in identifying appropriate market niches and areas with the greatest potential for allocating stops and designing routes.

Keywords: customized bus; demand-responsive transit; travel impedance; built environment; machine learning; stop-to-stop;

1. Introduction

Public transport carries substantial responsibility and holds immense potential to reduce greenhouse gas emissions (Litman, 2015). However, simply having public transportation available is not enough. It needs to be reliable, convenient, and affordable to be a viable alternative to private cars (Banister, 2005). Customized buses (CBs) represent an important innovation in public transport (PT), offering a flexible and user-centric solution that can complement traditional fixed-route bus services, the services help make public transport a more attractive option for a wider range of people (Liu and Ceder, 2015). By offering a reliable and convenient public transport option, CB services may encourage people to switch from private vehicles to public transport. This could reduce the total number of vehicles on the road, leading to less traffic congestion and lower emissions per passenger.

In recent years, CBs have been increasingly promoted in cities all over the world, such as Shanghai, Manchester, and San Francisco, since their first launch in Qingdao in August 2013 (Liu and Ceder, 2015). CBs are recognized as an ideal PT service for the foreseeable future, holding the promise of reducing traffic pollution, easing traffic pressure, enhancing the passenger experience, and optimizing urban transit systems (Liu and Ceder, 2015; Wang et al., 2019; Wang et al., 2020). In contrast to traditional PT services with fixed routes, stop layouts, fleet and crew scheduling (Petit and Ouyang, 2022), CB services offer a more reliable, cost-effective (e.g. when compared to taxis), and more convenient (with fewer or no stops) transit service by aggregating travelers with similar travel address and travel time through app-based platforms. Moreover, CB providers can deploy their service capacity in a more efficient and demand-driven manner based on actual travel requests collected from online surveys (Huang et al., 2020).

However, given the prevalence and advantages of CB services, little is known about how CB services perform across time and space within the urban transit system. Several recently launched CB schemes have experienced a dramatic drop in the number of CB subscriptions and user registrations (Wang et al., 2019), casting doubt on the long-term future of CBs. The sustainable development of CBs urgently requires more thorough and case-specific studies. While prior studies primarily focus on investigating the factors driving the demand for PT and demand-responsive transport (DRT) services, the factors associated with the demand for CB services have not been extensively studied in the literature. Blindly pursuing the ongoing development of CBs without gaining a deeper comprehension of their travel patterns and the factors influencing CB demand may result in a regrettable misuse of public or private resources, especially since more and more local authorities are now considering CBs as a plausible solution to reverse the recent declines in traditional bus ridership (Wang et al., 2014).

The existing literature on CBs has mainly focused on elaborating the conceptual framework of CB (Liu and Ceder, 2015; Liu et al., 2016), route mining (Li et al., 2019; Lyu et al., 2016), and network design or operation optimization algorithms (Huang et al., 2020; Li et al., 2018; Lyu et al., 2016; Lyu et al., 2019; Petit and Ouyang, 2022). There is a limited number of studies that have conducted empirical research on the spatiotemporal patterns of CB use and the determinants influencing their ridership.

(Huo et al., 2021; Yan et al., 2020). The literature investigating the factors affecting the demand for CBs has generally been based on questionnaire surveys (Li et al., 2021). However, data obtained from revealed or stated preference surveys can only reflect possible choices made by respondents based on predefined attributes and not respondents' actual choice behaviors. Additionally, it is difficult for surveys to capture respondents' choice behaviors over a long period due to high tracking costs and privacy concerns (Wang et al., 2020). Other studies have tried to overcome these limitations by directly utilizing recorded data, such as PT smart card data (Guo et al., 2019; Li et al., 2018) or taxi trajectory data (Lyu et al., 2016) in order to investigate CB demand. However, the findings are still controversial as CBs have unique characteristics of service in terms of passenger travel behavior, route design, and service operation (Liu and Ceder, 2015).

Moreover, numerous studies on CBs or their predecessor, demand-responsive transit (DRT), primarily concentrate on the stop level. There is a scarcity of research examining the effects of the built environment and travel impedance variables on ridership at the stop-to-stop level and at different times of the day. Travel demand is generally characterized by consistent densities over time and/or space (Daganzo, 2010; Petit and Ouyang, 2022; Wu et al., 2020). For instance, travel demand is typically concentrated during peak hours and near high-interest centers (e.g. central business districts), while it is sparse during off-peak hours and in less populated suburban areas. Studies on the correlations between the built environment and metro ridership affirm that the built environment's effect varies based on origin-destination factors and the time of day (Ewing and Cervero, 2010). However, to our knowledge, no studies have yet investigated the factors influencing CB ridership at the stop-to-stop level and across different times of the day, despite their significance for offering sustainable CB services. The rapid evolution of CBs in recent years should motivate academia to address this research gap.

Regarding methodology, the conventional approach for exploring the critical determinants of transit demand is based on statistical models, with the assumption that the relationships between transit demand and explanatory variables are linear or log-linear (Chiou et al., 2015; Deepa et al., 2022; Ewing and Cervero, 2010). Recent studies have shown that most explanatory variables (e.g. the built environment) might have nonlinear effects on transit demand (Gan et al., 2020). For example, increasing development intensity and density in low-density areas is an effective strategy to reduce car dependence, but the critical question is what level of density is appropriate to reduce car dependence effectively. Is there any threshold or interval at which an effect is exerted by increasing density? Therefore, it is more important and meaningful to specify the nonlinear and threshold correlations between explanatory variables and transit demand, as they can better identify the most effective influences of these variables for improving transit usage (Ding et al., 2019; Shao et al., 2020). In this context, machine learning approaches offer a substantial advancement over conventional statistical models, enabling a more adaptable model structure capable of capturing irregular, non-linear relationships among variables (Yan et al., 2020). However, to our knowledge, only one study published in mainstream academic journals has leveraged machine-learning approaches to investigate the potential factors influencing the performance of CBs (Wang et al., 2023a). More efforts are required to draw a more generalized and robust conclusion.

In light of this context, the present study aims to enrich the existing literature on CB ridership by examining the spatiotemporal patterns of CBs and identifying key factors that influence CB ridership at the stop-to-stop level from a practical performance standpoint. This research seeks to address three key questions: (1) What are the spatiotemporal patterns of CBs? (2) Which factors have a significant impact on CB ridership at the stop-to-stop level? (3) Are the effects of the investigated factors nonlinear? This paper addresses the need for empirical research that assesses the actual performance of CBs using advanced analytical methods (Liu and Ceder, 2015; Sanaullah et al., 2021; Wang et al., 2014). The insights gained from this study offer CB operators and planners valuable guidance for identifying suitable market niches, designing routes, and allocating stops for CBs.

This study offers three novel contributions to literature. Firstly, it is among the first to analyze the spatiotemporal patterns of CB services. Secondly, it expands upon the limited literature on CB ridership at the stop-to-stop level by comprehensively examining the effects of various influential factors, such as travel impedance and built environment variables, across distinct times of day (morning and evening peak hours). Lastly, this research delves into the nonlinear effects of selected explanatory variables on CB ridership using the XGBoost and SHAP techniques and compares these findings to the conventional multiplicative model to highlight consistencies and discrepancies.

The remainder of the paper is structured as follows: Section 2 provides an overview of the evolution of CBs and presents a review of the key literature on DRT factors influencing demand. Section 3 outlines the methods employed in this study. Section 4 presents the analysis and results. Section 5 discusses the results. Section 6 concludes the paper.

2. Literature review

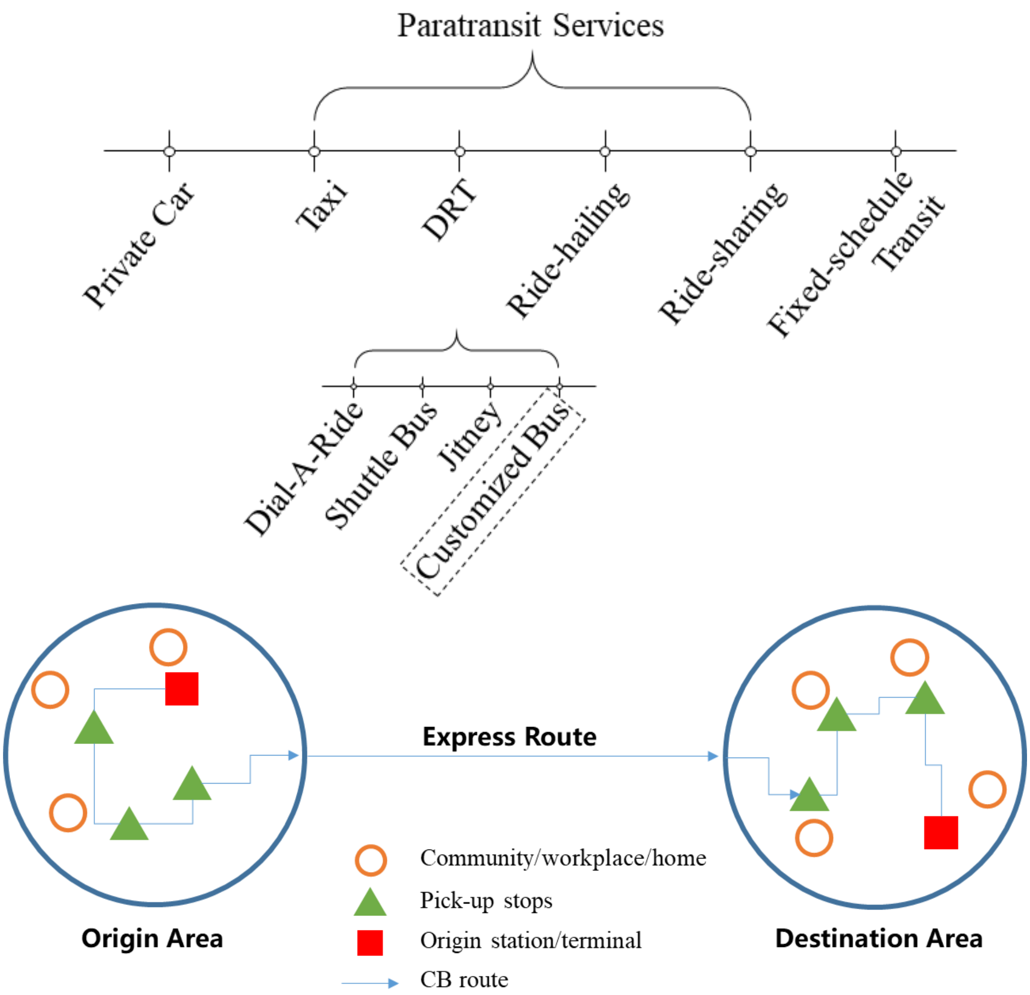
2.1 The historical development of customized buses

The basic concept of customized bus services is not new and has existed since the introduction of car-pooling services in Zurich in 1948 (Shaheen et al., 1998). Early attempts were mainly motivated by economic reasons, allowing individuals to enjoy the benefits of private cars without the costs and responsibilities of ownership. Although similar experiments were initiated elsewhere, the scale of car-sharing practices remained small, and most of them failed due to high maintenance and management costs (Wagner and Shaheen, 1998).

The earliest systematic studies concerning the integration of a car-sharing system into the public transport system, known as ‘subscription bus services’ or ‘dial-a-ride’, were undertaken by Kirby and Bhatt (1974). The services were mainly designed to serve commuters traveling between suburbs and downtown to compensate for areas with poor public transport provision. The services were impressively successful in places where they were attempted for peak-period commuting trips exceeding 10 miles or to suburban workplaces (Bautz, 1975). Some cities in the United States, such as Los Angeles, California, and Ventura, have successfully implemented subscription bus services and realized benefits. Potts et al. (2010) conducted systematic research on the merits of flexible public transport services and produced a practical guide on how public transportation providers can take advantage of flexible PT in their services in different spatial contexts.

In recent years, increasing individual aspirations for mobility have generated a demand for a variety of transport services, such as ‘ride-hailing’, ‘dockless bike-sharing’, and ‘eScooters’, while the rapid development of advanced information and communication technology has substantially changed the way such demand is satisfied (Huang et al., 2020). Customized buses, a relatively innovative form of public transport service, have been proposed to better match travelers’ needs and transport services through apps or online platforms.

As shown in Fig.1(a), customized buses are one type of DRT service (Huang et al., 2020), sharing similar service characteristics (e.g. pre-subscription) with shuttle buses, feeder buses, and other on-demand shared mobility services. Compared with other DRT services, customized buses have the following unique characteristics: (1) advanced subscription (e.g. smartphone app) and (2) aggregated demand with a flexible route/stop/schedule. Hence, customized bus services allow users to obtain a more direct (e.g. one-stop or no-stop) service (as shown in Fig.1(b)) at a cheaper cost compared to taxis or ride-hailing services.



(a) Service category of CB.

(b) Typical example route of CB.

Fig. 1. The service characteristic of CB.

While CB services offer explicit advantages and have been theoretically examined, their long-term financial sustainability remains uncertain, as only a limited number of schemes have achieved commercial viability (Davison et al., 2014). Existing research primarily concentrates on-demand analysis, route design, and operational optimization algorithms (Huang et al., 2020; Li et al., 2019; Lyu et al., 2016; Lyu et al., 2019), focusing mostly on technological and methodological aspects. However, little is known regarding their real-world performance and the factors contributing to their success. To effectively implement and maintain CB development, it is essential to comprehend their spatiotemporal performance and identify the critical factors affecting CB demand. Such insights will greatly benefit operators and potential users of customized bus services.

2.2 Factors associated with customized bus ridership

Numerous studies have assessed the determinants related to the performance of public transit usage rates. Li et al. (2013) reported that average travel time and service reliability act as the main factors contributing to the usage of bus services. Redman et al. (2013) also indicated that service reliability significantly affects public transit performance. In a literature review, Chiou et al. (2015) identified several factors affecting public transit service performance, including demographic and socio-economic factors, land use, car ownership, and public transport provisions.

Multiple studies have investigated the variables affecting the performance of DRT. For example, Nelson and Phonphitakchai (2012) showed that, besides fare level, user demographics (e.g. gender and age) and the overall booking process significantly influenced DRT demand. Wang et al. (2014) employed a multi-level modeling approach in Greater Manchester and found that DRT service demand was higher in neighborhoods with lower car ownership, lower population density, and higher proportions of white people. More recently, Sanaullah et al. (2021) conducted a temporal and spatial analysis of on-demand transit in Belleville, Canada, revealing that travel time, waiting time, and dissemination areas with higher population density and lower median income were associated with higher DRT demand.

Nonetheless, only a handful of studies have explored the factors associated with the performance of customized buses. To our knowledge, Liu et al. (2016) pioneered in comparing the overall performance of public transit, customized buses, and private cars in terms of travel costs, travel time, and fuel consumption for commuting trips in Auckland and France. Their findings indicated that customized bus services were more effective in areas with longer commuting distances, higher population densities, and insufficient public transit accessibility. However, the study's scope was limited by the number of candidate trips (i.e., 200 trips in total), and a larger sample size would be necessary to reach a more robust conclusion.

Li et al. (2019) carried out a study on the determinants of demand for CBs in Shanghai, utilizing revealed and stated preference survey data collected from 1,007 respondents. The findings indicated that fare, travel time, and travel time fluctuations negatively impacted CB demand. Conversely, part-time students, commuters aged 31-40 years, and commuters with high education qualifications were more inclined to opt-in for CBs. Their study enhanced our understanding of how individual sociodemographic attributes may be associated with the demand for CBs. However, it did not explore the impacts of transit supply-related variables (e.g. PT accessibility) and built environment variables on CB demand. Moreover, the analysis was based on data collected from surveys, which can only reflect probabilistic choices made by respondents based on predefined attributes and may not represent actual travel behavior, especially for those who have never used CB services before.

On the other hand, Huo et al. (2021) explored the factors influencing CB ridership based on actual subscription data for a CB service in Xiongan New Area, China. Their findings showed that scenic spots, governmental agencies, and peak periods positively affected CB ridership, while company locations, medical facilities, poor visibility, rain, and weekends were negatively associated with CB ridership. However, their study had a small sample size, covering only one month and 44 stops in the study area, making it difficult to generalize the results. Additionally, Xiongan New Area is a newly developing town established in 2017, and the findings may not be generalizable to other cities.

Recently, Wang et al. (2023a, 2023b) examined the correlations between the built environment and CB usage. Using a large sample of subscription data from Dalian, they discovered that the effects of the built environment on CB usage largely differ from those in conventional transit systems, and most built environment variables exhibit nonlinear effects on CB usage rate. However, as they noted, their findings may not be generalizable to other cities or CB services, and additional research is required to investigate the factors influencing CB performance and ridership in various contexts.

In sum, studies related to key findings on the factors influencing DRT and CB services are summarized in Table 1.

Table 1. Key findings on the factors influencing DRT or CB services.

Study	City	Data	Dependent variable	Methods	Key findings
Wang et al. (2014)	Greater Manchester, UK	17 months DRT subscription records	Annual average number of trips per LSOA	Multilevel model	• DRT services was higher in areas with low car ownership, low population density, high proportion of white people, and high levels of social deprivation.
Liu and Ceder (2015)	Auckland and Paris	100 trips in Paris and 100 trips in Auckland	Difference arrival time, travel time, travel cost, fuel consumption	Quantify operational performance measures	• CB can provide a useful alternative for commuter trips in Auckland and Paris. • For increased commuter trips, CB proved to be more efficient than the PC and PT modes.
Jain et al. (2017)	Melbourne, Australia	Victorian Integrated Survey of Travel and Activity	DRT susceptibility	Correlation of parameters	• Demand patterns of DRT are caused by the spatial variation of demographic characteristics, and travel behavior over the city.
Li et al. (2019)	Shanghai, China	840 questionnaires	Commute mode choice	Multinomial logit models	• Fare, in-vehicle time, and travel time fluctuation negatively affect the potential demand for CB. • Male, low-income, or poorly ducated people show less tendency to shift to CB.
Sanaullah et al. (2021)	Belleville, Canada	9 months DRT subscription records	DRT pick-up and drop-off counts	GIS and K-means clustering	• Higher population density, lower median income, or higher working-age percentages tend to have higher DRT trip attraction levels.

Huo et al. (2021)	Xiongan, China	9 months CB subscription records	The number of CB reservations at the pick-up stations	Poisson regression model	• Scenic spot, governmental agency, morning peak period, and evening peak period are positively correlated with the CB ridership. • Company, medical facility, visibility, rain, and weekend have a negative correlation with CB ridership.
Wang et al. (2023a)	Dalian, China	12 months CB subscription records	Passenger subscription frequency	GBDT model	• Local accessibility at the residence and workplace are the most important correlates of CB use, followed by the proximity of workplace to bus stop. • Distance from workplace to transit stops, distance from residence to business centers, and population density influence CB use differently from traditional transit use.
Wang et al. (2023b)	Dalian, China	22 months CB subscription records	Passenger subscription frequency	Random-effects negative binomial regression model	• A larger residential population, lower employment density, more residential land uses, less administrative land uses, poor connectivity to road networks and parking supply, better accessibility to different facilities, and long distances from urban business centers encourage CB use.

2.3 From stop level to stop-to-stop level

Prior studies have predominantly concentrated on the stop (origin/destination) level, while research at the stop-to-stop or station-to-station level can be viewed as an extension of transit ridership studies (Gan et al., 2020). Although the chosen explanatory variables (e.g. the 5Ds) remain the same, their effects are examined in terms of both origin and destination. Ewing and Cervero (2010) contended that employment densities at destinations might be as important as, or even more important than, population densities at origins for transit and walking trips. This implies that the focus of the built environment and travel behavior literature on population and employment density at the station level may not be ideal for transit-oriented design. Choi et al. (2012), Zhao et al. (2014), and Gan et al. (2020) validated this hypothesis by showing that the effects of the built environment vary between origin and destination, as well as throughout different times of the day. For instance, Choi et al. (2012) discovered that employment density at destination stations was more crucial than population density in explaining metro ridership during morning peak hours, while population density and employment at destination stations were equally significant in accounting for metro ridership during evening peak hours. Similar findings were reported by Zhao et al. (2014) and Gan et al. (2020).

Moreover, stop-to-stop ridership is correlated with travel impedance variables, such as trip distance/time, trip fare, and transfer time, which are key determinants in travelers' decision-making. Therefore, studies at the stop-to-stop level are more advanced in reflecting travelers' actual travel behavior. From a modeling perspective, the stop-to-stop level model allows ridership to be explained by variables associated with both origin and destination stations (Choi et al., 2012; Sohn and Shim, 2010). Furthermore, travel impedance variables or station context variables can be integrated into the gravity model for analysis (Choi et al., 2012).

2.4 A need for nonlinear analysis

Previous studies have extensively explored potential factors influencing the performance of PT, DRT, and CBs using statistical models. These models have the advantage of using a simple formulation and established assumptions, allowing easy interpretation of model estimates. However, a prior expectation in these studies is that the relationship between the dependent variable and independent variables is nonlinear. The models often assume a linear-in-parameters (or log-linear at best) specification within regression (Huo et al., 2021; Wang et al., 2021) or logit models (Li et al., 2019). If these assumptions are false, the parameter estimates for predictors may be incorrect, leading to inappropriate conclusions (Yang et al., 2021).

In practice, independent variables frequently exhibit irregular and nonlinear relationships with dependent variables (Gan et al., 2020; Shao et al., 2020; Yan et al., 2020). As a result, recognizing the nonlinear influences of explanatory variables can better guide transportation planners in understanding the effective range of these variables and help prioritize service enhancement strategies for CB providers (Gan et al., 2020; Shao et al., 2020). Recent research has shown that machine learning-based models surpass linear regression models in terms of model fit and predictive accuracy (Yang et al., 2021).

In summary, this literature review highlights several key findings related to CBs: (1) recent interest in CB services has increased, but the failed cases in some cities suggest that achieving success is not easy, and a thorough understanding of travel patterns and key factors affecting CB ridership is urgently needed (Davison et al., 2014); (2) built environment and travel impedance variables are important factors associated with transit usage, but their coefficients and effects on CBs ridership have not been specifically investigated in the literature, especially at the stop-to-stop level; (3) machine learning-based methods provide an opportunity to explore the salient and nonlinear effects between the dependent variable and independent variables.

Building on the literature, this study is aimed to fill research gaps and improve our understanding of CBs by employing machine learning-based methods and actual subscription data for CBs over two years in Shanghai. Specifically, we aimed to investigate: (1) the spatiotemporal patterns of CBs, and (2) the importance and the extent of the impacts of travel impedance and built environment variables on CB ridership.

3. Methods

3.1 Study area and variables

The geographic focus of this study was the city of Shanghai, one of the largest and most densely populated cities in China. As of 2019, Shanghai had a population of 24.9 million and is a highly urbanized area spanning 6,340.5 km^2. To address the increasing travel demand in the city, Shanghai authorities have implemented various measures, both on the supply and demand sides, to encourage more use of public transit, including the introduction of a CB service in 2017.

This study analyzed user subscription data from Yidong, Shanghai's largest CB service provider, covering a total of 1.15 million CB subscriptions from May 2017 to October 2019. Each subscription record contained attributes such as the unique identifier of the traveler, trip date, start time, fare, pick-up (origin) stop, and drop-off (destination) stop. To prepare the data for analysis, all stops were first geocoded using the Map Location interface tool on maplocation.sjfkai.com. The total number of trips for each origin-destination (O-D) pair was calculated, serving as the dependent variable in the model. The variable represented an accumulated count over the entire study period, rather than on a monthly basis, which avoided potential serial correlation issues within the model.

This resulted in a total of 2,277 O-D pairs for morning peak hours and 2,172 O-D pairs for evening peak hours. The total number of daily CB trips was 1,164, the average daily number of trips was 567 during the morning peak and 412 during the evening peak.

Based on the literature review in Section 2, this study selected two categories of explanatory variables: (1) travel impedance variables and (2) built environment variables. The travel impedance variables include trip distance, trip departure time, and trip fare, which were obtained directly from the subscription data. Additionally, using the API Web Services from Amap (the equivalent of Google Maps in China), we generated several new attributes based on the observed O-D pairs, including the trip fare for a taxi service for the same O-D journey. It is important to highlight that socio-economic factors are a recognized category of determinants that can affect CB ridership (Nelson and Phonphitakchai, 2012; Wang et al., 2014; Wang et al., 2023), but due to privacy concerns, the socio-economic profiles of users were not available in the dataset, so we were unable to include these variables in our model.

Regarding the built environment variables, this study computed the 5D variables (i.e., density, diversity, design, destination accessibility, distance to transit), a widely recognized framework for categorizing aspects of the built environment (Ewing and Cervero, 2010). Density was measured using population and employment density data obtained from China Unicom, one of the largest mobile providers in China. These data provide more fine-grained population density information compared to census data and have been validated repeatedly for their high quality (Yang et al., 2023). Diversity was measured using land-use entropy (Cervero and Kockelman, 1997) based on point of interest (POI) data from Amap, which have been shown to represent the layout characteristics of urban functional facilities and services at a fine-grained level (An et al., 2019). Design was assessed using road density data acquired from Open Street Map (OSM). Destination accessibility evaluates the ease of reaching trip attractions. Since accessibility can be either regional or local, the distance to the city center was employed to represent regional accessibility, while the number of jobs and amenities were utilized as proxies for local destination accessibility. To measure amenity accessibility, we used the number of POIs within a predefined distance in categories such as beauty, catering, training, shopping, entertainment, and sport. Previous studies have shown that these types of POIs have a significant impact on metro ridership (An et al., 2019). This variable was included in our analysis to examine whether these impacts also apply to CB services. Distance to transit was measured as the distance to the nearest metro station and the total number of bus stops. Due to data availability, the frequency and number of transit services were not considered in this study.

After checking for multicollinearity, we excluded some factors (e.g. trip time, expense saving compared to a taxi, time-saving compared to PT service, and employment density) that had VIF scores greater than 10. Table 2 presents the descriptive statistics for all the variables investigated in this study.

Table 2. Descriptive statistics for variables.

Variable	**Mean**	**St.dev.**	**Source**
Dependent variable			
Stop-to-stop ridership (morning peak)	225.3	1409.2	Yidong
Stop-to-stop ridership (evening peak)	172.9	1104.9	Yidong
Travel impedance variables			
Trip distance (km)	18.3	15.0	Measured in ArcGIS
Trip fare (CNY/km)	1.7	3.9	Yidong
Built environment variables (radiu = 500 m)			
Population density (10,000 persons per km^2)	0.57	0.99	China Unicom
Land use mix (index)	1.57	0.67	Amap
Road density (km/km^2)	5.21	2.50	Open Street Map
Job accessibility (count)	133.80	195.62	Zhaopin (China)
Amenity accessibility (count)	137.67	195.68	Amap
Distance to the city center (km)	26.55	15.49	Measured in ArcGIS
Distance to metro station (km)	2.68	5.33	Measured in ArcGIS
Number of bus stops (count)	10.89	10.17	Amap

3.2 Analytical approach

3.2.1 Machine learning approach: eXtreme Gradient Boosting

Fig.2 provides an overview of the modeling framework. 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 (Chen and Guestrin, 2016; Liu et al., 2021; Yang et al., 2021; Zhang et al., 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 our 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.

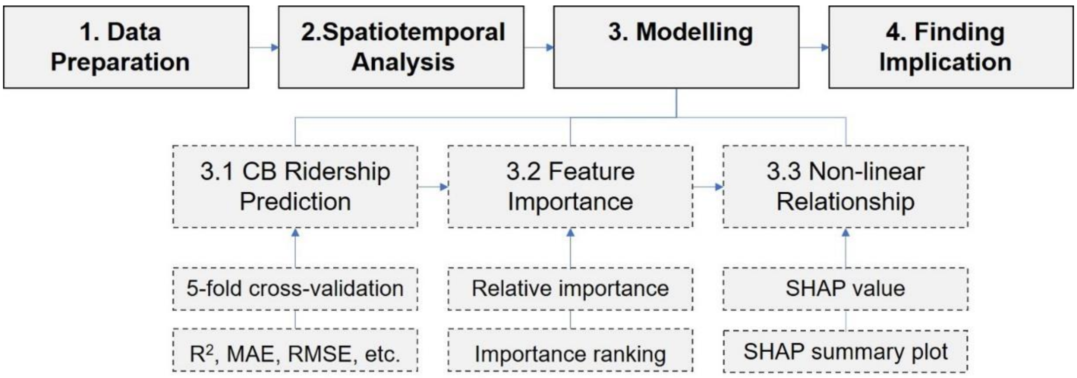


Fig. 2. Analysis framework.

Mathematically, assume the dataset consists of n samples, and the trees are constructed from m dimensions of the features (i.e., input variables) $x_i (x_i \in R_m)$. For each tree, the outcome (i.e., response variable) Y_i exists, and the tree is built based on K additive functions. Therefore, the XGBoost model can be formulated as:

$$\widehat{Y}_i = \hat{f}(X_i) = \sum_{k=1}^{K} f_k(X_i) ,\cdot f_k \in F \tag{1}$$

where $\mathcal{F} = \{f(X) = \omega_{q(X)}\} (q: \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T)$ refers to the space of all regression trees; q represents the tree structure, and T represents the total number of leaves in each tree. f_k is an individual tree configured with structure q and leaf weight ω.

A limitation of XGBoost and machine learning models, more generally, is that in order to obtain the best performance from XGBoost it often requires careful 'tuning' of the hyperparameters (Yang and Shami, 2020). If not properly regulated, XGBoost may overfit the training data, which can result in poor generalization to unobserved data. Also, the primary goal of models like XGBoost is to maximize predictive accuracy, rather than providing interpretable parameter estimates and measures of uncertainty that are common in traditional statistical models (Li, 2022). For tasks that require understanding the effect of individual predictors or quantifying uncertainty, traditional statistical models may be more appropriate. This is the main reason that we included both the XGBoost and conventional statistical regression models in our study. By using both machine learning model and and conventional statistical regression model, we can take advantage of the strengths of each and provide a form of model validation or robustness check.

During the hyperparameter tuning process, we evaluated the XGBoost models with varying values for tree complexity (3, 4, 5, 6) and shrinkage (0.1, 0.05, 0.01, 0.005) using a fivefold cross-validation procedure. We allocated 90% of the dataset as learning samples and the remaining 10% as testing samples. The number of trees was examined by setting the indicator from 2,000 to 10,000 at intervals of 1,000. Ultimately, the XGBoost models gained the best hyperparameters with 10,000 trees, a learning rate of 0.01, and a depth of 5 to prevent overfitting.

Moreover, many machine learning algorithms can tolerate multicollinearity to some extent, however, the issue of multicollinearity does persist in machine learning algorithms (Garg and Tai, 2013) and thus it is worth checking for multicollinearity for several reasons. Firstly, in the context of machine learning, feature importance is often assessed, but if two variables are highly correlated, the model may split the importance between these variables, making it difficult to ascertain the individual importance of each variable. Secondly, if two features are highly correlated, removing one of them may not result in a significant decrease in model performance. By removing highly correlated variables, the model can be made simpler and its estimation faster without a significant loss of predictive accuracy. Thirdly, multicollinearity can cause instability in the model predictions because it can result in large changes to the predicted outcomes for small changes in the input variables, which is an inherent characteristic of machine learning algorithms (Dormann et al., 2013). This is the main reason why a multicollinearity test had been performed since machine learning algorithms were used.

3.2.2 Interpretation of machine learning: SHapley Addictive exPlanations

To interpret the XGBoost regression model and understand how travel impedance and the built environment metrics are associated with CB ridership, this study applied the SHapley Addictive exPlanations (SHAP) technique. The SHAP technique is an emerging machine-learning method proposed by Lundberg and Lee (2017) and measures the contributions of each input variable in the XGBoost model and examines the extent to which those variables affect CB ridership. The Shapley value of a variable i is measured as follows (Shapley, 1997):

$$\varphi_i = \sum\nolimits_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_s(x_s) \right] \tag{2}$$

where S represents the variable subset, F is the set of all variables, and $f_{S\cup\{i\}}$ and $f_S(x_S)$ denote the model trained with and without feature i, respectively.

The Shapley value, φ_i, for a given feature i can be positive, negative, or zero, indicating whether that feature increases the prediction, decreases the prediction, or does not affect the prediction, respectively, compared to the average prediction. For instance, positive SHAP value ($\varphi_i > 0$) indicates that the presence of the feature i increases the prediction for the specific instance compared to the average prediction. In other words, the feature has a positive contribution to the predicted outcome.

Previous studies have used partial dependence plots to display the direction of the studied variable and its marginal effect on the predicted outcome, these plots cannot explain the XGBoost method (Yang et al., 2021). XGBoost and SHAP are often paired in machine learning because they together offer a comprehensive solution for prediction and interpretability (Lundberg et al., 2020; Ziakopoulos et al., 2023). The resulting SHAP value demonstrates how a model influences the target value, as well as how independent feature values within the model are related to one another. Thus, the SHAP method offers the context-specific importance of the variable compared to global interpretation techniques (Li, 2022).

The local interpretation approach facilitated by SHAP allows the analysts to extract spatial relationships in applications dealing with spatial data, particularly when observations are geocoded. Through both simulated and empirical examples, Li (2022) found that spatial effects can be interpreted by mapping SHAP values derived from interaction effects related to location features. These values showed consistency with estimates produced by spatial statistical models like the Spatial Lag Model (SLM) and Multi-scale Geographically Weighted Regression (MGWR) at the parameter level. Therefore, the SHAP method emerges as a valuable alternative to traditional spatial statistical models and performs better when complex spatial and non-spatial effects (e.g. non-linearities, interactions) coexist and are unknown.

3.2.3 The multiplicative model

We also employed the conventional statistical regression model in this study, the multiplicative model, which is a frequently used method for modeling station-to-station patronages (Choi et al., 2012; Gan et al., 2020; Zhao et al., 2014). A multiplicative model is a form of a mathematical model where the effects of individual factors are multiplied together rather than added (Wooldridge, 2015). This model differentiates itself from traditional count data models such as Poisson or Negative Binomial by assuming the effects of predictors are not just additive, but also multiplicative. The multiplicative model can be expressed as follows:

$$T_{ij} = \emptyset \prod_{p=1}^{P} X_{ip}^{\alpha_p} \prod_{p=1}^{P} X_{jp}^{\beta_p} \prod_{q=1}^{Q} M_{ijq}^{\theta_q} \tag{3}$$

where T represents the trips from a stop i to stop j, X_{ip} is the pth independent variable regarding the pick-up stop i, and X_{ip} is the pth independent variable regarding the drop-off stop. M_{ijq} is the qth independent variable related to the trip characteristics of a CB trip from stop i to stop j; $\emptyset$ is the scale parameter; and α_p, β_p, and θ_q are the model parameters to be estimated. P is the number of independent variables in the built environment, and Q is the number of independent variables in the trip characteristics. The equation has been transformed into the linear form by taking the natural logarithm on both sides of Eq. (3).

Several advantages make the multiplicative model preferable over count data models. Firstly, it better performs in capturing interaction effects, illustrating how predictors collectively influence the outcome variable rather than merely highlighting their independent contributions, which is particularly beneficial for many station-to-station studies (Choi et al., 2012). Secondly, the multiplicative model exhibits flexibility by its ability to transform into an additive model through a logarithmic transformation. It simplifies the model, making it easier to solve, interpret, and visualize. The transformation also facilitates the interpretation of coefficients as effects of unit changes in predictors, while holding other variables constant (Dios Ortuzar and Willumsen, 2011; Wooldridge, 2015).

Logarithmic transformation of the dependent variable was used in the models to normalize the dataset, this transformation helped to achieve a normal distribution of the data. To facilitate a comparison between the XGBoost model and the multiplicative model, we implemented a natural logarithm transformation of the dependent variable in the XGBoost models.

4. Results

4.1 Exploratory spatiotemporal analysis of customized buses in Shanghai

The total numbers of monthly CB subscriptions from May 2017 to October 2019 are shown in Fig.3. During the early stage between May 2017 and August 2017, the CB service only attracted a few passengers. However, from October 2017, the number of subscriptions increased significantly after the launch of the online subscription app. The monthly subscriptions continued to grow and reached a record high of 78,037 per month in July 2019.

This upward trend in subscriptions during the study period reflects the growing acceptance of CB services in Shanghai. It is noteworthy that there were no aggressive marketing strategies (e.g. advertising, personal selling, or sales promotion) by CB service providers, nor were any subsidies provided by the local government. The increasing patronate purely came from the acceptance of travelers, which indicates a genuine demand and necessity for CB services in the city.

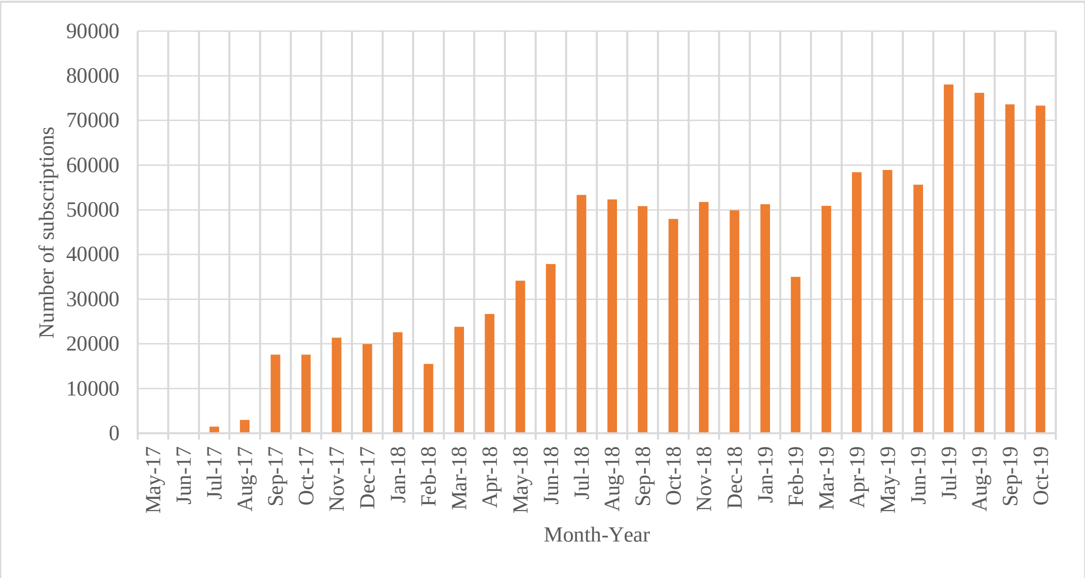


Fig. 3. The numbers of monthly subscriptions from May 2017 to October 2019.

The daily travel distribution is shown in Fig.4 and illustrates a strong regularity of CB services, with recurrent patterns over the study period. The daily travel patterns using CB services featured two peaks. A morning peak was observed from 6 to 8 a.m., which was one hour earlier than the conventional morning peak time of 7 to 9 a.m. in Shanghai. One of the main reasons for this phenomenon is that CB users typically live in suburban areas and tend to leave earlier to arrive at their workplace on time after long-distance commutes. The evening peak occurred around 6 p.m., which aligns with the evening peak for Shanghai, however, the peak share of subscriptions during the evening peak was only half that of the morning peak, and the use of CB spread until 21:00-22:00. This discrepancy may arise from people's varying sensitivity to time flexibility during these two peak periods (Oakil et al., 2016). For example, CB users might engage in various activities, such as dining, shopping, or other recreational pursuits, before heading home due to the increased flexibility in travel time, thus making the choice of a CB service less of a priority.

In comparison, the time for home-to-work commuting in the morning is somewhat inflexible, and CB users decide to use CB services due to their time-efficient advantages (e.g. more direct routes, guaranteed departure, and arrival times) compared to conventional bus services.

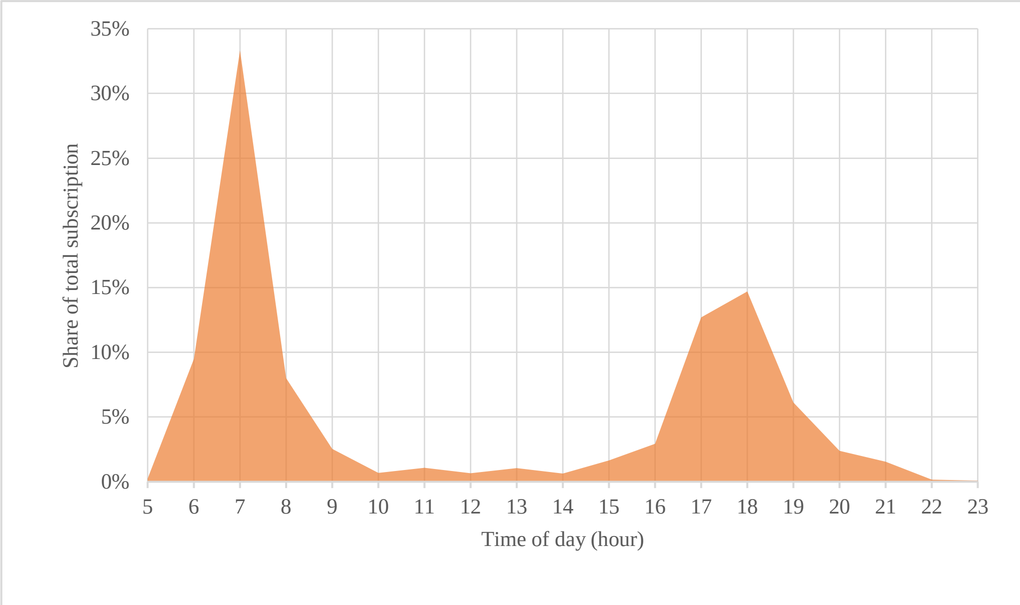
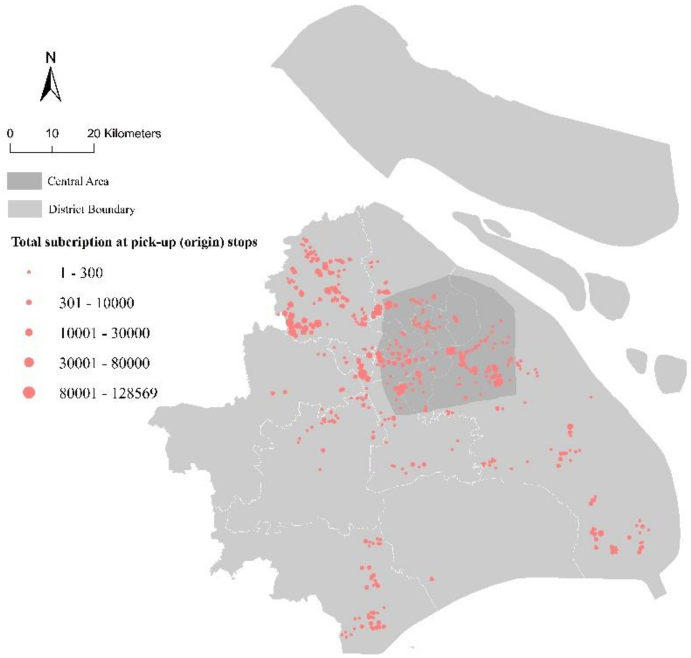
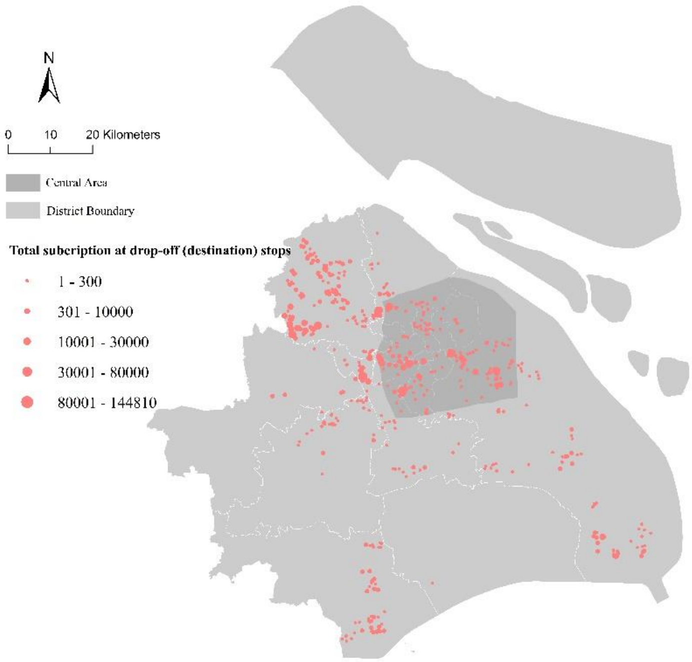


Fig. 4. Distribution of subscriptions in the daytime.

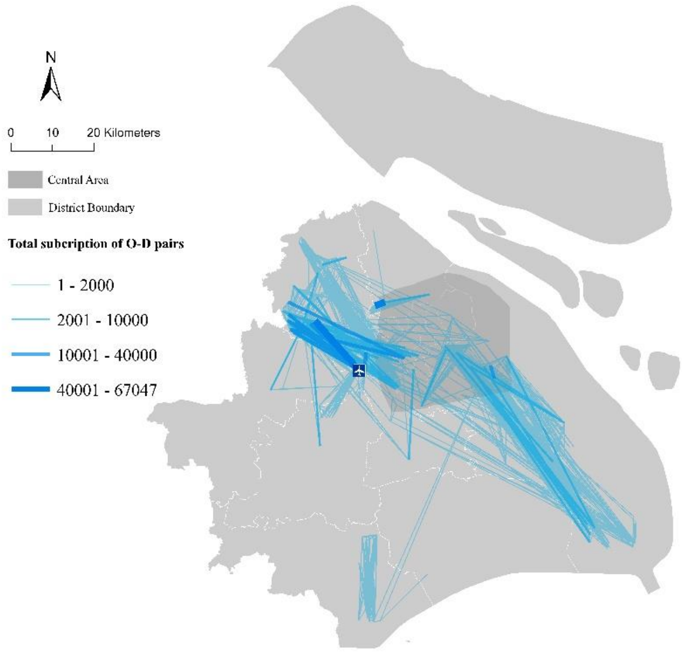
The spatial distribution of CB services can be seen in Fig. 5. First, from Fig. 5(a) and 5(b), it is apparent that the spatial distributions of pick-up and drop-off stops are quite similar. Stops with high pick-up rates also exhibit similar levels of drop-offs, suggesting a balanced distribution of CB demand. In practice, CB service operators initially design their routes and stops based on user travel requests. These routes and stops tend to stabilize once a sustained demand is observed. Secondly, Fig.5(c) illustrates the spatial distribution of O-D trip flows. CB services in Shanghai mainly cater to long-distance travel between suburban areas and the central city, with an average trip distance of approximately 20.9 km.



(a) Spatial distribution of pick-up (origin) stops



(b) Spatial distribution of drop-off (destination) stops



(c) Spatial distribution of trip flows

Fig. 5. Spatial distribution of customized bus services.

To validate and understand the spatial interaction patterns of CB services, a spider chord diagram was used to visualize the actual time-of-day origin-destination travel flows. The dataset did not include information on stop categories, so we manually identified the categories of both origin and destination stops based on the main categories of land and POI, resulting in 2,558 stops being categorized into 13 types.

As shown in Fig.6, the spatial interaction patterns for the CB service demonstrated strong regularity during the morning and evening peak hours. Most flows in the morning peak came from residential stops and dispersed to various destinations, including transport hubs, community centers, and offices. During the study period, trips to the airport and railway station were among the most popular services, indicating that the CB service not only serves commuting trips between suburbs and downtown but also acts as an attractive transport hub transfer service. Many of which occurred between the airport and various other destinations, with these origin-destination pairs typically spanning a distance of 20 km. Community center stops, which are neighborhood centers with various daily activities, had high trip production/attraction levels, suggesting that the CB service also provides one-to-many and many-to-one express services, reducing riders' overall waiting time and enhancing service efficiency by collecting more passengers at one time.

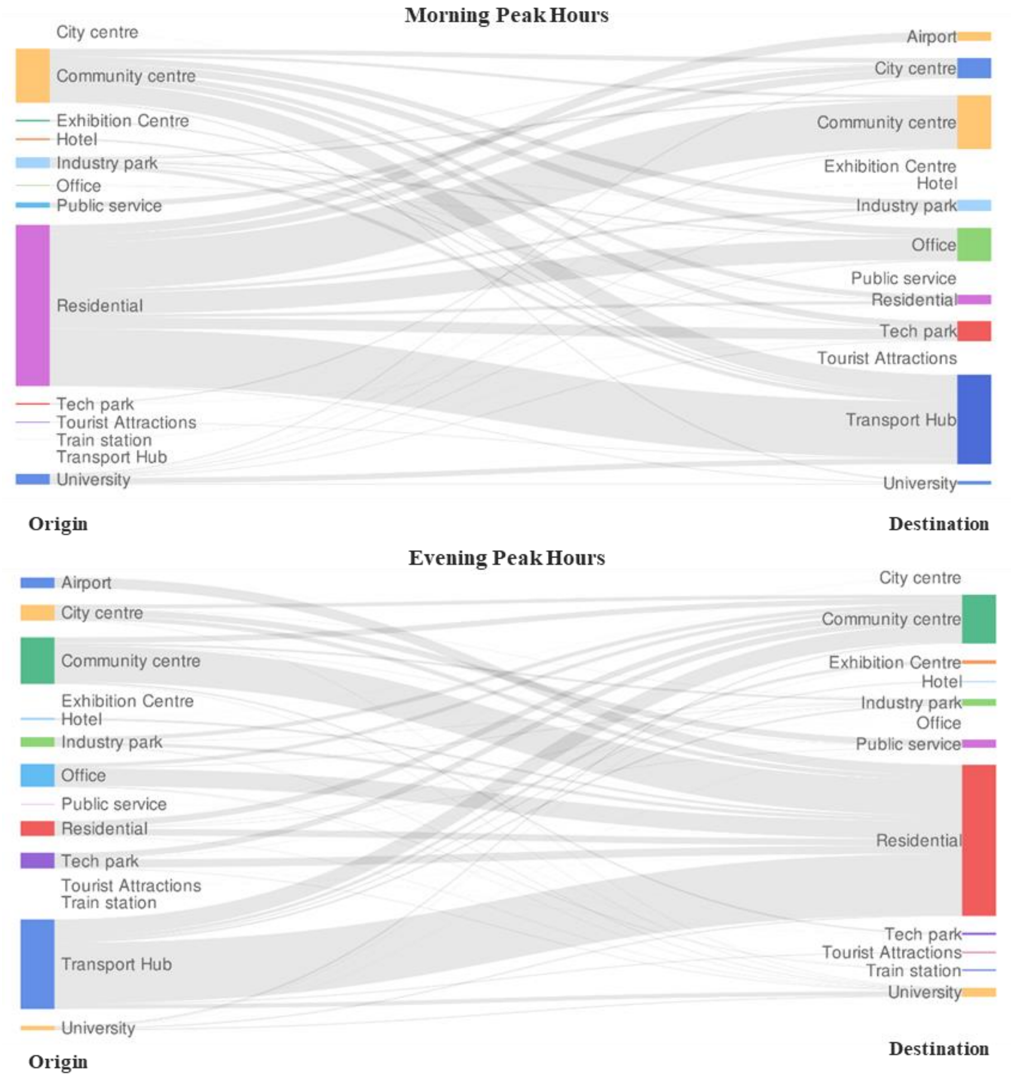


Fig. 6. Travel flows for CBs in the morning and evening peak hours.

4.2 The relative importance of independent variables

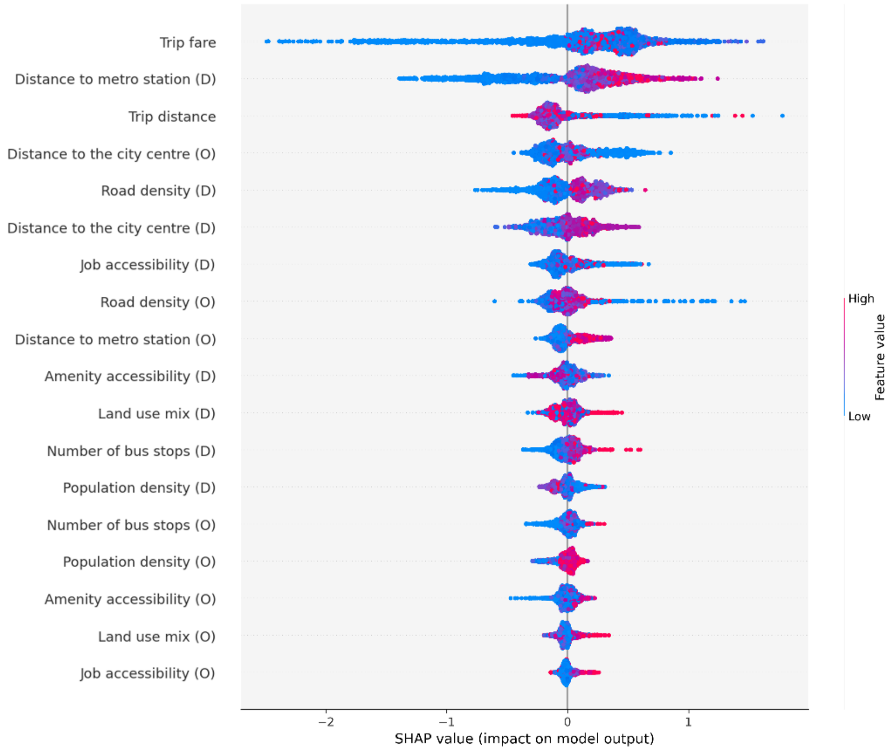
Table 3 and Fig. 7 present the relative importance and ranking of independent variables for predicting stop-to-stop CB ridership. A higher relative importance (RI) value indicates stronger predictive power. The findings reveal that for the morning and evening peaks, the combined contributions of built environment variables on the origin stop side were approximately 38% and 45%, respectively, in predicting stop-to-stop ridership. On the other hand, the built environment variables on the destination stop side contributed 45% and 39%, respectively. These results suggest that for the morning rush hour, the built environment on the destination stop side has a greater impact on stop-to-stop ridership. Whereas, for the evening rush hour, the opposite holds for the origin stop side. Furthermore, the combined contributions of the two travel impedance variables for the morning and evening peaks were 18% and 16%, respectively, indicating that travel impedance variables have a constant and significant effect on the stop-to-stop CB ridership.

Table 3. Estimates of the multiplicative model and the XGBoost model

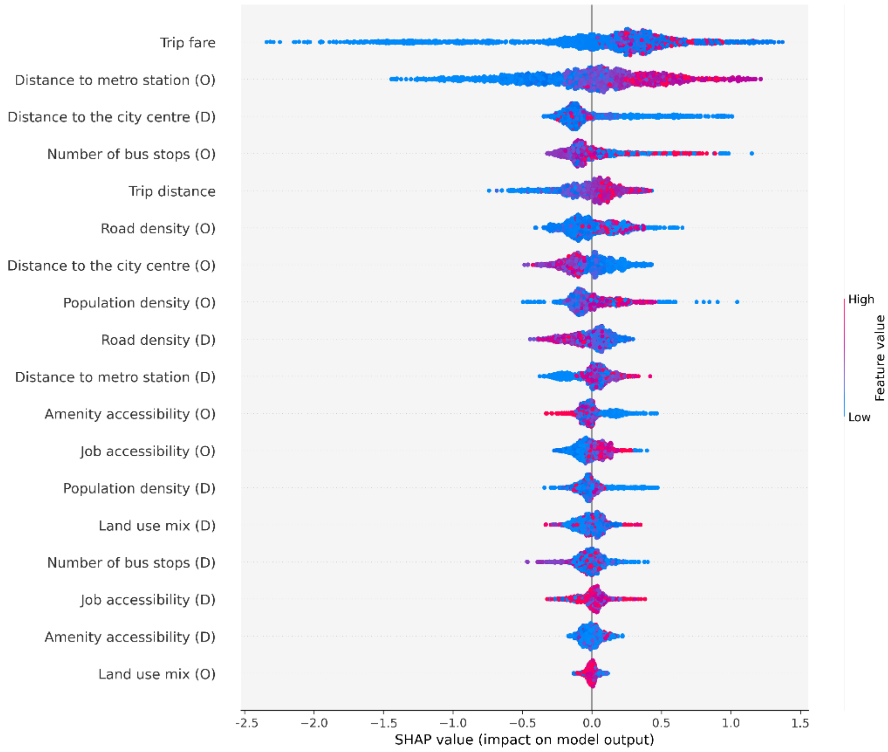
	Morning Peak Hours Model				**Evening Peak Hours Model**			
	Multiplicative Model		**XGBoost Model**		**Multiplicative Model**		**XGBoost Model**	
	Coef.	**S.E.**	**RI**	**Ranking**	**Coef.**	**S.E.**	**RI**	**Ranking**
Travel impedance								
Trip distance	0.45***	0.075	7.45%	3	0.42***	.073	6.81%	5
Trip fare	-0.30***	0.069	10.13%	1	-0.26***	.050	9.12%	1
Origin stop								
Population density	-0.03	.061	4.31%	15	0.01	.046	5.25%	8
Land use mix	-0.01	.075	3.99%	17	0.04	.074	3.29%	18
Road density	0.01	.024	5.03%	8	0.03	.021	6.29%	6
Job accessibility	0.08***	.001	3.82%	18	-0.09***	.001	4.48%	12

Amenity accessibility	-0.03	.000	4.25%	16	0.20***	.000	4.76%	11
Distance to the city centre	0.14***	.000	7.04%	4	0.13***	.000	5.92%	7
Distance to metro station	-0.05	.000	4.79%	9	0.02	.000	8.46%	2
Number of bus stops	0.03	.006	4.56%	14	-0.14***	.006	6.83%	4
Destination stop								
Population density	0.02	.056	4.57%	13	-0.04	.068	4.38%	13
Land use mix	0.07***	.078	4.73%	11	0.01	.074	4.31%	14
Road density	0.02	.020	6.41%	5	0.05*	.025	5.18%	9
Job accessibility	-0.10***	.001	5.51%	7	0.07**	.002	4.23%	16
Amenity accessibility	0.20***	.000	4.77%	10	-0.05	.000	4.19%	17
Distance to the city center	0.08*	.000	6.40%	6	0.12***	.000	7.39%	3
Distance to metro station	0.02	.000	7.54%	2	-0.11**	.000	4.80%	10
Number of bus stops	-0.11	.006	4.70%	12	-0.05*	.007	4.31%	15
(Pseudo) R^2	0.133		0.476		0.128		0.478	
MAE	2.795		1.193		2.960		1.206	
RMSE	1.994		1.493		1.945.		1.482	
N of Obs	2277				2051			

Significance codes: '***' $p < 0.01$; '**' $p < 0.05$; '*' $p < 0.1$; RI: Relative Importance



(a) Morning peak hours



(b) Evening peak hours

Fig. 7. SHAP summary plots of independent variables.

The three most important independent variables affecting stop-to-stop CB ridership during morning peak hours (6:00-8:00) were trip fare, distance to the nearest metro station at the destination stop side, and trip distance with relative importance values of 10.13%, 7.54%, and 7.45%, respectively. During evening peak hours (17:00-19:00), trip fare, distance to the metro station at the origin stop side, and distance to the city center at the destination stop side were the most influential variables with relative importance values of 9.12%, 8.46%, and 7.39%. This indicates that travel impedance (trip fare and distance) and destination accessibility (such as proximity to the city center) have significant and consistent impacts on stop-to-stop CB ridership, regardless of the peak period.

Population density had a minor effect on CB ridership, with relative importance values of 4.31% (rank 15) and 4.57% (rank 13) at the origin and destination stops during the morning peak. During the evening peak, population density has a smaller impact, with relative importance values of 5.25% (rank 8) and 4.38% (rank 13) at the origin and destination stops, respectively. This suggests that population density is not a decisive determinant of CB demand.

The land use mix has small and similar relative importance values for both morning and evening peaks, with values below 3.99%. In terms of road density, a higher relative contribution was observed during evening peak hours (5.03% and 6.41% vs. 6.29% and 5.18%). Compared to other built environment variables, job accessibility, and amenity accessibility have relatively small relative importance values for both stop sides and peak periods. During the morning peak hours, job accessibility had the highest relative importance on the destination side (rank 7), contributing 5.51%. On the other hand, during the evening peak hours, amenity accessibility had the highest relative importance on the origin side (rank 11), contributing 4.76%.

The distance to transit is a crucial factor affecting CB ridership, although its relative importance differs by stop side and peak period. During morning peak hours, the distance to the nearest metro station has lower relative importance on the origin side (rank 9), contributing 4.79%, but higher relative importance on the destination side (rank 2), contributing 7.54%. Similar findings were observed for the number of bus stops. These results imply that the impact of factors related to transit accessibility can vary considerably depending on the stop side and peak period.

Furthermore, we compared the performance of the XGBoost models with multiplicative models. Our results indicate that the XGBoost model performed better, showing a higher pseudo-R^2 and lower root-mean-square error (RMSE) and mean absolute error (MAE). Having said that, the main objective of the study was not to achieve better forecasting performance, and the ranking of the relative importance of independent variables in the XGBoost models can be very sensitive to the data quality, therefore, we further extended our analysis by comparing the results with those from multiplicative models.

Some interesting consistencies and divergences between the two models were observed. Firstly, variables such as trip distance and trip fare, which had significant correlations with the dependent variable in the multiplicative models, also demonstrated strong contributions in the XGBoost models. Secondly, some variables, such as job accessibility, that had high relative importance in the XGBoost models showed different results in terms of significance and coefficients in the multiplicative models. This difference may have been due to the different modeling mechanisms between machine learning-based models and conventional statistical models. For example, multiplicative models assume a log-linear relationship between log-transformed dependent and log-transformed variables. In contrast, the XGBoost model uses machine learning algorithms to estimate possible nonlinear relationships more feasibly. Therefore, the XGBoost model is better suited to capturing the complex relationships that exist between the dependent variable and independent variables when they are nonlinear. As a result, the XGBoost model provides more accurate predictions than the multiplicative models.

4.3 Nonlinear relationships with customized bus ridership

Although XGBoost models can identify explanatory variables with significant correlations to stop-to-stop ridership, they cannot determine the magnitude of the impact that independent variables have on the dependent variable. To address this, we utilized SHAP plots for each independent variable to conduct a more comprehensive analysis of the nonlinear and threshold effects of the selected explanatory variables on stop-to-stop CB ridership.

SHAP plots are valuable tools for visualizing the connection between independent variables and dependent variable. They illustrate the independent variable on the *x*-axis and its corresponding marginal effect, represented by the Shapley value, on the *y*-axis. Unlike partial dependence plots that provide an average effect for an independent variable, SHAP plots display the variance in the effect for each independent variable with each dot of the feature value on the *y*-axis. This allows for a more detailed understanding of the relationship between independent variables and the dependent variable.

Fig.8 illustrates the relationship between travel impedance variables (trip distance and trip fare) and stop-to-stop CB ridership during morning and evening peak hours. The plots reveal a nonlinear association between travel impedance and ridership, independent of stop side and peak hours. Furthermore, trip distance on both sides displays a positive correlation with stop-to-stop CB ridership, which aligns with the outcomes of the multiplicative models in Table 3.

For both morning and evening rush hours, trip distance displays a positive, albeit fluctuating correlation with stop-to-stop ridership when it ranges from 0 to around 30 km. Once it surpasses 30 km, an increase in trip distance does not guarantee an increase in ridership. These results indicate that travelers tend to use CB services more for long-distance trips, but there is a threshold distance beyond which the effects of travel distance on CB ridership are insignificant.

Additionally, there is a negative correlation between trip fare and stop-to-stop ridership, which also corresponds with the results of the multiplicative models. The correlation is apparent when the trip fare ranges from 1 to 5 CNY/km, beyond which the effect becomes negligible. These findings contribute to the current understanding of the influence of trip distance and fare on stop-to-stop CB ridership. It is generally acknowledged that CB services are more attractive to passengers with long-distance commutes, and the trip fare is inversely related to CB ridership. Our study further provides insights into the specific range of the effects of travel impedance variables on stop-to-stop CB ridership.

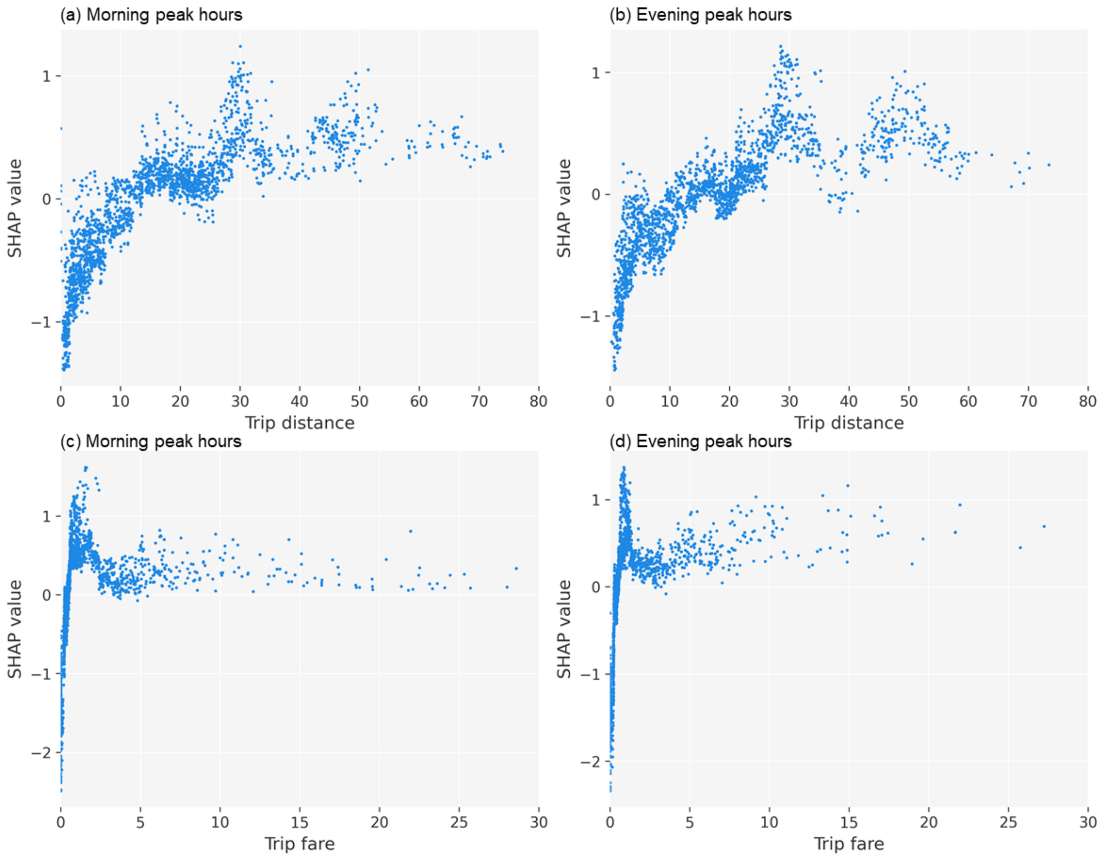


Fig. 8. Threshold effects of travel impedance variables on CB ridership.

To investigate the link between population density and stop-to-stop CB ridership, we generated SHAP plots, as illustrated in Fig. 9. The plots demonstrate a dispersed and unstable relationship between population density and stop-to-stop CB ridership on different stop sides during different rush hours. These results are consistent with those obtained from the multiplicative models presented in Table 3 and align with the research of Wang et al. (2023a), which found that population density has a negligible and statistically insignificant effect on stop-to-stop CB ridership.

High population density typically has a positive effect on the use of public transit demand (Ewing and Cervero, 2010). However, our study suggests that the positive effect of population density on public transit demand cannot be applied to CB services. One possible explanation for this discrepancy is that CB services are primarily designed to complement traditional transit systems in low-density areas, meaning that the association between density and service demand may not be the same as for public transit services. In densely developed areas, travelers tend to have more transit options and are more likely to prefer public transit services over CB services.

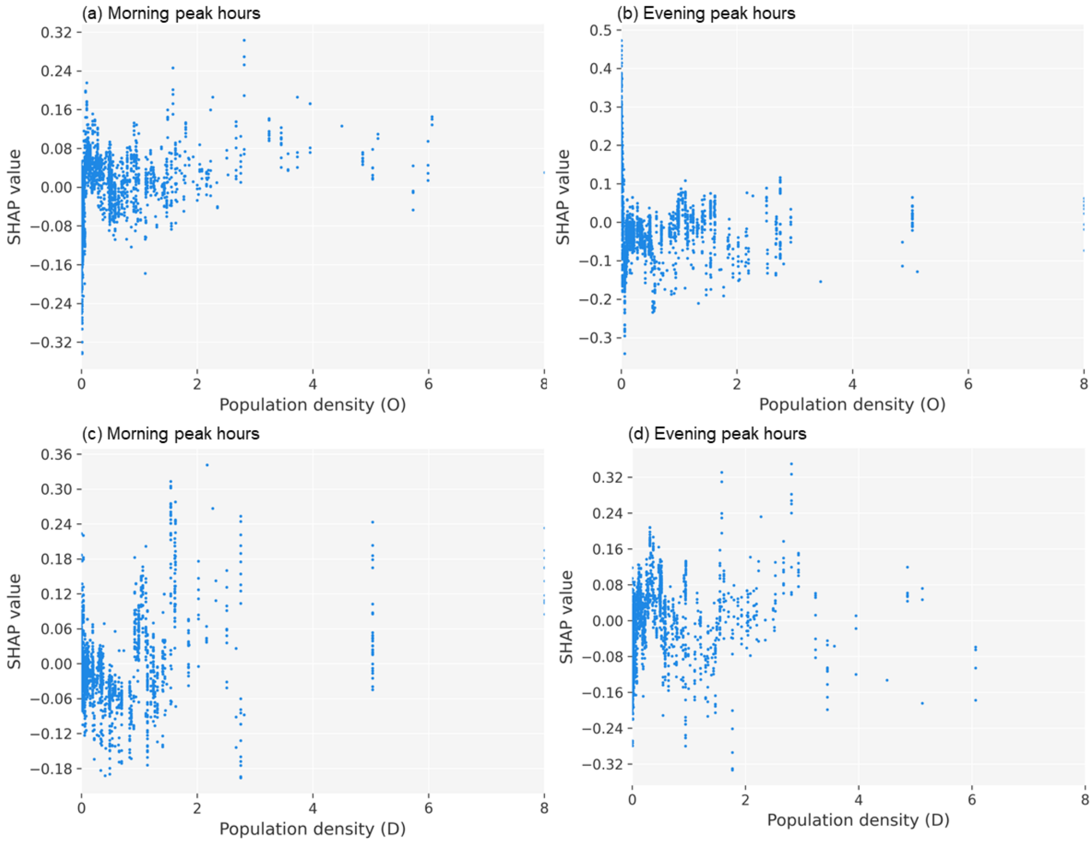


Fig. 9. Threshold effects of population density on CB ridership.

The relationship between land use mix and stop-to-stop CB ridership on the origin and destination sides during the two rush hours is depicted in Fig. 10. Fig. 10(a) indicates a nonlinear pattern with a steady increase followed by a peak and then a subsequent decline (inverted-U shape). Fig. 10(b) also shows a nonlinear pattern with an initial increase followed by a decrease once the land use mix index exceeds 1.5.

These results are partially consistent with those obtained from the multiplicative models presented in Table 3, which found a significantly positive association between land use mix at the destination side during morning peak hours. However, our results further suggest that this positive effect has a threshold, meaning that once the land use mix exceeds 1.5, more diverse land use does not necessarily lead to an increase in CB ridership. Additionally, Fig.10(c) and 10(d) show that the land use mix for the evening rush hours is not related to stop-to-stop CB ridership on either side, as the dots in the plots are nearly horizontal and dispersedly distributed.

Overall, the findings in this study indicate that the impact of land use mix on stop-to-stop CB ridership varies depending on the stop side and time of day. Nonlinear patterns were observed, and there is a threshold effect that must be taken into account when evaluating the influence of land use mix on CB ridership.

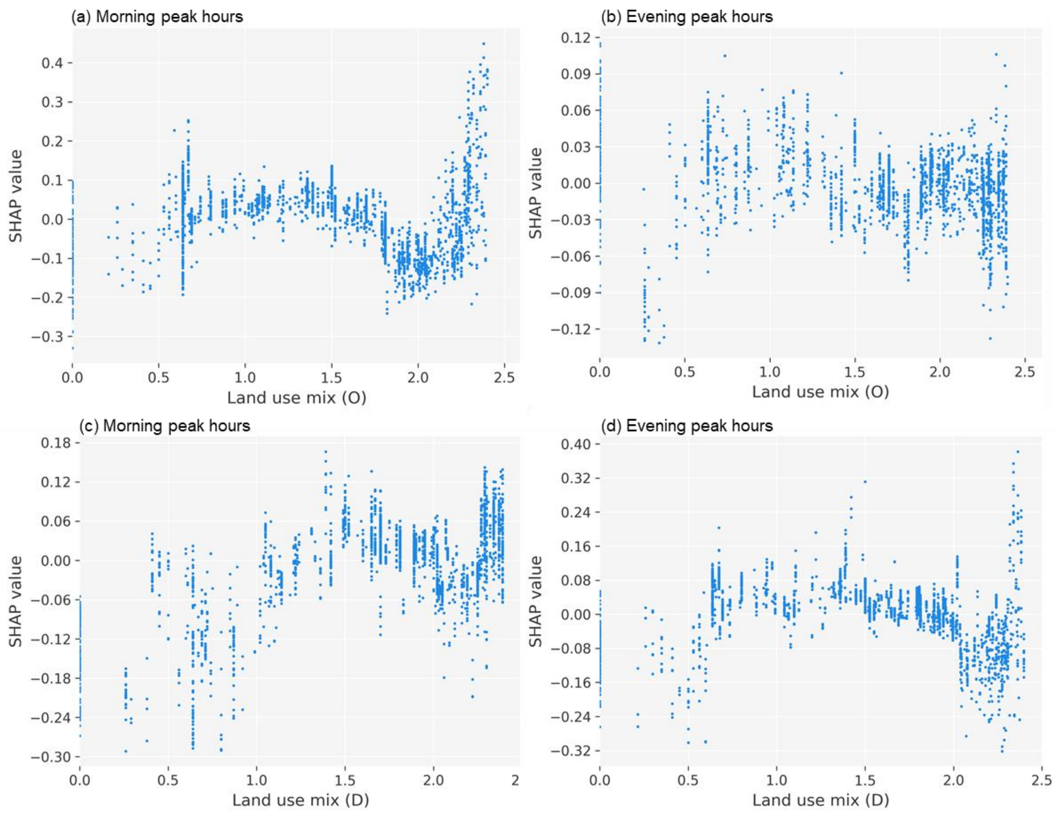


Fig. 10. Threshold effects of land use mix on CB ridership.

Fig. 11 provides a summary of the effects of road density on stop-to-stop CB ridership on both the origin and destination sides during the two rush hours. Fig. 11(a) and 11(d) exhibit comparable patterns, with dots horizontally scattered on the plots, indicating that stop-to-stop CB ridership is unrelated to road density on the origin side during morning peak hours or on the destination side during evening peak hours. Fig. 11(b) and (c) display similar nonlinear patterns, with an initial decrease followed by an increase at 6 km/km^2. The estimation results suggest that the impact of road density on stop-to-stop CB ridership has multiple dimensions. While increased road density has been found to significantly increase public transit demand (Ewing and Cervero, 2010), our results show different interrelationships for CB services. One possible explanation is that areas with poor road connectivity tend to generate more potential market demand for CB services, as they provide more accessible services compared to public transit. Similarly, Wang et al. (2023b) found that urban areas with poorer street connectivity were conducive to CB use in Dalian.

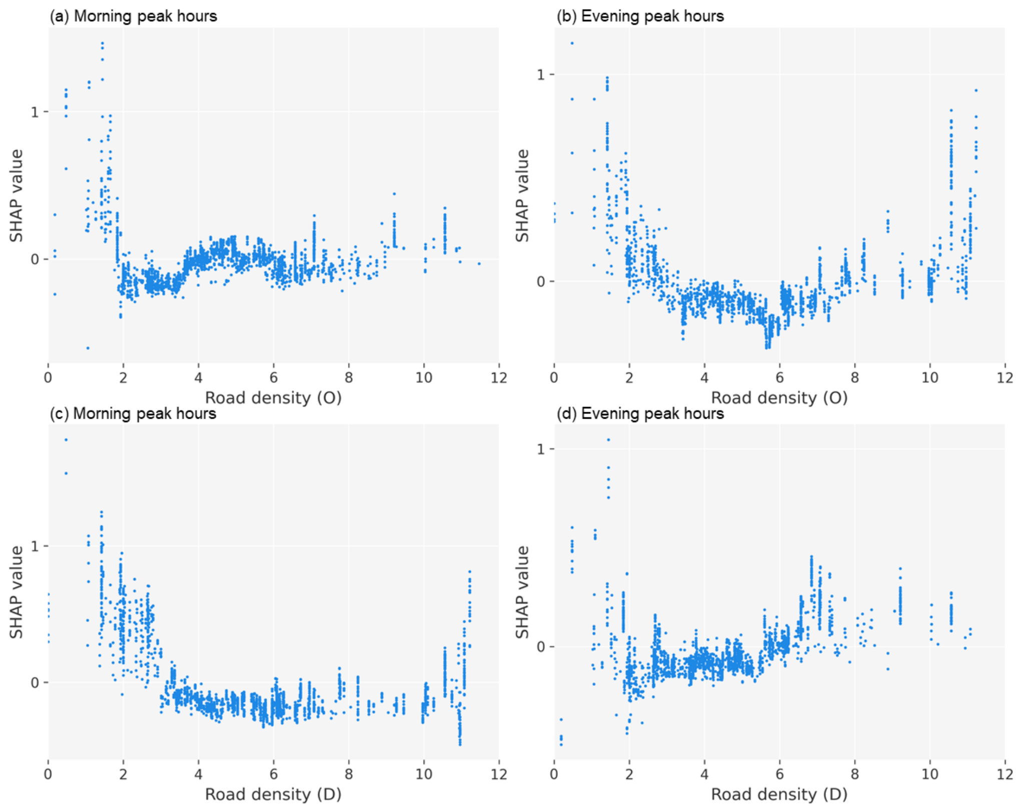


Fig. 11. Threshold effects of road density on CB ridership.

Fig.12 illustrates the nonlinear associations between job accessibility and stop-to-stop CB ridership. The plots show that the effects of job accessibility on stop-to-stop CB ridership on the origin and destination sides fluctuate. For example, Fig.11(c) indicates that job accessibility on the destination side during morning rush hours has a negligible effect on CB ridership when it ranges from 0 to 100. Stop-to-stop ridership then increases significantly when job accessibility is between 100 and 200 and becomes dispersedly distributed after 200. These fluctuating effects contribute to the low ranking (ranks 11–16) for job accessibility in forecasting stop-to-stop CB ridership.

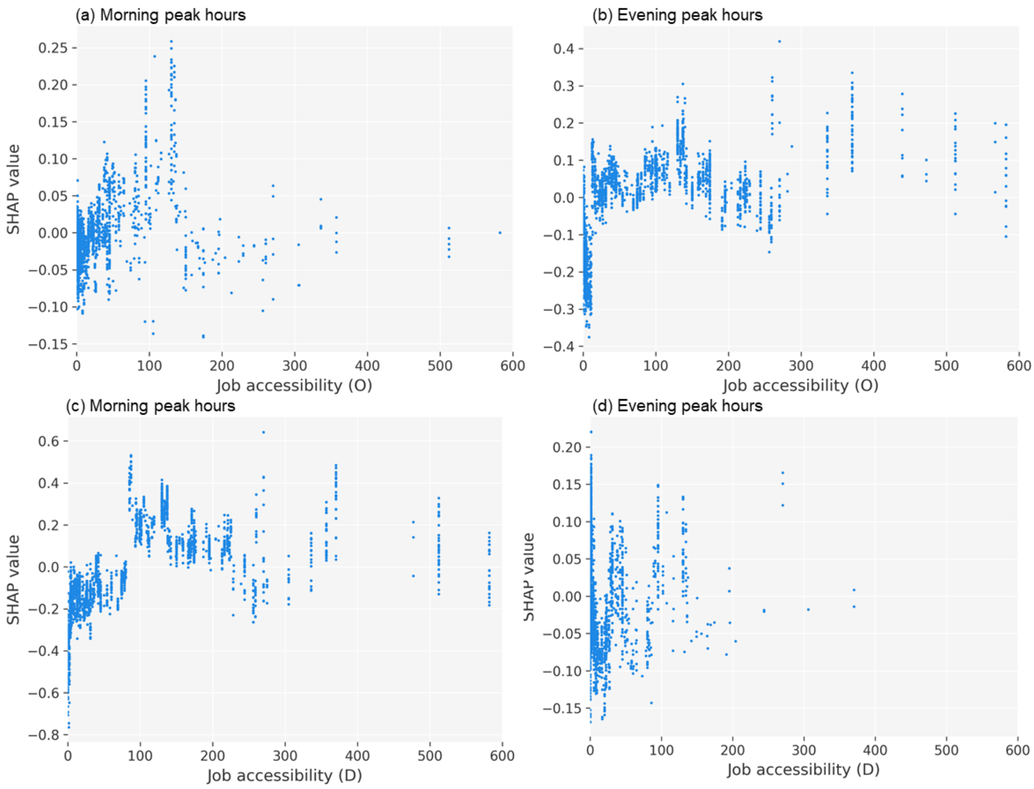


Fig. 12. Threshold effects of job accessibility on CB ridership

Fig.13 displays the effects of amenity accessibility on stop-to-stop CB ridership. Compared to job accessibility, the plots show a more stable distribution. Fig. 13(a) and 13(b) display comparable patterns, indicating that amenity accessibility has an insignificant effect on CB ridership on the origin side during morning rush hours and on the destination side during evening peak hours. Fig. 13(b) and 13(c) also exhibit similar nonlinear patterns, with a gradual increase followed by a peak and then a subsequent decline, with the turning point occurring at around 600. As a result, the amenity accessibility index has a threshold effect on stop-to-stop CB ridership. A positive correlation with stop-to-stop ridership is only present when the index falls within the range of 200 to 600, as demonstrated in the results from the multiplicative model presented in Table 3.

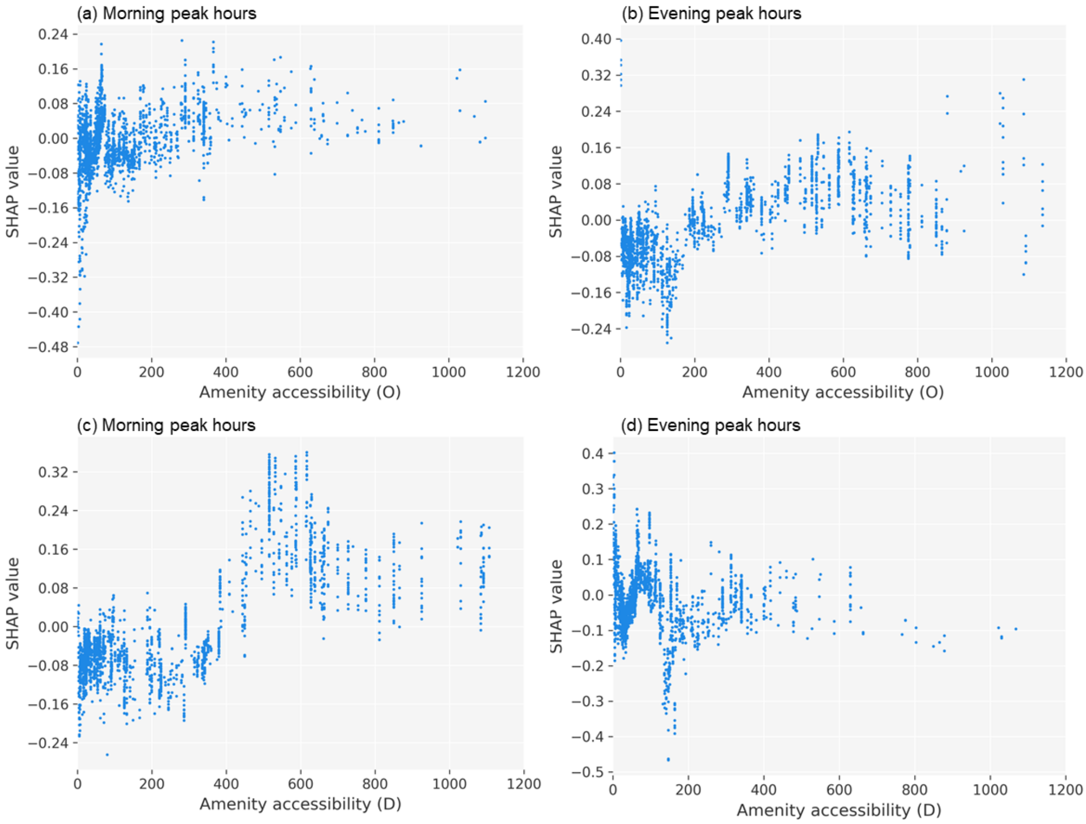


Fig. 13. Threshold effects of amenity accessibility on CB ridership.

Fig. 14 displays the SHAP plots for distance to the city center. Generally, Fig. 14(a) and 14(d) suggest that the distance to the city center on the origin side has a positive impact on stop-to-stop CB ridership during morning rush hours, while the distance to the city center on the destination side is positively associated with stop-to-stop CB ridership during the evening rush hours. Conversely, Fig. 14(b) and 14(c) imply that a greater distance to the city center on the destination side reduces stop-to-stop CB ridership during morning rush hours, whereas the distance to the city center on the origin side is negatively related to stop-to-stop CB ridership during the evening peak hours. These results align with the main directions of CB services as an alternative to commuting transit mode, with the main directions of CB service during the two peak periods consisting of inbound commuting trips moving into the city center during morning peak hours and outbound trips moving away from the city center during evening peak hours.

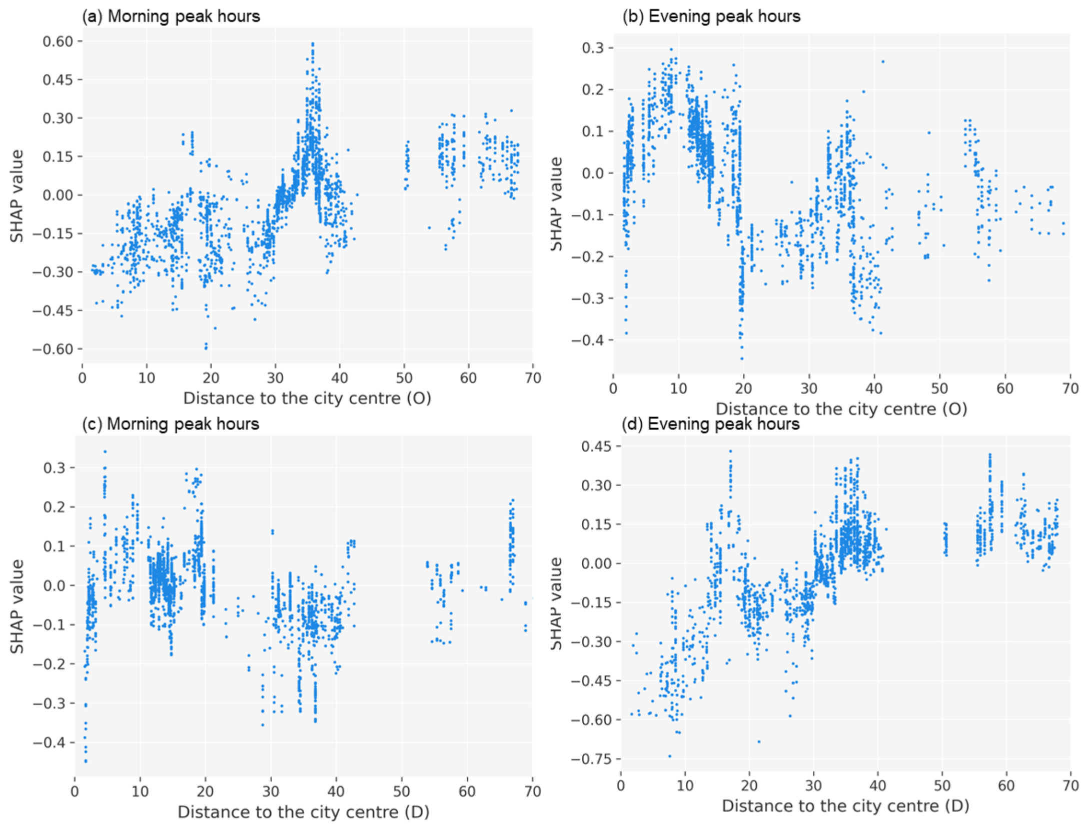


Fig. 14. Threshold effects of distance to the city center on CB ridership.

Fig. 15 illustrates the effects of distance to the nearest metro station on stop-to-stop CB ridership. The four plots display similar nonlinear patterns, with a gradual decrease followed by a plateau once the distance to the nearest metro station exceeds 0.4 km. One possible explanation for this observation is that within the metro catchment area, the metro is a competitive mode of transportation with CBs, and given the higher level of service frequency and reachability, the metro would be the preferred option for travelers. Furthermore, based on the vertical axis ranges, the effects of distance to the nearest metro station at the origin during morning peak hours and at the destination during evening peak hours are smaller than those at the destination during morning peak hours and the origin during evening peak hours. This suggests that the effects of distance to the metro station from the workplace are more critical to CB ridership than the effects of distance from the residence.

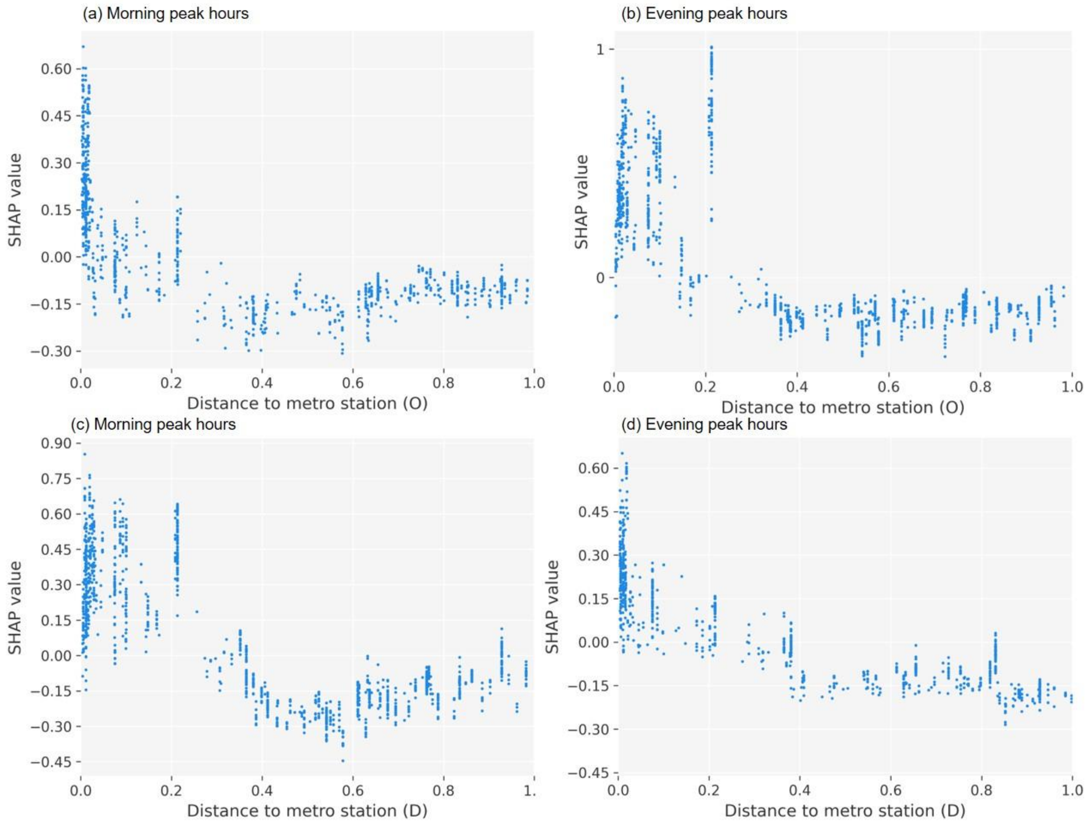


Fig. 15. Threshold effects of distance to metro station on CB ridership.

Fig. 16 displays the relationship between stop-to-stop CB ridership and the number of bus stops on the origin and destination sides during rush hours. The number of bus stops has a positive influence on stop-to-stop ridership when it falls within the range of 0 to 10 on the origin side during morning rush hours. However, Fig. 16(b) and 16(c) demonstrate that the correlation between the number of bus stops is negative on the destination side during morning rush hours and on the origin side during evening rush hours when the number of bus stops increases from 0 to 15. These findings are consistent with the research of Wang et al. (2023a), which suggests that CBs supplement

PT in areas with limited coverage by conventional bus transit, and that the provision of PT services near CB stops has no negative impact on CB demand. CB service providers should make use of existing PT provisions, such as bus stops, to establish their CB pick-up or drop-off stops.

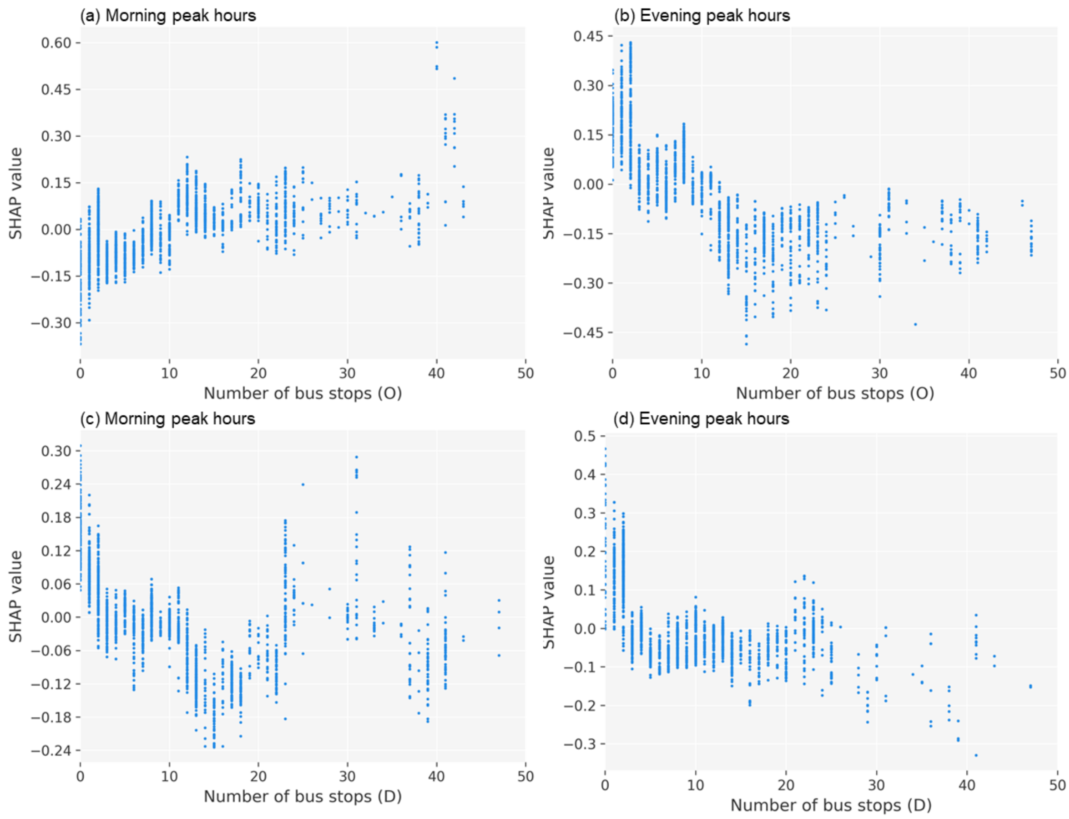


Fig. 16. Threshold effects of distance to transit on CB ridership.

5. Discussion

This study contributes to the enhancement of customized service development and generated several interesting and novel findings. Firstly, in the initial exploratory analysis, regular and recurrent travel patterns were observed for CB services, which challenges the common perception that CB services are more likely to exhibit irregular and dynamic demand patterns with high spatial demand disparity (Petit and Ouyang, 2022). Substantial literature studied routing and scheduling modification problems in

CB service with real-time requests (Huang et al., 2020). These researchers argue that due to the inherent complexity arising from the dynamic decision-making process, the ad hoc on-demand CB transit system design problem should be tackled using sophisticated simulation-based approaches and analytical models. Nevertheless, our findings indicate that this issue might be overestimated, numerous regularities are embedded in this complex and dynamic system, which are also worth further investigation. Moreover, the spatial distribution of O-D flows indicates that CB services not only serve long-distance commuting travel but also train station/airport transfer travel. This finding is consistent with Davison et al. (2012), highlighting the great potential and popularity of CB in such a niche market.

Secondly, we discovered that numerous built environment factors significantly impact the use of CB. For instance, our findings indicate that the proximity of metro stations at the destination during morning rush hours and at the origin during evening hours ranks second in importance among all variables in the XGBoost models. This suggests that CB services may have a competitive or complementary relationship with metro services, rather than with conventional bus services. In Shanghai, the average travel distance for conventional buses is approximately 9.7 km, while metro transit covers an average distance of about 18.3 km. This is strikingly similar to the average travel distance of CB in our study, which is 20.9 km. Considering the travel distances, metro and CB services may compete for the same market. However, since metro services are limited or non-existent in suburban and rural areas, CB services can be promoted as a transportation alternative when expanding the metro system is not feasible or economically viable. This highlights the complementary role of CB services in the public transportation system (Davison et al., 2012). In Shanghai, improving public transport accessibility in suburban and rural areas presents a significant challenge for the government. Public transport agencies often struggle with high operating costs in areas with lower demand density but larger service coverage. In such situations, implementing more flexible door-to-door CB services could better accommodate users' travel behavior, reduce lengthy access distances to and from stations, and most importantly, save money while maintaining public accessibility for all groups. Based on the findings of our empirical study, a primary local policy recommendation is to advance 'institutional reform' by integrating transport provision across various sectors (Davison et al., 2012). Concurrently, the government can offer subsidies, funding, or permits to support CB service provision in the commuting corridors between rural and urban areas. This approach would not only supply essential public transport services to those without access to personal vehicles but also contribute to the development of sustainable transportation systems.

Moreover, we found that the impact of certain built environment factors on CB ridership may not align with their effects on traditional bus ridership. For example, existing literature demonstrates a significantly positive correlation between population density and conventional bus usage rates (Chiou et al., 2015), as well as metro ridership (Shao et al., 2020). Nevertheless, our study shows that population density has a negligible and insignificant relationship with CB ridership, irrespective of stop sides and time of day. This observation also applies to other built environment factors such as land use mix and destination accessibility. These findings suggest that the behavioral mechanisms dictating how the built environment influences CB travel behavior may not be the same as those for public transportation (PT) travel. In other words, as a specialized bus service catering to more personalized travel demands, a more effective strategy for CB providers to attract more passengers would be to concentrate on coordinating with specific and loyal users, rather than merely allocating service stops in densely populated areas.

Thirdly, a series of SHAP plots demonstrated that most explanatory variables exhibited nonlinear correlations with CB ridership. This nonlinearity suggests that the effects of explanatory variables differ at various scales. For instance, we found that CB ridership is most likely to reach its peak when the trip distance is around 30 km. Beyond 30 km, an increase in trip distance does not necessarily lead to an increase in CB ridership. This observation adds nuance to the existing literature, which posits that CBs are more attractive to passengers with long-distance trips (Liu and Ceder, 2015; Wang et al., 2021). Our results further delineate the effective range of that distance, which can inform CB providers' route design. Moreover, our findings can aid CB service providers in determining where to deploy capacity and stop to attract more passengers. We discovered that the distance to the nearest metro station, when less than 400 m, is negatively associated with stop-to-stop CB ridership. Consequently, CB stops should steer clear of areas with immediate access to metro transit services. CB providers often face the challenge of deciding where to deploy stops to maximize passenger attraction (Wang et al., 2023a). Our findings suggest that they should consider route design and stop allocation concurrently. To attract riders efficiently, they should avoid short-distance service routes, as these can be better served by conventional bus systems with higher schedule frequency and shorter time and lower fare. Additionally, they need to avoid areas with good accessibility to public transport (especially metro services) at destinations during morning rush hours and at origins during evening rush hours.

Fourthly, regarding methodology, this study reaffirms the benefits of machine learning algorithms in predicting and providing a more in-depth description of relationships. Furthermore, this study compared the results of XGBoost with those of traditional statistical regression models and found both similarities and differences. In terms of similarities, variables that demonstrate significant correlations with the dependent variable in the multiplicative model tend to have a strong impact in the XGBoost models. However, we also found situations where variables exhibit a significant correlation in the multiplicative model, but their contribution is low in the XGBoost model. We hypothesize that this divergence might stem from differences in the underlying mechanisms. In the XGBoost model, the relative importance of variables is measured by the mean absolute Shapley values of all observations. The fluctuations and signs of SHAP values do not affect the relative importance of variables, but they significantly influence the coefficients and significance of variables in the multiplicative model, which assumes a linear relationship between dependent and independent variables. Similar findings and discussions can be supported by Yang et al. (2021). However, this does not imply that traditional linear regression models are useless. On the contrary, using machine learning models in conjunction with traditional linear regression models can enhance the robustness of the analysis results by validating the results between the two models. This is particularly relevant because the literature has raised concerns about the possibility of machine learning algorithms producing counterintuitive outcomes due to sparse data, outliers, and/or unbalanced spatial distributions of independent variables, especially when using partial dependence plots as an interpretive method (Gan et al., 2020). By comparing the results of both the machine learning model and the conventional statistic model, we can take advantage of the strengths of each and provide a form of model validation or robustness check.

It is worth noting that the dependent variable in both the multiplicative regression models and the XGBoost model is the total subscriptions for each origin-destination (OD) pair from May 2017 to October 2019. It does not represent weekly, monthly, or seasonal subscriptions of all OD pairs during the study period. Consequently, the models in this study do not have the issue of serial correlation, since the dependent variable was not examined over a predetermined sequence of time points. However, for modeling CB ridership over a specific period, where serial correlation could be significant (Wang and Noland, 2021), machine learning models like XGBoost can still provide accurate predictions. Despite not explicitly modeling the correlation, the flexibility and ability of the XGBoost model to discern complex, non-linear patterns in high-dimensional data make it a solid choice for prediction tasks (Praveen et al., 2020). This includes scenarios where serial correlation could be crucial, such as forecasting CB ridership over a certain period. In future work, it could be intriguing to conduct a time-serial analysis of the data, for instance, calculating the serial correlation and incorporating that into a regression model.

Lastly, while our study utilized Shanghai as a case study, many conclusions drawn and findings identified are transferable to other urban contexts. Cities around the world share common urban development patterns and transportation challenges (Banister, 2005). Our study used the 5D built environment variables and travel impedances, which are relevant in any urban context (Ewing and Cervero, 2010), insights gained from Shanghai and the built environment study can apply to cities with similar characteristics. Nevertheless, it is worth noting that the non-linear and threshold effects of the built environment can vary depending on the context of different cities. For instance, Shanghai is a mega-city with a population of 24.9 million and covers a highly urbanized area of 6340.5 km^2. The threshold of travel distance (30 km) that we found to maximize CB ridership in this study cannot be simply applied to other cities. CB providers would need to conduct a case-specific study to identify an appropriate threshold for the local CB service design. Besides this consideration, the research design, methodological approaches, and empirical findings used in this study remain transferable.

6. Conclusions

Using Shanghai as a case study, this research investigated the spatiotemporal patterns and key built environment factors influencing CB ridership on the origin and destination sides and across different rush hours of the day. By utilizing long-term subscription data and open-source big data, we applied the XGBoost and SHAP models to investigate the nonlinear effects of travel impedance and the built environment on CB ridership. The selection of this research design was driven by the existing research gap and the aspiration to examine the relationships between the built environment and CBs ridership more comprehensively. This analysis is grounded in the belief that the relationship is non-monotonic and varies with the side of stop (origin and destination), as opposed to the continuous and homogeneous nature often assumed by the literature.

Our research underscores that the built environment indeed exerts significant impacts on CB transit ridership. However, the magnitude and nature of these impacts vary depending on the side of the stop, and substantially from those observed in regular transit systems. These variations and differences are primarily driven by the unique service characteristics inherent in CB transit, including its flexibility, on-demand booking capability, and its crucial role in meeting users' travel needs. This is especially true for individuals living in suburban and rural areas where the availability and frequency of public transit services are considerably lower compared to central urban areas. The incorporation of CB transit systems effectively addresses the underserved transit needs of suburban and rural populations, thus contributing to more equitable and inclusive urban mobility strategies. These insights provide essential guidance not only for CB transit providers in identifying suitable market niches but also for policymakers. They should consider adopting public transport strategies to integrate CB transit as a pivotal complementary component in the urban mobility system. Such integration would assist in accomplishing the sustainable development goals outlined in the New Urban Agenda (UN-HABITAT, 2016).

Several questions merit further exploration. Future research should investigate the factors that influence CB ridership during different times of the day, such as peak and off-peak periods, as well as weekdays versus weekends. CB services on different days may attract unique user groups with distinct travel purposes and serve different spatial areas. More importantly, the dataset we utilized did not include the sociodemographic profiles of CB users due to the data availability. These factors can have a substantial impact on CB ridership, particularly when it comes to understanding the attributes and choice preferences of highly loyal users, and the key factors that foster passenger loyalty towards CB services. Furthermore, the longitudinal nature of subscription data presents a potential opportunity to measure passenger loyalty, providing valuable insights for improving service design and customer engagement. Additional research in this area would be both intriguing and valuable. Lastly, it is worth noting that Shanghai is a megacity with a complex and distinctive transport network and public transportation facilities. 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.

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.

Banister, D., 2005. Unsustainable transport: city transport in the new century. Taylor & Francis.

Bautz, J.A., 1975. Subscription service in the United States. Transportation. 4, 387–402.

Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: Density, diversity, and design. Transp. Res. Part D: Transp. Environ. 2, 199–219.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In Proceedings of the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 13–1, 785–794.

Chiou, Y.C., Jou, R.C., Yang, C.H., 2015. Factors affecting public transportation usage rate: Geographically weighted regression. Transp. Res. Part A: Policy Pract. 78, 161–177.

Choi, J., Lee, Y.J., Kim, T., Sohn, K., 2012. An analysis of Metro ridership at the station-to-station level in Seoul. Transportation. 39, 705–722.

Daganzo, C.F., 2010. Structure of competitive transit networks. Transp. Res. Part B: Methodol. 44, 434–446.

Davison, L., Enoch, M., Ryley, T., Quddus, M., Wang, C., 2012. Identifying potential market niches for Demand Responsive Transport. Res. Transp. Bus. Manag. 3, 50-61.

Davison, L., Enoch, M., Ryley, T., Quddus, M., Wang, C., 2014. A survey of demand responsive transport in Great Britain. Transp. Policy. 31, 47-54.

De Dios Ortuzar, J., Willumsen, L.G., 2011. Modelling Transport. John Wiley & Sons.

Deepa, L., Pinjari, A.R., Nirmale, S.K., Srinivasan, K.K., & Rambha, T., 2022. A direct demand model for bus transit ridership in Bengaluru, India. Transp. Res. Part A: Policy Pract. 163, 126–147.

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. J. Transp. Geogr. 77, 70–78.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 36(1), 27-46.

Ewing, R., Cervero, R., 2010. Travel and the built environment: A meta-analysis. J. Am. Plan. Assoc. 76, 265–294.

Gan, Z., Yang, M., Feng, T., Timmermans, H.J., 2020. Examining the relationship between built environment and metro ridership at station-to-station level. Transp. Res. Part D Transp. Environ. 82, 102332.

Garg, A., Tai, K., 2013. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. Int. J. Model. Identif. Control. 18(4), 295-312.

Guo, R., Guan, W., Huang, A., Zhang, W., 2019. Exploring potential travel demand of customized bus using smartcard data. In Proceedings of the 2019 IEEE Intelligent

Transportation Systems Conference (ITSC). 27–3, 2645–2650.

Huang, D., Gu, Y., Wang, S., Liu, Z., Zhang, W., 2020. A two-phase optimization model for the demand-responsive customized bus network design. Transp. Res. Part C Emerg. Technol. 111, 1–21.

Huo, Y., Qiu, Z., Xin, Y., Dai, S., Pan, R., Wen, Y., 2021. Exploring the influencing factors on demand-responsive customized bus ridership using multi-source data. In Proceedings of the 2021 6th International Conference on Transportation Information and Safety (ICTIS), 22–24, 293–299.

Jain, S., Ronald, N., Thompson, R., Winter, S., 2017. Predicting susceptibility to use demand responsive transport using demographic and trip characteristics of the population. Travel Behav. Soc. 6, 44-56.

Kirby, R.F., Bhatt, K.U., 1974. Guidelines on the Operation of Subscription Bus Services, National Technical Information Service: Springfield.

Li, D., Ye, X., Ma, J., 2019. Empirical analysis of factors influencing potential demand of customized buses in Shanghai, China. J. Urban Plan. Dev. 145, 05019006.

Li, J., Chen, X., Li, X., Guo, X., 2013. Evaluation of public transportation operation based on data envelopment analysis. Procedia-Soc. Behav. Sci. 96, 148–155.

Li, J., Lv, Y., Ma, J., Ouyang, Q., 2018. Methodology for extracting potential customized bus routes based on bus smart card data. Energies. 11, 2224.

Li, Z., 2022. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Comput. Environ. Urban Syst. 96, 101845.

Li, Z., Liu, K., Wang, J., 2021. Exploring observed and unobserved preference heterogeneity in choice behavior of demand responsive customized bus service. J. Urban Plan. Dev. 147, 05020038.

Litman, T., 2015. Evaluating public transit benefits and costs. Victoria, BC, Canada: Victoria Transport Policy Institute.

Liu, J., Wang, B., Xiao, L., 2021. Non-linear associations between built environment and active travel for working and shopping: An extreme gradient boosting approach. J. Transp. Geogr. 92, 103034.

Liu, T., Ceder, A.A., 2015. Analysis of a new public-transport-service concept: Customized bus in China. Transp. Policy. 39, 63–76.

Liu, T., Ceder, A.A., Bologna, R., Cabantous, B., 2016. Commuting by customized bus: A comparative analysis with private car and conventional public transport in two cities. J. Public Transp. 19, 55–74.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), 4–9.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell., 2(1), 56-67.

Lyu, Y., Chow, C.-Y., Lee, V.C., Li, Y., Zeng, J., 2016. T2CBS: Mining taxi trajectories for customized bus systems. In Proceedings of the 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 10–14, 441–446.

Lyu, Y., Chow, C.-Y., Lee, V.C., Ng, J.K., Li, Y., Zeng, J., 2019. CB-Planner: A bus line planning framework for customized bus systems. Transp. Res. Part C: Emerg. Technol. 101, 233–253.

Nelson, J.D., Phonphitakchai, T., 2012. An evaluation of the user characteristics of an open access DRT service. Res. Transp. Econ. 34, 54–65.

Oakil, A.T.M., Nijland, L., Dijst, M., 2016. Rush hour commuting in the Netherlands: Gender-specific household activities and personal attitudes towards responsibility sharing. Travel Behav. Soc. 4, 79–87.

Petit, A., Ouyang, Y., 2022. Design of heterogeneous flexible-route public transportation networks under low demand. Transp. Res. Part C: Emerg. Technol. 138, 103612.

Potts, J.F., Marshall, M.A., Crockett, E.C., Washington, J., 2010. A guide for planning and operating flexible public transportation services. TCRP Rep. 140.

Praveen, B., Talukdar, S., Mahato, S., Mondal, J., Sharma, P., Islam, A.R.M., Rahman, A., 2020. Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. Sci. Rep. 10(1), 1-21.

Redman, L., Friman, M., Gärling, T., Hartig, T., 2013. Quality attributes of public transport that attract car users: A research review. Transp. Policy. 25, 119–127.

Sanaullah, I., Alsaleh, N., Djavadian, S., Farooq, B., 2021. Spatio-temporal analysis of on-demand transit: A case study of Belleville, Canada. Transp. Res. Part A: Policy Pract. 145, 284–301.

Shaheen, S., Sperling, D., Wagner, C., 1998. Carsharing in Europe and North American:

Past, Present, and Future, University of California Transportation Center.

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. J. Transp. Geogr. 89, 102878.

Shapley, L.S., 1997. A value for n-person games. Class. Game Theory. 69, 295.

Sohn, K., Shim, H., 2010. Factors generating boardings at metro stations in the Seoul metropolitan area. Cities. 27, 358–368.

UN-HABITAT, 2016. The New Urban Agenda. http://habitat3.org/wp-content/uploads/NUA-English.pdf (accessed 21.05.2023).

Wagner, C., Shaheen, S., 1998. Car Sharing and Mobility Management: Facing new challenges with technology and innovative business planning. J. World Transp. Policy Pract. 4, 39–42.

Wang, C., Quddus, M., Enoch, M., Ryley, T., Davison, L., 2014. Multilevel modelling of Demand Responsive Transport (DRT) trips in Greater Manchester based on area-wide socio-economic data. Transportation. 41, 589–610.

Wang, H., Noland, R.B., 2021. Bikeshare and subway ridership changes during the COVID-19 pandemic in New York City. Transp. Policy. 106, 262–270.

Wang, J., Cao, X.J., Liu, K., Wang, D., 2023a. Exploring the nonlinear effects of built environment characteristics on customized bus service. Transp. Res. Part D: Transp. Environ. 114, 103523.

Wang, J., Liu, K., Yamamoto, T., Wang, D., Lu, G., 2023b. Built environment as a precondition for demand-responsive transit (DRT) system survival: Evidence from an empirical study. Travel Behav. Soc. 30, 271–280.

Wang, J., Yamamoto, T., Liu, K., 2019. Role of customized bus services in the transportation system: Insight from actual performance. J. Adv. Transp. 2019, 6171532.

Wang, J., Yamamoto, T., Liu, K., 2020. Key determinants and heterogeneous frailties in passenger loyalty toward customized buses: An empirical investigation of the subscription termination hazard of users. Transp. Res. Part C: Emerg. Technol. 115, 102636.

Wang, J., Yamamoto, T., Liu, K., 2021. Spatial dependence and spillover effects in customized bus demand: Empirical evidence using spatial dynamic panel models. Transp. Policy. 105, 166–180.

Wooldridge, J.M., 2015. Introductory econometrics: A modern approach. Cengage learning.

Wu, L., Gu, W., Fan, W., Cassidy, M.J., 2020. Optimal design of transit networks fed by shared bikes. Transp. Res. Part B: Methodol., 131, 63–83.

Yan, X., Liu, X., Zhao, X., 2020. Using machine learning for direct demand modeling of ridesourcing services in Chicago. J. Transp. Geogr., 83, 102661.

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. Accid. Anal. Prev., 158, 106153.

Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing. 415, 295–316.

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

Zhang, X., Yan, X., Zhou, Z., Xu, Y., Zhao, X., 2022. Examining spatial heterogeneity of ridesourcing demand determinants with explainable machine learning. arXiv Prepr., arXiv:2209.07980.

Zhao, J., Deng, W., Song, Y., Zhu, Y., 2014. Analysis of Metro ridership at station level and station-to-station level in Nanjing: An approach based on direct demand models. Transportation. 41, 133–155.

Ziakopoulos, A., Kontaxi, A., Yannis, G., 2023. Analysis of mobile phone use engagement during naturalistic driving through explainable imbalanced machine learning. Accid. Anal. Prev. 181, 106936.

Declaration of Competing Interest

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

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). We would like to thank Yidong for providing the data. The authors are grateful to the editor, guest editor, and anonymous referees for their guidance and constructive comments that helped improve this manuscript.