Journal of Transport Geography 110 (2023) 103604

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Data-driven interpretation on interactive and nonlinear effects of the correlated built environment on shared mobility

Kun Gao [a], Ying Yang [a,b,*], Jorge Gil [a], Xiaobo Qu [a,c]

[a] Department of Architecture and Civil Engineering, Chalmers University of Technology, Gothenburg SE-412 96, Sweden
[b] School of Behavioural and Health Sciences, Australian Catholic University, Australia
[c] State Key Laboratory of Automotive Safety and Energy, School of Vehicle and Mobility, Tsinghua University, Beijing, China

ARTICLE INFO

Keywords:
Interpretable machine learning
Built environment factors
Shared mobility systems
Interactive effects

ABSTRACT

Understanding the usage demand of shared mobility systems in different areas of a city and its determinants is crucial for planning, operation and management of the systems. This study leverages an unbiased data-driven approach called accumulated effect analysis for examining the complex (nonlinear and interactive) effects of correlated built environment factors on the usage of shared mobility. Special research emphasis is given to unraveling the complex effects using an unbiased and data-driven approach that can overcome the impacts of correlations among built environment factors. Based on empirical analysis of synthetic data and a field dataset about dockless bike sharing systems (DLBS), results demonstrate that the method of partial dependency analysis prevalent in the relevant literature, will result in biases when investigating the effects of correlated built environment factors. In comparison, accumulated local effect analysis can appropriately interpret the effects of correlated built environment factors. The main effects of many built environment factors on the usage of DLBS present nonlinear and threshold patterns, quantitively revealed by accumulated local analysis. The approach can reveal complex interaction effects between different built environment factors (e.g., commercial service and education facility, and metro station coverage and living facility) on the usage of DLBS as well. The interactions among two built environment factors could even change with the values of the factors rather than invariant. The outcomes offer a new approach for revealing complex influences of different built environment factors with correlations as well as in-depth empirical understandings regarding the usage of DLBS.

1. Introduction

Transport takes up approximately a quarter of global greenhouse gas emissions (Edelenbosch et al., 2017; McCollum et al., 2018; Chi et al., 2022), and addressing this source plays a vital role in realizing net-zero emissions. One recognized measure for reducing emissions in the transport sector is facilitating sustainable mobility. Shared mobility systems, such as bike-sharing, e-scooter sharing, ride-hailing, and car-sharing systems, have been rapidly expanded in recent years to promote sustainable travel behavior (Attard, 2022; Gao et al., 2021c; Li et al., 2021a, 2021b; Giuffrida et al., 2023). Besides environmental benefits, emerging shared mobility systems could contribute to increased transport efficiency, reduced user expense, and urban accessibility (Becker et al., 2020; Gao et al., 2021b; Laporte et al., 2018; Li et al., 2022; Arias-Molinares et al., 2021; Roman et al., 2021). One of the core measures to ensure the operational efficiency of shared mobility systems is accurate usage demand estimation and prediction in spatio-temporal dimensions (Ortúzar, J.d.D., 2021), which is the most fundamental component for planning dispatching, and rebalancing to realize supply-demand matchup. Therefore, it is essential for urban planners and shared mobility operators to precisely understand the travel demand in different areas of the city and their determinants for operational efficiency and scientific planning (Coretti Sanchez et al., 2022). It is noted that the built environment factors play important roles in affecting travel behavior, usage patterns, and ridership of different transport modes (Cheng et al., 2022a; Ding et al., 2019a; Ding et al., 2021; Hasnine et al., 2020; Hu et al., 2021). Meanwhile, investigating the impacts of various built environment factors on travel demand and usage patterns has been one of the focuses of urban and transport planning and management.

The necessity to apprehend the impacts of diverse built environment factors on demand motivates many researchers to develop modeling

* Corresponding author.
E-mail address: Ying.Yang@acu.edu.au (Y. Yang).

https://doi.org/10.1016/j.jtrangeo.2023.103604
Received 12 January 2023; Received in revised form 6 April 2023; Accepted 7 May 2023
Available online 23 May 2023

approaches and conduct empirical analysis based on multiple-source data. To model and interrogate the influences of various factors on demand for shared mobility systems quantitatively, some previous studies employed conventional correlation analysis or regression models with linear relationship assumptions (e.g., Ordinary Least Squares regression) based on different datasets (e.g., Etminani-Ghasrodashti and Hamidi, 2019; McKenzie, 2019; Ma et al., 2020; Torrisi et al., 2021; Yang et al., 2020; Li et al., 2021a, 2021b). For example, Yang et al. (2020) applied a semi-parametric geographically weighted regression to investigate the influences of different built environment factors on the ridership of different users. Li et al. (2021a, 2021b) revealed the spatially varying impacts of built environment factors on transit ridership using geographically weighted regression with linear model specifications based on a dataset in Guangzhou. Liu et al. (2022a, 2022b) used a Global Linear Regression model with consideration of temporal heterogeneity to understand the divergent effects of land use on bike-sharing usage at different hours of the day and different days of the week. However, it is argued that linear assumptions are very strong hypotheses and may lead to seriously biased results on account that the effects of many factors may present nonlinearity and the existence of threshold effects (Du et al., 2022; Huang et al., 2021; Tu et al., 2021; Wang et al., 2022; Liu et al., 2022a, 2022b). Hence, some studies devote efforts to modeling and revealing nonlinear effects or so-called threshold effects. Presuming the nonlinear effect of a built environment factor (e.g., exponential or power relations) based on expert judgment is one option but not preferred by researchers as it is arbitrary to determine the nonlinear patterns in advance instead of extracting them from data. For instance, Liu et al., 2022a, 2022b analyzed the factors influencing demand for ride-hailing using a quantile regression where the non-stationary effects of each factor were considered as well. The local estimates at five quantiles of each factor were obtained and found to be significantly different, indicating nonlinear effects.

Therefore, some studies utilized semi-parametric statistical approaches such as generalized additive mixed models considering autocorrelation and heteroscedasticity (Cheng et al., 2022a; Hu et al., 2021; Lin et al., 2020) and quantile regression approaches considering heteroscedasticity (Li et al., 2020; Cheng et al., 2022b) for analysis. These semi-parametric methods extend the conventional models to consider potentially nonlinear effects of different factors by embedding polynomial spline estimators or quantile regression instead of global regressions and present quite good performances compared to conventional methods with linear assumptions. For instance, Lin et al. (2020) used historical trip data from docked-station bike-sharing systems in Beijing of 2016 and a log-linear regression to examine the effects of different factors on station-level usage demand, including land use and transport infrastructure. Cheng et al. (2022b) utilized a quantile regression method to investigate the nonlinear effects of factors on the transfer ridership between bike-sharing and urban transit systems and indicated pretty good performances of using the method. Cheng et al. (2023) employed a generalized additive mixed model to explore the influences of built environment factors on demand for integration station-based bike-share systems and free-floating bike-share systems with public transit using data from Nanjing. Their results reported the plateau effect of several factors, which were interpreted well by the used approaches. Although semi-parametric methods can model nonlinearity to some extent, they still need presumed model formulation and equations to represent nonlinear effects. Therefore, it still has the risk of getting biased results due to arbitrary model assumptions.

Instead, recent studies have started to leverage new data-driven methods based on machine learning to investigate the nonlinear effects of different factors on demand for different transport modes (Chen and Ye, 2021; Ding et al., 2019b; Ding et al., 2018; Pérez-Fernández and García-Palomares, 2021; Tu et al., 2021; Wang et al., 2022; Xu et al., 2021; Wagner et al., 2022). Just to name a few, Ding et al. (2018) adopted gradient-boosting decision trees to test the divergent effects of built environment factors on driving distance on weekdays and weekends. Noticeable nonlinearity in the effects of different features was observed, which violated the linear assumptions. Ding et al. (2019b) employed gradient-boosting decision trees and Partial Dependence Analysis (PDA) to unravel the nonlinear effects of built environment factors on the ridership of Metrorail in Washington. Li et al. (2021a, 2021b) took advantage of the extreme gradient boosting model to investigate the non-linear relations of the built environment with the demand for using active mobility for working and shopping trips. Xu et al. (2021) used a random forest model to depict the association of built environment factors with ride splitting adoption rate and identified nonlinear effects of the population structure, household income, and walking environment. Wang et al. (2022) applied the random forest for modeling the influences of various factors on trip volume and utilized partial dependence analysis to reveal the nonlinear effects of different factors. Zhuang et al. (2022) utilized gradient-boosting decision trees and interpretation methods, including relative importance and PDA, for analyzing the nonlinear effects of demand for bike sharing at the street level. The impacts of several factors such as PM2.5 emissions, street greenness view index and sky view index presented nonlinear and effects on the demand of using bike sharing at the streets. Based on data from Zhongshan of China, Shao et al. (2022) scrutinized the nonlinear and interaction effects of built environment on car ownership using gradient boosting decision trees. One noteworthy finding was the threshold (nonlinear) relationships between built environment and car ownership. The typical procedure is using supervised machine learning algorithms to fit the relationships between features and demand for a specific shared mobility system (e.g., bike sharing, ride-hailing, or e-scooter sharing) and use feature importance or PDA to investigate the nonlinear effects of built environment factors. This framework generally presents superiority in contrast to conventional models with specific formulations.

Notwithstanding, the relevant existing literature regarding interrogating the influences of various built environment factors on demand for shared mobility has several shortcomings to be tackled. Firstly, the prevalently used method PDA has a strong assumption that the explanatory variables are perfectly independent when interpreting the relationship of an explanatory variable with the dependent variable (e.g., demand for bike sharing in a specific area). If the prerequisite is not satisfied, the results from PDA will be problematic due to the extrapolation issue of machine learning methods (Molnar, 2020), which has never been explicitly addressed in the relevant literature. However, different built environment factors naturally have some degree of correlation as different aspects of the built environment interact with each other (Brownson et al., 2009; Cervero et al., 2009; Ewing and Cervero, 2010). For instance, residential areas are generally surrounded by commercial services such as restaurants and grocery stores in Chinese cities. As a consequence, using PDA for interrogating the complex effects of the correlated built environment factors on demand for shared mobility has a high risk of obtaining biased and even wrong results. Secondly, existing studies hardly decipher the interactive effects of different built environment factors on demand for shared mobility and mostly presumed independence among various built environment factors. The arbitrary hypotheses may result in imprecise results about the effects of the built environment factors if interaction effects do exist. The effects of built environment factors on mobility (Gao et al., 2021a; Schüle and Bolte, 2015) may not be independent and could be interactive, which have hardly been quantitatively modeled and investigated using data-driven methods in the contexts of shared mobility systems. For instance, the influences of land use characteristics (e.g., commercial land use) and transport facilities (e.g., accessibility to public transit) on demand for shared mobility may not be independent but interactive. Meanwhile, such complex (nonlinear and high-dimension) interaction effects are hard to be modeled and investigated by conventional regression approaches.

To improve the abovementioned research gaps, this study leverages an unbiased data-driven approach for examining the complex (nonlinear and interactive) effects of the correlated built environment factors on demand for shared mobility. Machine Learning (ML) models based on data-driven mechanisms are utilized to learn the complex relationships between built environment factors and demand for shared mobility. ML models can appropriately tackle the deficiencies of conventional methods by automatically learning the nonlinear and interactive effects of different factors. More importantly, a data-driven interpretation technique that can eliminate the biases due to correlated features is leveraged to explain "black-box" ML and reveal the nonlinear and interactive effects unbiasedly. Special research emphasis is given to unraveling the complex effects using an unbiased and data-driven approach that can overcome the impacts of correlations among built environment factors. Synthetic data are firstly utilized to validate the reliability and unbiased merits of the used method. Afterward, a case study about dockless bike-sharing systems in Shanghai is conducted using the proposed method for empirical analysis. The outcomes offer a novel approach for revealing complex influences of different built environment factors with correlations and in-depth practical understandings. These are useful inputs for rebalancing, planning, and management of shared mobility systems in a cost-effective way.

The remaining sections are structured as follows. Section 2 elaborates on the technical details of the methodology, followed by empirical data and analysis in Section 3. Section 4 presents the results and discussions, and concluding remarks are provided in the last section.

2. Methodology

The methodological framework consists of two stages: 1) a model to quantitative depict the effects of different built environment factors on the demand for shared mobility; 2) an approach to reveal the relationship of each factor with the demand for shared mobility. For modeling the complex impacts of various built environment factors, we make the best of data-driven modeling through supervised machine learning. As a rule of thumb, different supervised machine learning algorithms in different categories are compared to select the best one. The aim is to find an appropriate ML model that can well model the relationships between different built environment factors and the dependent variable, which is the prerequisite for precisely and quantitatively interpreting the effect of each factor. Afterward, data-driven interpretation techniques are used to interpret the trained models and reveal the quantitative effects of various built environment factors on demand for shared mobility systems. Traditional methods, such as Multiple Linear Regression, have presumed model specifications and could be interpreted intuitively and directly. Nonetheless, the presupposed model specifications also confine the capacity of such methods to model complicated nonlinear and interactive effects. Instead, we utilize a data-driven interpretation technique to explain "black-box" ML and decipher the nonlinear and interactive effects in an unbiased way. Especially, differing from commonly used methods such as PDA, our approach can eliminate the adverse impacts and biases due to correlated features.

2.1. Selecting the ML regressor

Modeling the effects of built environment factors on demand for shared mobility is a typical regression problem and can be tackled by various supervised machine learning algorithms. We examine different prevalent ML algorithms for comparisons, including ElasticNet, Support Vector Regression (SVR), Multi-layer Perceptron (MLP), Random Forest (RF), and XGBoost. They are representative models with different mechanisms and architectures. A Multiple Linear Regression is used as the reference model. More technical details about the algorithms are available in Gao et al. (2021a, 2021b, 2021c, 2021d).

ElasticNet is a regularized regression that integrates the penalties from the lasso and ridge techniques by learning from their shortcomings to improve the regularization. ElasticNet has been widely used for regression problems involving high-dimensional data (De Mol et al., 2009). SVR is another popular regression algorithm with robust characteristics for outliers (Awad and Khanna, 2015). SVR tries to construct a hyperplane that can fit the maximum number of points within a specific bandwidth. The aim of training an SVR is not to minimize the exact value of predictive errors but to cover most points in a hyperplane with bandwidth. MLP is an artificial neural network where each perceptron connects to all perceptrons in the next layer (Tang et al., 2015). The transformations between perceptrons are formulated as equations with activation functions and node weights. Key hyperparameters, including activation function, the number of hidden layers, learning rate, and the number of perceptrons, are well-tuned based on grid search methods. Random Forest is a classic tree-based machine-learning model (Breiman, 2001). The merit of RF is that it utilizes a set of trained decision trees to facilitate robustness and generality (Gomes et al., 2017; Zhao et al., 2020). By randomizing samples and features across many trees, RF has a great ability to reduce sensitivity to noisy data and address redundant prediction (Breiman, 2001). Three main hyperparameters are critical for the performance of RF: the number of decision trees, the tree depth, and the feature number in each splitting node. The three hyperparameters are well-tuned using grid search methods and cross-validation. XGBoost is another popular tree-based algorithm that builds the decision trees in a boosting manner. This refers to the fact that in XGBoost trains every new tree to improve the deficiencies of the previous trees rather than developing each tree independently as that in RF (Chen and Guestrin, 2016). XGBoost has been widely applied for regression with superior performances in many aspects. Key hyperparameters in XGBoost, including learning rate, number of trees, the minimum sum of instance weight (hessian) needed in a child, maximum depth of a tree, minimum loss reduction, Subsample ratio of the training instances, and subsampling parameters are well-tuned by cross-validation.

Several performance metrics are used to quantify the predictive performances of different ML algorithms, including Coefficient of determination (R^2), Root Mean Square Error (RMSE), and Explained Variances (ER). We adopt the average performances during five-folder cross-validation as final performance surrogates. The technical details of the models are not described herein to avoid repetition and are readily available in the cited references.

2.2. Deciphering nonlinear and interaction effects in a data-driven way

This section elaborates on the interpretation method to quantify the relationship of each factor with the usage demand of shared mobility. The emphases are to analyze the nonlinear and interactive effects of different built environment factors with correlation in a data-driven manner. More specifically, the relationship between a built environment factor and the usage demand of a shared mobility system (i.e., dependent variable), and the interactive effects among different built environment factors, are investigated. The most popular method for investigating nonlinear effects in the relevant literature is Partial Dependence Analysis (Ding et al., 2018; Pérez-Fernández and García-Palomares, 2021; Tu et al., 2021; Wang et al., 2022). PDA calculates the marginal effect of a feature on the predicted dependent variable based on a trained ML model and available sample data. PDA can be mathematically expressed as

$$\begin{gathered} PD(x_I) = E_{x_L}(f(x_I, x_L)) = \int f(x_I, x_L) dP(x_I, x_L) \\ X = x_I \cup x_L \\ x_I = \complement_X x_L \end{gathered} \tag{1}$$

X is the feature vector of all features in the analysis. The $\mathbf{x_I}$ is the vector of analyzed features on which the partial dependence analysis will execute. $\mathbf{x_L}$ is the complementary feature set of $\mathbf{x_I}$ and consists of all features besides the analyzed features, namely $\mathbf{x_I} = \complement_{\mathbf{X}} \mathbf{x_L}$. $f(\bullet)$ is the trained algorithm that models the relationships between features and the dependent variable. In the context of this study, $f(\bullet)$ could be any supervised ML algorithm (e.g., Random Forest and XGBoost), as long as the algorithm has good enough predictive performances. To calculate the marginal effect of a feature x_I on the dependent variable, PDA quantifies the predicted values of the dependent variable by the trained ML model over the values of other features in $\mathbf{x_L}$

$$\widetilde{PD}(\mathbf{x_I}) = \frac{1}{K}\sum_{k=1}^{K} f\left(\mathbf{x_I}, \mathbf{x_L}^k\right) \tag{2}$$

where $\{\mathbf{x_L}^1, \mathbf{x_L}^2, \ldots, \mathbf{x_L}^k, \ldots, \mathbf{x_L}^K\}$ are the feature values of $\mathbf{x_L}$ in the used data. Given a specific value of x_I, the corresponding estimated value of the dependent variable is the mean value of enumerating $f(\bullet)$ over the joint values of the given x_I and all $\mathbf{x_L}$ (namely $\{\mathbf{x_L}^1, \mathbf{x_L}^2, \ldots, \mathbf{x_L}^k, \ldots, \mathbf{x_L}^K\}$) in the data. By changing the value of the given x_I and repeating the same process, the continuous or discrete relationship between x_I and the dependent variable can be estimated. This calculation process could be very intensive if the sample size of data is very large. Thanks to the fact that the calculation process is based on available data and the trained ML model without presumptions, PDA can interpret the effects of x_I on the dependent variable in a data-driven way, which is the most significant merit of PDA compared to conventional methods (Hastie et al., 2009).

Note that the PDA in Eq. (1) reflects the effects of $\mathbf{x_I}$ on the dependent variable after considering the average effect of other features in $\mathbf{x_L}$ rather than the independent effects of $\mathbf{x_I}$. This leads to biases in analyzing correlated features. In PDA, the relationship between features $\mathbf{x_I}$ and the dependent variable is estimated by averaging the predicted dependent variable over distributions of features in $\mathbf{x_L}$. Namely, the features in $\mathbf{x_I}$ and $\mathbf{x_L}$ are presumed to be independent during the calculation process. If the prerequisite is violated, the results from PDA may be biased due to the inherent flaw of lacking extrapolations of ML algorithms. For a specific example, investigated features are the population density (x_1) and accessibility to public transit (x_2) in an area, and the dependent variable (y) is the morning peak-hour travel demand for public transit in the area. In the training data of 20,000 samples, the range of x_1 and x_2 are from 1 to 50,000 people per mi^2 and from 0 to 100, respectively. x_1 and x_2 are positively related as areas with a large population are generally provided with public transit services. In the training data, there will not be points that have a very large value of x_1 (e.g., 40,000) and have a small value of x_2 (e.g., 1), because such cases are unrealistic. However, both the population density and the accessibility to public transit indeed affect the apartment price. Let us assume that a multiple-layer perceptron model is fitted for the relationship between $\{x_1, x_2\}$ and y. If the PDA is utilized to interpret the effects of x_1 on y, the $\widetilde{y}$ when is $x_1 = 40000$ is calculated by Eq. (3) over the distributions of x_2 in the data.

$$\frac{1}{20000}\sum_{i=1}^{20000} \widetilde{y}\left(x_1 = 40000, x_2^i\right) \tag{3}$$

During the calculation process, there will be implausible points such as $(x_1 = 40000, x_2 = 1)$. This causes severe biases in the estimated $\widetilde{y}$, because such points do not exist in the training data and ML algorithms such as the multiple-layer perceptron has awful extrapolation ability (Gao et al., 2021d; Hooker, 2004). The potential big issue of PDA has been heavily overlooked in the literature regarding analyzing the effects of built environment factors. However, built environment factors are inherently correlated to some extent. When this assumption of feature independence does not hold, PDA will have severe biases (Molnar, 2020).

To deal with the aforenoted issues, this study introduces a new method of interpreting the main and interactive effects of correlated built environment factors on demand for shared mobility systems. It can overcome the preceding problems of PDA and obtain unbiased results when features are correlated. The approach is called Accumulated Local Effect (ALE) analysis

$$\begin{aligned} A\widetilde{E}(x_I) &= \int_{min(x_I)}^{x_I} E_{x_L|x_I}\left[\frac{\partial f(x_I, x_L)}{\partial x_I} | x_I = s_I\right] ds_I - c_{x_I} \\ &= \int_{min(x_I)}^{x_I} \int \frac{\partial f(x_I, x_L)}{\partial s_I} P(x_L|s_I) dx_L ds_I - c_{x_I} \end{aligned} \tag{4}$$

where $\frac{\partial f(\mathbf{x_I}, \mathbf{x_L})}{\partial \mathbf{x_I}} | \mathbf{x_I} = s_I$ denotes the gradient of $f(\bullet)$ at (s_I,$\mathbf{x_L}$) and represents the local effect of $\mathbf{x_I}$ on $f(\bullet)$ at the point. $c_{\mathbf{x_I}}$ is the specific term for centralizing the results so that the estimated $A\widetilde{E}(\mathbf{x_I})$ has a zero mean over the distributions of $\mathbf{x_L}$. In this manner, ALE analysis reflects how the value of $f(\bullet)$ changes with investigated features $\mathbf{x_I}$, rather than the absolute value of $f(\bullet)$ on $\mathbf{x_I}$, which is the key difference between ALE and PDA. The basic idea of ALE is to calculate the local effect $f\left(\frac{\partial f(\mathbf{x_I}, \mathbf{x_L})}{\partial \mathbf{x_I}} | \mathbf{x_I} = s_I\right)$ at (s_I,$\mathbf{x_L}$), and calculate the average local effect across all values of $\mathbf{x_L}$ with the weight $P(\mathbf{x_L}|s_I)$. By repeating the local effect at different values of $\mathbf{x_I}$ (e.g., in ascending order) and accumulating the calculated local effect from $min(\mathbf{x_I})$ to $\mathbf{x_I}$, the relationship that how $f(\bullet)$ changes with $\mathbf{x_I}$ can be estimated. When averaging the local effect at a certain point, the conditional density of $\mathbf{x_L}$ (i.e., $P(\mathbf{x_L}|s_I)$) rather than the marginal density of $\mathbf{x_L}$ (namely $P(\mathbf{x_L})$) is used. This mechanism solves the extrapolation issue of PDA by avoiding unseen points in the calculation process and leverages the paired difference analysis to eliminate the nuisance feature based on statistical settings (Apley and Zhu, 2020; Molnar, 2020). This further evades the potential biases due to the omitted nuisance features and dependencies among features, namely blocking the confounding impacts of other correlated effects.

We assume $f(\bullet)$ in Eq. (4) to be differentiable. However, $f(\bullet)$ may be a trained ML model and is not mathematically differentiable. In such cases, the below approximation method is used. Let us assume a dataset with J features. The range of feature j is $\mathscr{S}_j = \left[x_{min,j}, x_{max,j}\right]$ for $\forall j \in \{1, 2, \ldots, J\}$. To estimate the effect of feature j on $f(\bullet)$, we divide $\mathscr{S}_j$ into H intervals with fine granularity. Let $\left\{N_j(h) = \left(z_{h-1,j}, z_{h,j}\right] : h = 1, 2, \ldots, H\right\}$ be the bounds of each partitioned interval. Generally, we can use the h^{th} quantile of feature j in the data as z_h. Let $m_j(h)$ be the number of data points in the interval $N_J(h)$ so that.

$$\sum_{h=1}^{H} m_j(h) = M \tag{5}$$

where M is the sample size of all data. For a specific value x for the feature j, let $h_j(x)$ be the index of the interval that x falls in. The main effect of feature j on $f(\bullet)$ is estimated by

$$\widehat{AE}(x_j) = \sum_{h=1}^{h_j(x)} \frac{1}{m_j(h)} \sum_{\left\{i: x_{\backslash j}^i \in N_j(h)\right\}} \left[f\left(z_{h,j}, x_{\backslash j}^i\right) - f\left(z_{h-1}, x_{\backslash j}^i\right)\right] - c_j \tag{6}$$

where $\mathbf{x}_{\backslash j}$ represents the features besides feature j. $f\left(z_{h,j}, \mathbf{x}_{\backslash j}^i\right)$ means replacing all x_j in the interval $h_j(x)$ with the upper bound value $z_{h,j}$ and keep all other features unchanged for all data points. $f\left(z_{h-1,j}, \mathbf{x}_{\backslash j}^i\right)$ means replacing all x_j with the lower bound value $z_{h-1,j}$ while all other features remain unchanged in the interval $h_j(x)$. The local effect of feature j is approximated by the discrepancy between $f\left(z_{h,j}, \mathbf{x}_{\backslash j}^i\right)$ and $f\left(z_{h-1,j}, \mathbf{x}_{\backslash j}^i\right)$ based on the data. As noted in Eq. (4), the results are centralized by subtracting the result by

$$\begin{aligned} c_j &= \frac{1}{M}\sum_{m=1}^{M} \sum_{\left\{i: x_{\backslash j}^i \in N_j(h)\right\}} \left[f\left(z_{h,j}, x_{\backslash j}^i\right) - f\left(z_{h-1}, x_{\backslash j}^i\right)\right] \\ &= \frac{1}{M}\sum_{h=1}^{H} \frac{1}{m_j(h)} \sum_{\left\{i: x_{\backslash j}^i \in N_j(h)\right\}} \left[f\left(z_{h,j}, x_{\backslash j}^i\right) - f\left(z_{h-1}, x_{\backslash j}^i\right)\right] \end{aligned} \tag{7}$$

The above ALE for one feature can be extended for two features (namely two-dimension analysis), which are two-way interaction effects of two features. Note that the obtained result from ALE for one feature reflects how the $f(\bullet)$ changes with the changes in the feature. The results from ALE for two features are not how the $f(\bullet)$ changes with the changes in the two features but the interactive effects of the two features. The overall effect of two features on the dependent variable can be expressed as

$$O(x_1 = v_{11}, x_2 = v_{21}) = M_{x_1}(x_1 = v_{11}) + M_{x_2}(x_2 = v_{21}) + I_{x1,x2}(x_1 = v_{11}, x_2 = v_{21}) \tag{8}$$

where O denotes the overall effect of x_1 and x_2. M_{x_1} and M_{x_2} are the main effect of features x_1 and x_2, respectively. $I_{x1,x2}$ represents the two-way interaction effect between x_1 and x_2. The interaction effect is estimated by two-dimension ALE analysis instead of $O(x_1 = v_{11}, x_2 = v_{21})$. If the interactive effects among two features do not exist, $I_{x1,x2}(x_1 = v_{11}, x_2 = v_{21})$ is zero. To estimate the interaction effect of two features (x_j, x_t), we divide the training sample space into a grid of H^2 rectangular cells. Let (h,p) be the indices of a rectangular cell where j and t correspond to x_j and x_t, respectively. Let $N_{j,t}(h,p) = (z_{h-1,j}, z_{h,j}] \times (z_{p-1,t}, z_{p,t}] : h,p \in (1,2,\dots,H)\}$ and $m_{j,t}(h,p)$ be the rectangular cell (h,p) and the number of sample points in the cell, respectively, to satisfy Eq. (9).

$$\sum_{h=1}^{H}\sum_{p=1}^{H} m_{j,t}(h,p) = M \tag{9}$$

The local two-way interactive effect between features j and t at the rectangular cell (h,p) without centralization is estimated by

$$\widehat{AE}(x_j, x_t) = \sum_{h=1}^{h_j(x_j)}\sum_{p=1}^{h_t(x_t)} \frac{1}{m_{j,t}(h,p)} \sum_{\{i: x^i_{\setminus j,t} \in N_{j,t}(h,p)\}} \Delta I\left(H, h, p, x^i_{\setminus j,t}\right) \tag{10}$$

$$\begin{aligned}\Delta I\left(H,h,p,x^i_{\setminus j,t}\right) &= \left[f\left(z_{h,j}, z_{p,t}, x^i_{\setminus j,t}\right) - f\left(z_{h-1,j}, z_{p,t}, x^i_{\setminus j,t}\right)\right] \\ &\quad -\left[f\left(z_{h,j}, z_{p-1,t}, x^i_{\setminus j,t}\right) - f\left(z_{h-1,j}, z_{p-1,t}, x^i_{\setminus j,t}\right)\right] \\ &= f\left(z_{h,j}, z_{p,t}, x^i_{\setminus j,t}\right) + f\left(z_{h-1,j}, z_{p-1,t}, x^i_{\setminus j,t}\right) - f\left(z_{h-1,j}, z_{p,t}, x^i_{\setminus j,t}\right) - f\left(z_{h,j}, z_{p-1,t}, x^i_{\setminus j,t}\right)\end{aligned} \tag{11}$$

Similar to Eq. (6) and (7), the interactive effect is centralized so that the main effects of x_j and x_t both have mean values of zero by Eqs. (12) and (13)

$$\begin{aligned} I\left(\widehat{AE}(x_j, x_t)\right) = \widehat{AE}(x_j, x_t) &- \sum_{h=1}^{h_j(x_j)} \frac{1}{m_j(h)} \sum_{p=1}^{H} m_{j,t}(h,p)\left\{\left(\widehat{AE}(z_{h,j}, z_{p,t}) - \widehat{AE}(z_{h-1,j}, z_{p,t}\right)\right\} \\ &- \sum_{p=1}^{h_t(x_t)} \frac{1}{m_t(h)} \sum_{h=1}^{H} m_{j,t}(h,p)\left\{\left(\widehat{AE}(z_{h,j}, z_{p,t}) - \widehat{AE}(z_{h,j}, z_{p-1,t}\right)\right\}\end{aligned} \tag{12}$$

$$\overline{I}\left(\widehat{AE}(x_j, x_t)\right) = I\left(\widehat{AE}(x_j, x_t)\right) - \frac{1}{M}\sum_{h=1}^{H}\sum_{t=1}^{H} m_{j,t}(h,p) I\left(\widehat{AE}(x_{h,j}, x_{p,t})\right) \tag{13}$$

The average differences in prediction analysis and the synthetic points in calculations of Eqs. (6) and (13) are based on conditional distributions of data in each interval $N_j(h)$ rather than marginal distributions. Hence, it can still interpret the effects of each feature even though features are correlated (Apley and Zhu, 2020; Molnar, 2020). In this regard, ALE analysis can provide theoretically unbiased estimations about how the dependent variable changes with the changes in a feature or several features. Technically, the interactive effect analysis using ALE can be extended to interpret more complex interactions such as three or five-dimension interactions. Nonetheless, results over three-dimensional are tough for humans to understand. Meanwhile, the interactive effects of over two features are not common and generally not substantial in the studies regarding the effects of built environments. Hence, we mainly explore the nonlinear effects of each built environment factor and two-way interactions among different factors. More comprehensive details about ALE are available in Apley and Zhu (2020) and Molnar (2020).

3. Data descriptions

We utilize two types of datasets in this study: synthetic datasets and a field dataset about bike-sharing systems. We do not merely use field datasets as there are no known ground truths about the effects of each built environment factor in field data. Thus, it is impossible to directly compare and validate which interpretation method is more aligned with reality. Alternatively, we first generate synthetic data to validate and demonstrate the reliability, superiority and unbiased merit of ALE compared to PDA for analyzing the effects of correlated built environment factors. Afterward, an empirical analysis of the impacts of built environment factors on demand for bike sharing is conducted.

3.1. Synthetic datasets

Two synthetic datasets are generated. The first synthetic dataset includes two features that are correlated with each other. The dependent variable y_1 is defined as

$$y_1 = 5x_1 + {x_2}^2 \tag{14}$$

where x_1 is in the range of 0 to 10. x_2 is randomly generated based on x_1 with a Pearson product-moment correlation coefficient of 0.82. The synthetic dataset is created to test the ability to interpret the nonlinear effects of a feature. One thousand points are simulated, and the relations between x_1 and x_2 are shown in Fig. 1. The second synthetic data includes six features, and the dependent variable y_2 is defined as

$$y_2 = z_1 + z_2 + 0.5z_3 + z_4 + 0.5z_5 + 0.3z_6 \tag{15}$$

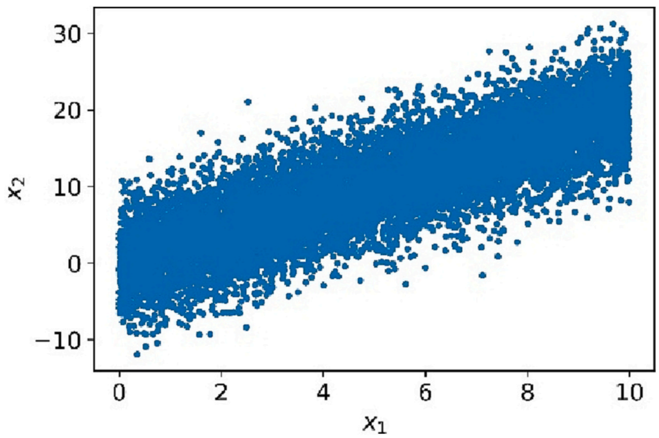


Fig. 1. Features in the first synthetic data.

where z_1 is in the range of 0 to 1. z_2 and z_3 are generated based on z_1. The correlation coefficients of z_2 and z_3 with z_1 are 0.8 and 0.7, respectively. z_4, z_5 and z_6 are independent features and have no correlations with other features. In this regard, three correlated features and another three uncorrelated features exist in the second synthetic data. Ten thousand points are created for the second synthetic data. A random forest algorithm is tuned and trained to model the relationships between explanatory variables and the dependent variable. In five-folder cross-validations, the average value of R^2 for the first and second synthetic datasets are 0.999 and 0.983, respectively. The high values of R^2 indicate the trained algorithms can properly fit the relationship between features and dependent variables. In the conventional regression, Variance Inflation Factors (VIFs) are generally calculated to avoid high multicollinearity. Therefore, some may argue that the highly correlated features may or should be excluded in the regression so that the problem due to correlated features can be solved by omitting the feature with a high VIF value. In this study, we will show that the criterions based on VIF cannot solve the biases due to correlated features in PDA. The VIF value of two features in the first synthetic data is 9.2. The VIF values of six features in the second synthetic are 8.1, 2.6, 1.6, 3.4, 3.4 and 3.3, respectively. The VIF values are all less than the generally used threshold (i.e., 10). Therefore, these features will be kept in the conventional regression process. Based on the trained algorithms, PDA and ALE are used separately to interpret the effects of each feature on the dependent variable. The results using the two different methods are compared in terms of alignment with the ground truths in Eqs. (14) and (15).

3.2. Field dataset about dockless bike-sharing systems

The empirical analysis uses the transaction data of a dockless bike-sharing system (DLBS) in Shanghai, China, which is one of the largest cities and has a population of over 24 million in an area of over 6300 km^2. The data cover more than 27 million trip transactions of 14 continuous days (from Aug 26, 2018 to Sep 08, 2018) from over 635,000 shared bikes. Each transaction represents a trip and has a trip ID, bike ID, timestamps, and starting and ending coordinates. Preprocessing is executed to filtrate outliers due to technical errors (e.g., connection failure) or abnormal user behavior. We divide the study area into a grid of 0.01 longitude × 0.01 latitude rectangles, which is around 1 km × 1 km rectangles. Each rectangle displayed in Fig. 2 is an analysis zone. We map the trips into the partitioned grid based on the starting coordinates of the trip, on account that the starting point reflects the demand location of using DLBS. The dependent variable is the daily demand for DLBS in each analysis zone during workdays (i.e., from Monday to Friday). Some grids are in special terrains such as rivers, lakes or wetlands, and thus have neglectable usage demand. Such areas are not valid for

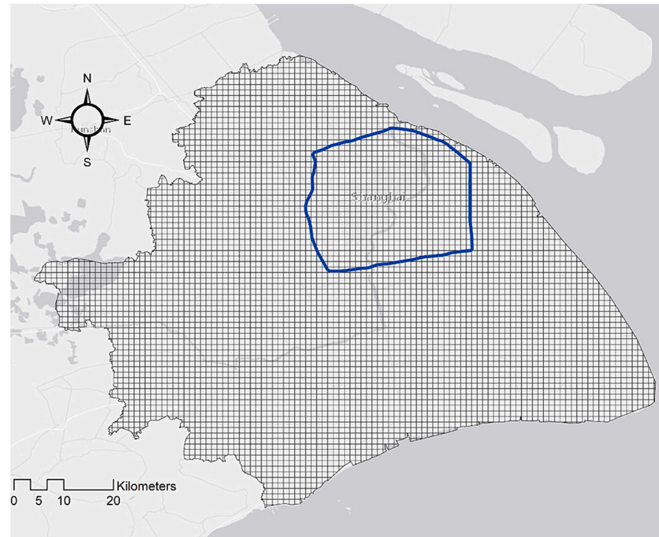


Fig. 2. The study area and analysis zones.

analysis. Therefore, we have excluded the zones with a daily usage demand of less than 20 to avoid biases. Finally, 2454 analysis zones are obtained for analysis. The spatial distribution of the daily demand for DLBS is demonstrated in Fig. 3. The spatial variation in demand for DLBS can be clearly observed, implying the effects of the built environment on the daily demand for DLBS. It can be easily observed that the data cover most of the municipal area of Shanghai and thus different populations in Shanghai. The data was from one of the main local operators, Mobike, who took over 40% market share of bike-sharing systems in Shanghai in 2018. These corroborate the representative of the data to represent the usage patterns of bike sharing in Shanghai.

For built environment factors in each analysis zone a, we consider 5-D built environment factors developed by Cervero et al. (2009). We utilize three data resources to measure these built environment factors in each analysis zone, including Point of Interest (POI) data from the local navigation platform Amap (2022), road network data from OpenStreetMap, and local population statistics. The used POI dataset covers over 1.12 million POIs in Shanghai. Each POI is represented by the element name, address, element types, and coordinates (longitude

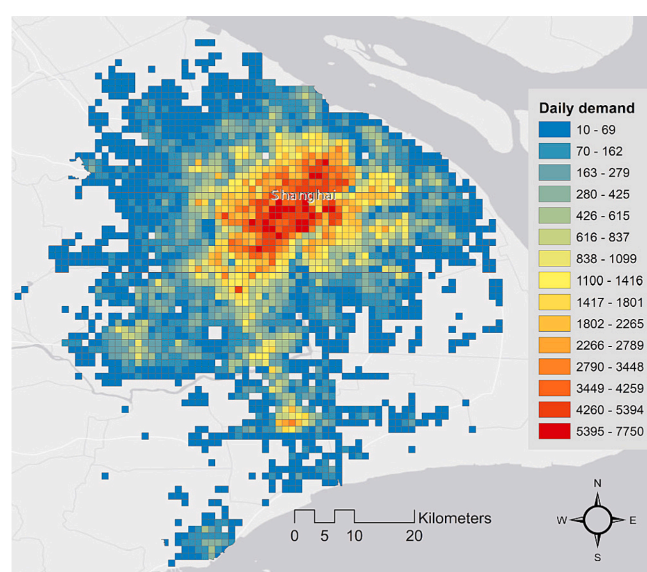


Fig. 3. Spatial distributions of daily usage demand of DLBS in workdays.

Table 1
The investigated built environment factors.

Explanatory variables	Unit
Density	
Population density (PD)	persons/km^2
Employment density (ED)	number/km^2
Diversity	
Commercial land use ratio (CLUR)	%
Living land use ratio (LLUR)	%
Public management and service land use ratio (PLUR)	%
Park and square land use ratio (PSLUS)	%
Industry land use ratio (ILUR)	%
Land use entropy (LUE)	
Design	
Motorway density (MOD)	km/km^2
Motorized road density (MRD)	km/km^2
Branch road density (BRD)	km/km^2
Bicycle lane density (BLD)	km/km^2
Destination	
Living facility density (LD)	amount/km^2
Commercial service density (CD)	amount/km^2
Industrial facility density (ID)	amount/km^2
Leisure facility density (LFD)	amount/km^2
Education service density (EFD)	amount/km^2
Park and square density (PSD)	amount/km^2
Parking lot density (PLD)	amount/km^2
Distance to transit	
Metro station coverage ratio (MSR)	%
Bus station coverage ratio (BSR)	%

Note: Details of calculating the variables in this table refer to Gao et al. (2021b).

and latitude). According to the POI type categories from the data provider, the POIs are categorized into 267 types for different utilization purposes. The details of POI categories are available on the website (Amap, 2022). We leverage the POI data to quantify the built environment factors about land use characteristics in Table 1. Six land-use types are finally considered referring to classification standards in China. The technical methods of using POIs to quantify the land use measurements are available in Gao et al. (2021b) and skipped for simplification. The road network data from OpenStreetMap is utilized to quantify road-relevant built environment factors in Table 1. The population statistics in Shanghai are used to measure population and employment density in each analysis zone. Based on the extracted workday daily demand and corresponding built environment factors in each analysis zone, the methodology in Section 2 is used to model and quantitatively interpret the effects of built environment factors on demand for DLBS.

4. Results and discussions

4.1. Comparative results based on synthetic data

The results of PDA and ALE based on the first synthetic data are demonstrated in Fig. 4. It can be observed in Fig. 4 (a) and (b) that PDA cannot well interpret and reproduce the actual effects of x_1 and x_2. Especially, the interpreted effects of x_1 by PDA has obvious biases as compared to actual effects. In comparison, the interpreted effects of both x_1 and x_2 by ALE are perfect in line with ground truths as visualized in Figs. 4 (c) and (d). The results based on the second synthetic data are summarized in Fig. 5. We merely present the results of features that are correlated. It should be noted that our synthetic analysis has used VIF values to screen out multicollinearity as conventional regressions do. Again, the interpreted results by PDA present considerable biases when correlations among features exist. Particularly, the biases by PDA are not systematical but could be either underestimation or overestimation. In

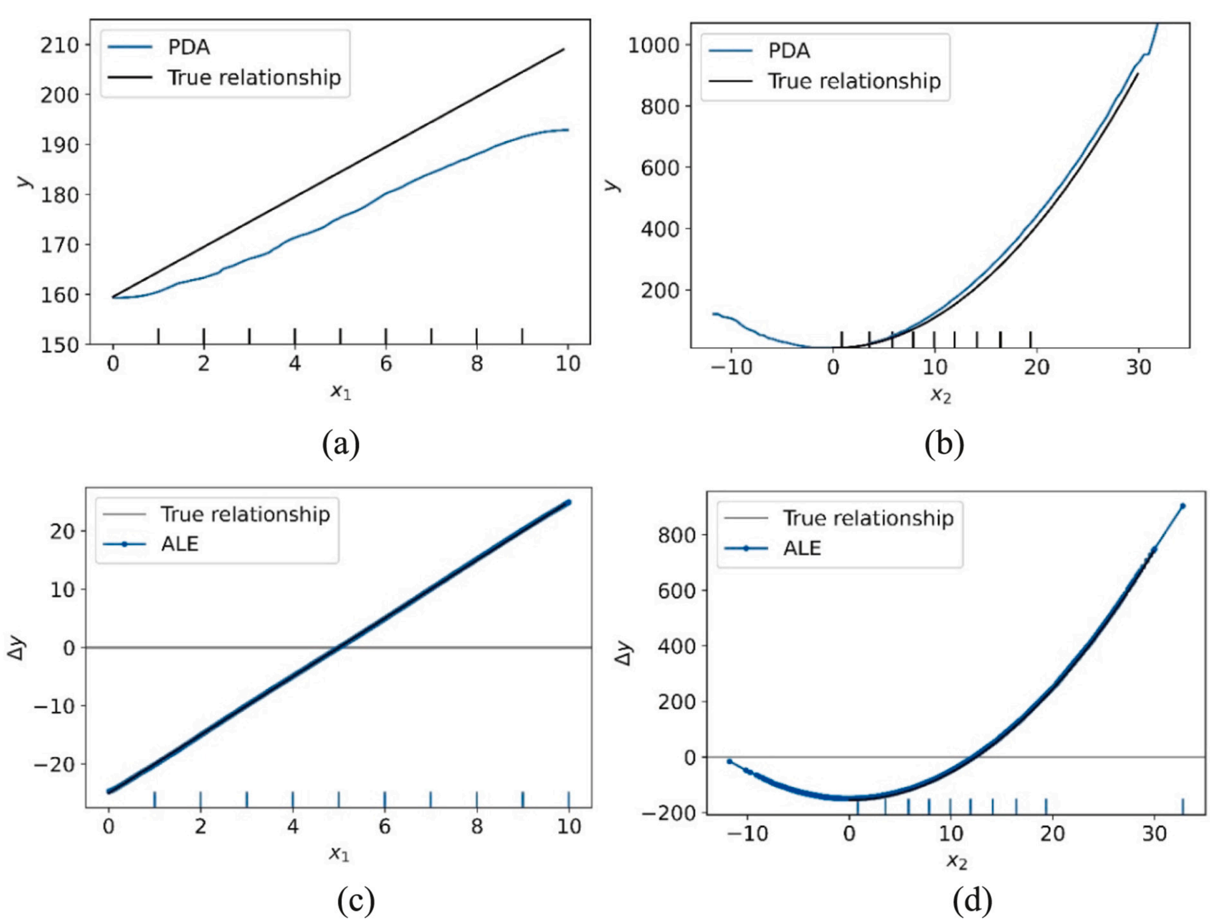


Fig. 4. Results based on the first synthetic data: (a) and (b) are the PDA results; (c) and (d) are the ALE results.

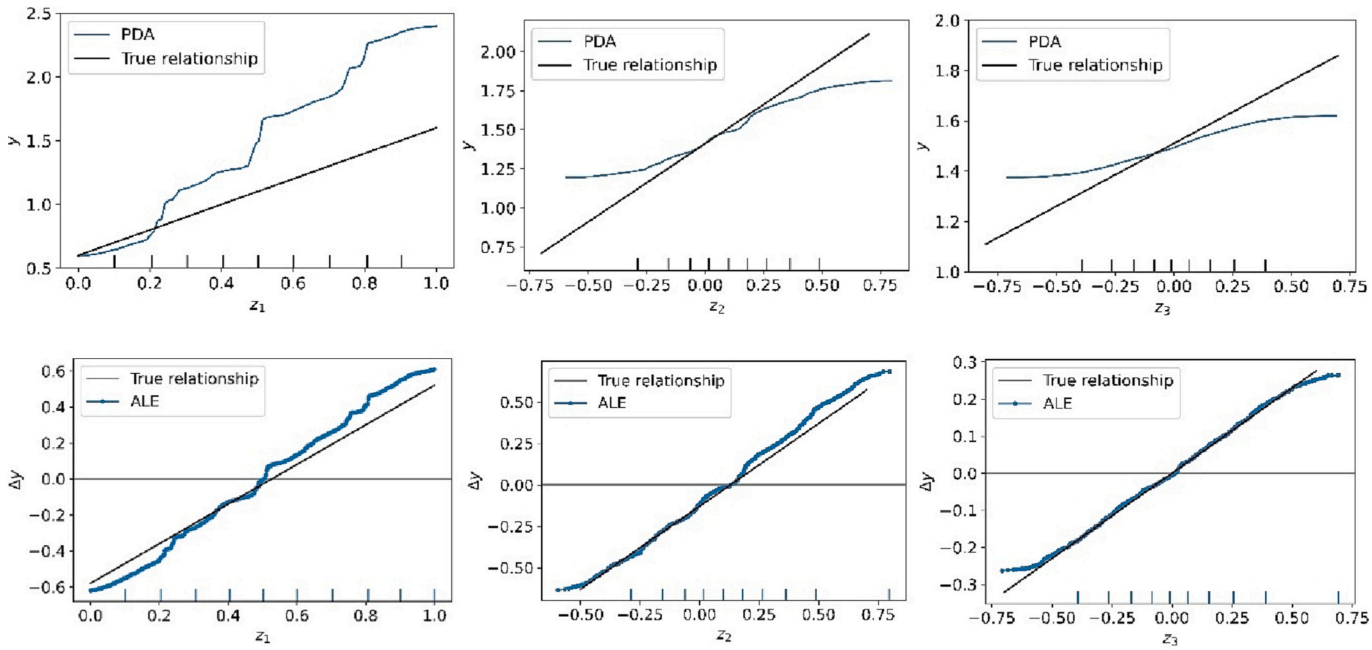


Fig. 5. Results based on the second synthetic data. The first and second rows are results from PDA and ALE, respectively.

contrast, the interpreted results by ALE in Fig. 5 can well reproduce the actual effects of correlated features (z_1, z_2, z_3). The results based on two synthetic data corroborate the shortage of PDA for investigating the effects of correlated factors and validate the ability of ALE to interpret the impact of correlated features. The results imply that the existing studies using PDA for investigating the effects of correlated built environment factors have a high risk of biased results. Especially, many studies do not check and pay attention to the correlations of features as correlations of features may not influence the predictive performances of many machine learning algorithms. Nevertheless, it is heavily overlooked that correlated features may not deteriorate the model fitness or predictive performances but would likely result in biases in the interpreted effects of different built factors, which are the emphasis of many empirical studies (rather than prediction).

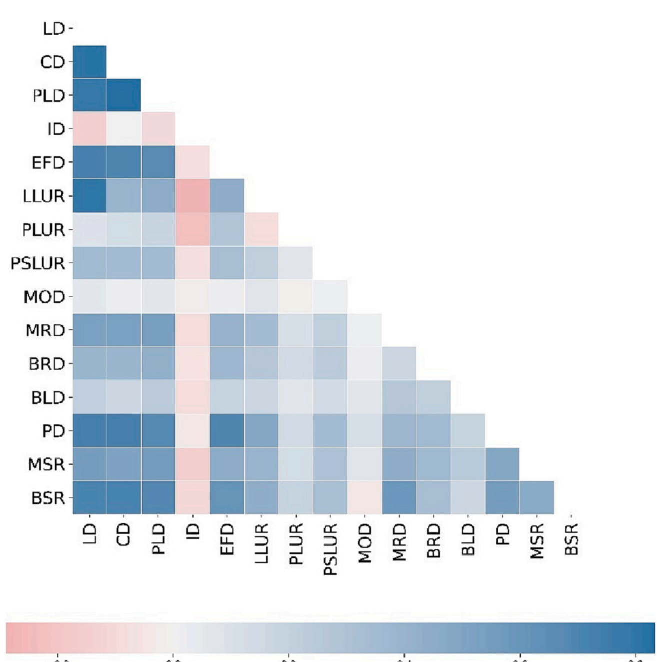


Fig. 6. Correlations of features.

4.2. Results of the empirical study of DLBS

The analysis aim is to interrogate the effects of various built environment factors on demand for DLBS during workdays in an analysis zone. Following the general feature selection process, we screen out features with a variance inflation factor larger than 7.5 to avoid multicollinearity. Permutation importance analysis is also used to filtrate features with trivial importance. These are targeted to prevent the argument that some correlated features should be screened out in standard feature selection processing, so that correlated features may not be used for final analysis, which to some extent counters the meaning and superiority of ALE. Fifteen features are kept for analysis after selections, and their spearman correlations are displayed in Fig. 6. After standard feature selection is executed, the remaining features still present some degrees of correlations. Importantly, it is inherent that built environment factors are, to some extent, correlated. For example, places with a lot of commercial services are supposed to have high employment density; places with a lot of living communities are supposed to be surrounded by daily commercial services in Chinese cities. If very strict criteria are used to filtrate correlated features (e.g., very low VIF value and correlation coefficients), some features with useful information would be filtered out. Therefore, it is ubiquitous that we need to consider correlated built environment factors in the analysis. Different supervised machine learning algorithms are compared to select the best candidate. The random forest algorithm has the best performance with the R^2 of 0.859, root mean squared error of 428.04, and the explained variance of 86.4%. Hence, it is selected as the ML model for the following interpretation analysis.

4.2.1. Nonlinear effect of built environment factors

We first analyze the main effects of each built environment factor using one-dimension ALE. The results are summarized in Fig. 7. Living facility density, commercial service density, and education service density are positively related to the demand for DLBS. These are easily understandable as DLBS are convenient tools for short-distance trips for various daily activities. A higher density of living facilities, commercial services and education services in an area generally means higher

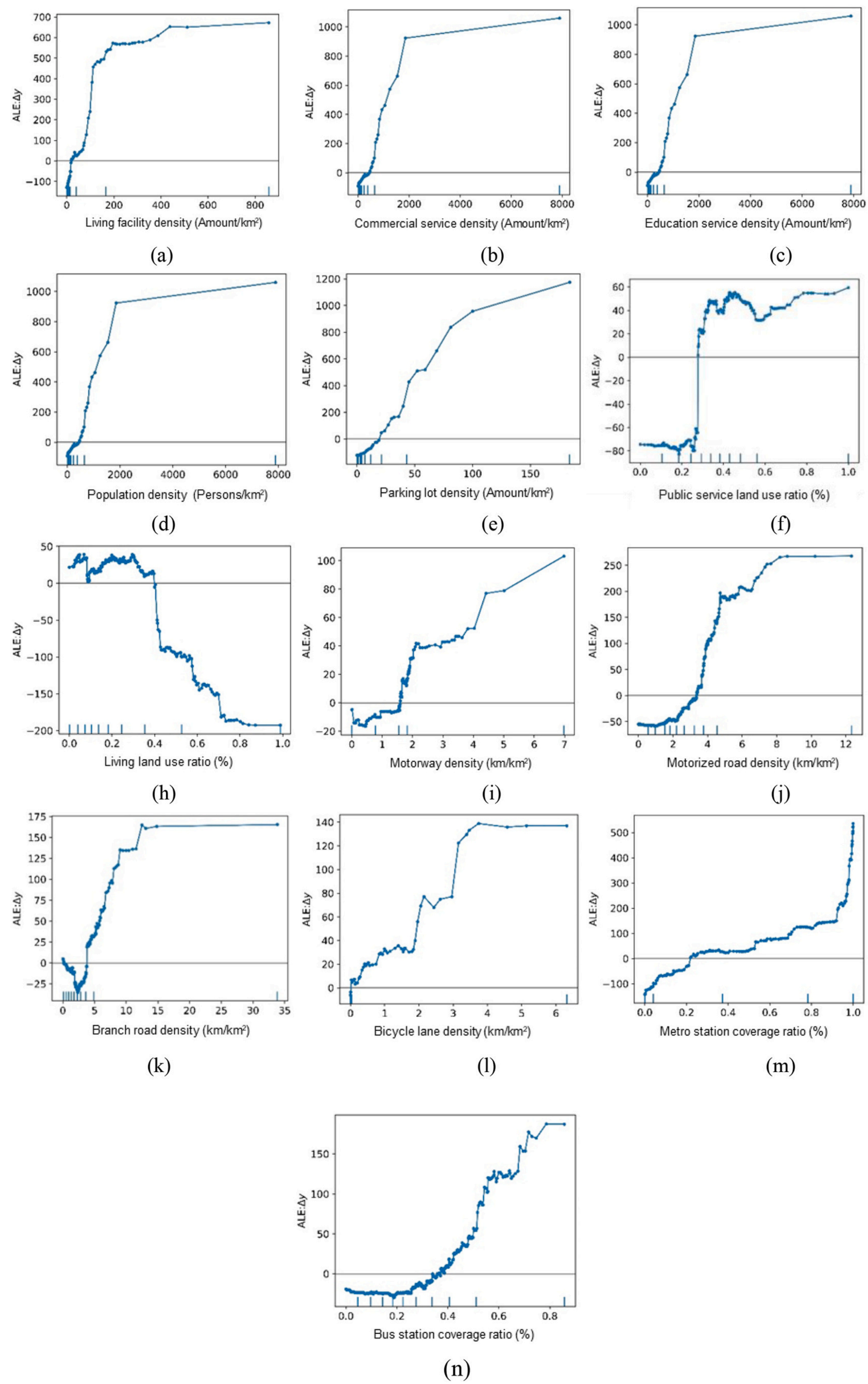


Fig. 7. Main effects of various built environment factors. The unit of Y-axis is the amount of trip per day.

demand attraction, and thus more demand for using DLBS to access these services. More importantly, the effects of the three built environment factors all present nonlinear and threshold effects. The nonlinear and threshold effects refer to the pattern that the demand for DLBS increases notably with the three factors but reaches a plateau when the values of the factors exceed certain thresholds. For instance, the demand for DLBS upsurges significantly with commercial service density in Fig. 7 (b), but would not change obviously when the density of commercial service is over 2000 per km^2. The nonlinear and threshold effects of built environment factors on travel demand of other transport modes have also been found in the literature, such as metro ridership (Ding et al., 2019b) and ride-sourcing (Bi et al., 2022). The thresholds in the effects of living area density and education service density are around 200 and 40 per km^2, respectively. The results imply that biased results will be acquired if linear effects are assumed for these factors. The results also indicate the superiority of data-driven methods in investigating the complex effects of built environment factors. For the effect of population density in Fig. 7 (d), a similar effect pattern can be observed. The demand for DLBS increases with population density but keeps almost unaffected when the population density exceeds 43,000 per km^2. The parking lot density is found to be positively related to the demand for DLBS (see Fig. 7(e)). This might be attributed to the fact that many designated parking zones or lots were provided for legally parking shared bikes in Shanghai. Therefore, analysis zones with many parking lots are expected to attract more usage of DLBS.

For land use ratio factors, the public service land use ratio (see Fig. 7 (f)) has trivial influences on the demand for DLBS when the value is less than 0.25. However, its effect surges sharply until the value of 0.4 and tends to be consistently unchanged. An apparent nonlinear threshold effect is observed as well. The effect of the living land use ratio in Fig. 7 (h) shows that it is negatively linked to the demand for DLBS, especially when the living land use rate is over 0.4. The potential reason is that many residential communities in China are surrounded by protective fences for safety reasons, differing from the residential communities in Western nations. DLBS are not allowed to enter these residential communities and can merely be parked outside of living areas (e.g., at the entrance). The area of the defined analysis zone in this study is around 1 $\times$ 1 km^2. If the living land use ratio in an analysis zone is very high, the analysis zone is likely to mainly cover the internal area of a residential community where DLBS cannot enter, which could result in lower demand in the analysis zone. It should be noted that if one area with a lot of residential buildings, its influences on the demand for DLBS should be the joint effect of living facility density and living land use ratio, namely the main effects of the two features and their interaction effects as shown in Eq. (8). Fig. 7 (i)-(l) demonstrate the effects of densities of different road types on demand for DLBS. The demand for DLBS is positively associated with the density of bicycle lane density, branch road density, motorized road density, and motorway density. The finding is in alignment with the results in the literature that indicates the increase in road density is helpful for the usage of shared micro-mobility (Chen and Ye, 2021; Huo et al., 2021). Differing from the literature, we distinguish the effects of different road types. The effects of bicycle lane density, branch road density and motorized road density present nonlinear effects. When the density of bicycle lane density, branch road density and motorized

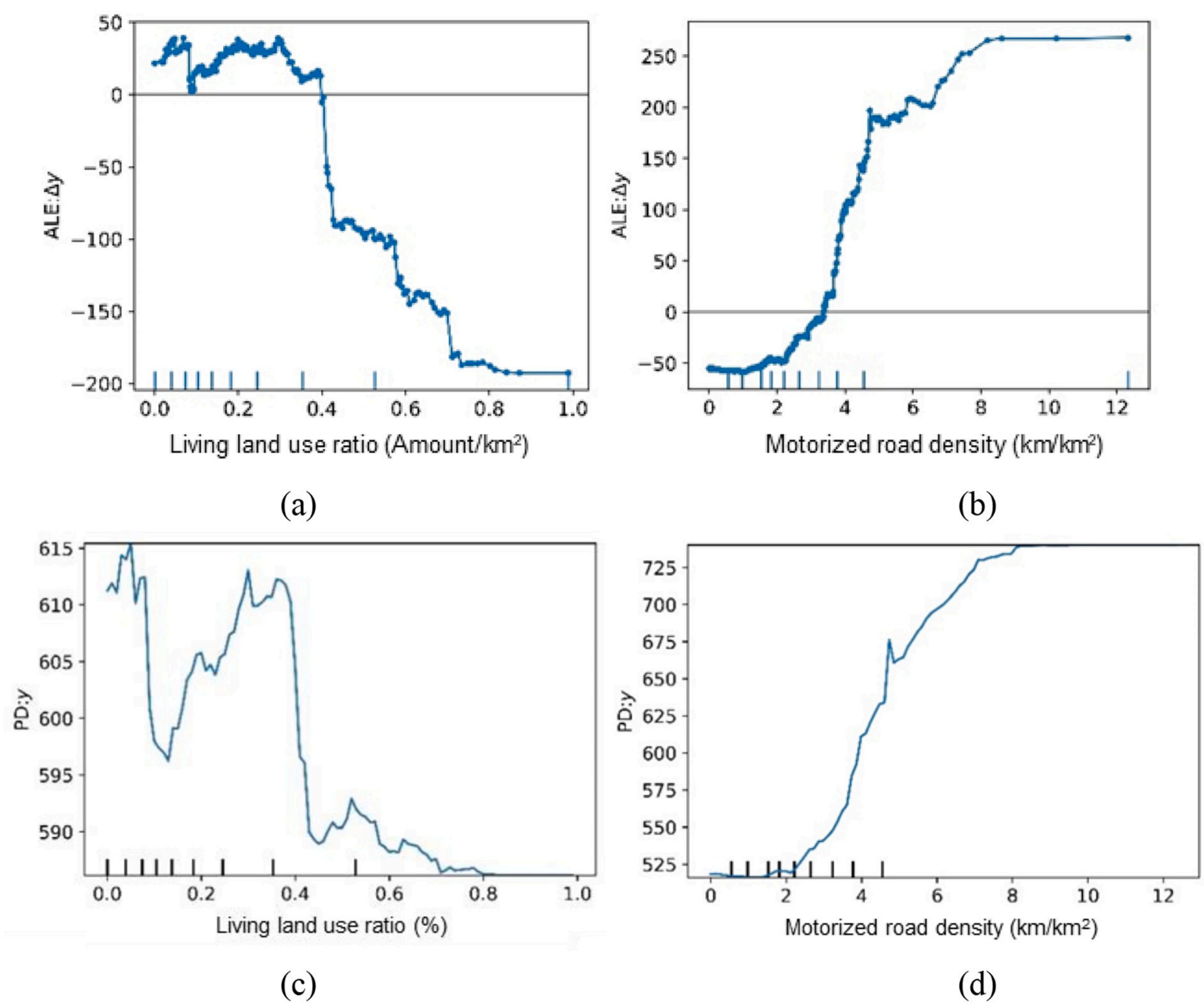


Fig. 8. Main effects of living land use ratios and motorized road density by ALE ((a) and (b)) and PDA ((c) and (d)).

road density are less than 4, 13 and 9 km/km^2, respectively, the demand for DLBS increases with the density of these roads but hardly varies after these thresholds. The metro station coverage ratio has a positive influence on the demand for DLBS, as shown in Fig. 7 (m). The reason should be that DLBS is a convenient transport tool for first- and last-mile access to and egress from metro stations, and the finding is corroborated by relevant literature (Guo and He, 2020; Huo et al., 2021). It is intriguing to find that the effect of the metro station coverage ratio turns more substantial when its value exceeds 0.9. The potential reason is that if the value of the metro station coverage ratio in an analysis zone is over 0.9, it is very likely that there is at least one metro station in the analysis zone, which is a demanding area for using DLBS as feeder transport modes. In contrast, the bus station coverage ratio has trivial effects when its value is less than 0.25 and has positive impacts on the demand for DLBS when the value is over 0.25, as shown in Fig. 7 (n). The effects of industry facility density and park and square land use ratio are negligible, so they are not discussed in detail herein.

As per the results, many built environment factors are found to have nonlinear and threshold effects on the demand for DLBS, which are hard to be identified by conventional methods with presumed model formulations. Even though PDA can also investigate the nonlinear effects, it has biases when analyzing correlated built environment factors, as demonstrated in Section 4.1. Fig. 8 further displays the differences in the impact of the same factors obtained by ALE and PDA. In the effect of the living land use ratio in Fig. 8 (a) and (c) from PDA, the results show unreasonable patterns with fluctuations that are hard to explain. The effect of motorized road density obtained by PDA shows a similar tendency as that obtained by ALE. However, the magnitudes of the impact interpreted by ALE and PDA are distinct. The predicted increase in the demand for DLBS is around 325 based on ALE but is around 220 based on PDA when the motorized road density increases from 0 to 9 km/km^2. Although the ground truths about the effect in the empirical study are impossible to know, ALE should have more credible and reliable results when correlated features exist according to the analysis results of synthetic data.

4.2.2. Interactive effects between built environment factors

The potential two-way interaction effects between built environment factors on demand for DLBS are investigated. The value of two-dimension ALE should be zero if no interaction effect exists between two factors, as indicated in Eq. (8). We enumerate all pairs of any two built environment factors (e.g., $15 \times 14 \div 2 = 105$ pairs). Only results with considerable interaction effects are discussed herein and are summarized in Fig. 9. Education facility density and commercial density have a noticeable interactive impact on demand for DLBS. Particularly, the interaction effect changes with the value of the two factors rather than invariant, as shown in Fig. 9 (a). When one of the two factors has a large value and another has a relatively smaller value, there is a negative interaction effect. Nevertheless, a considerable positive interaction is observed when both features have large values. The potential reason is that DLBS is very prevalent in student groups as a low-budget and convenient transport mode for daily activities such as shopping, eating, and entertainment, so the demand for DLBS would be higher as compared to other analysis zones if one analysis zone has a high density of both commercial service and schools/universities. Similar synergistic effects can be observed between the metro station coverage ratio and commercial service density (see Fig. 9 (b)) and between parking lot density and metro station coverage ratio (see Fig. 9 (c)). The results indicate that jointly promoting the two built environment factors can generate additional effects for improving the usage of DLBS. Fig. 9 (d) shows noticeable interactions between commercial service density and parking lot density. When the parking lot density is very high, the effect of commercial service density is not as notable as when the parking lot density is low. In contrast, the effect of parking lot density is more substantial when the commercial service density is high. Fig. 9(e) illustrates that the changes in living facility density have a larger influence on the changes on the demand for DLBS when the metro station coverage ratio is high. This may be attributed to the popularity of DLBS as a feeder to metro stations in Chinese cities (Guo and He, 2020; Radzimski and Dzięcielski, 2021). When one analysis zone has high accessibility to metro stations, it is expected that there will be much more

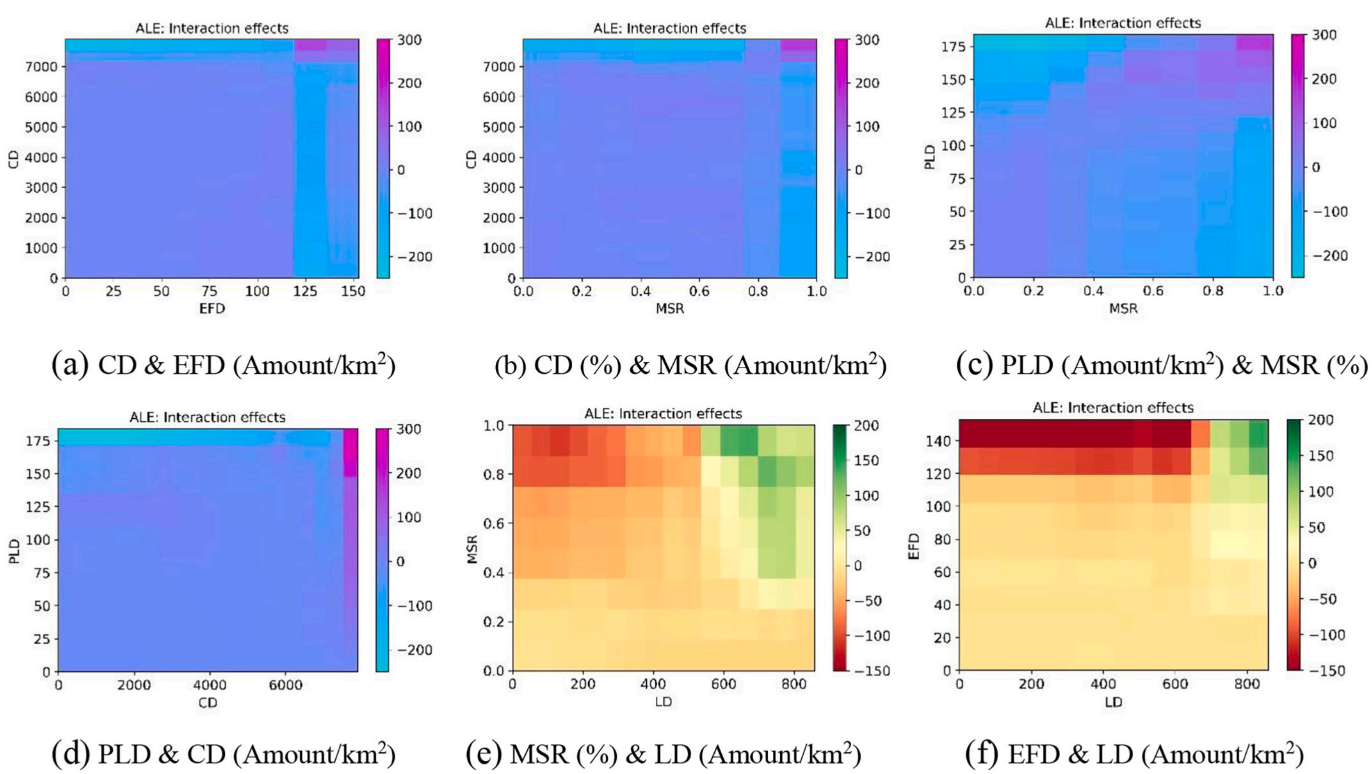


Fig. 9. Two-way interactions among built environment factors.

demand of using DLBS to get access and egress to metro stations for people in surrounding living areas. A similar interactive effect can be found between living facility density and education facility density in Fig. 9(f). The changes in living facility density lead to more considerable differences in demand for DLBS when education facility density is high. Herein, we mainly illustrate the existence of significant interactions among different built environment factors and the ability of the proposed method to interpret the complex interactive effects of built environment factors. Several other interactions are revealed in the analysis and provided in the appendix but are not elaborated on for simplification.

A lot of built environment factors jointly affect the demand for DLBS in one area, and their accumulated effects are supposed to be complex and interact with each other. Solely changing one of the built environment factors may result in an increase or decrease in the demand for DLBS, but the combined effect of simultaneously changing two built environment factors is not necessarily the direct sum of the pure effects of the two factors. The complex interaction effects are very tough (if possible) to be modeled or investigated in conventional regression methods by adding presumed interaction terms (e.g., linearly two-way interaction, $\alpha \times x_1 \times x_2$). The data-driven ML algorithm and interpretation techniques like ALE present more powerful abilities for modeling and unraveling such interaction effects than conventional models.

5. Conclusions

This study leverages an unbiased data-driven approach for examining the complex (nonlinear and interactive) effects of correlated built environment factors on demand for shared mobility. Particular emphasis has been given to unraveling the complex effects in an unbiased and data-driven way that can block the adverse impacts of correlations among built environment factors. Supervised machine learning algorithms are used to model the influences of different built environment factors. A data-driven interpretation technique based on the paired difference analysis and accumulated local effects to eliminate the nuisance feature is leveraged to interpret the nonlinear and interactive effects of built environment factors. Synthetic data are first utilized to validate the reliability and unbiased nature of the proposed approach. Afterward, an empirical case study about bike-sharing systems in Shanghai is conducted. The results are analyzed to shed light on the effects of built environment factors on demand for DLBS. The main contributions include:

- Accumulated local effect analysis can successfully interpret the nonlinear and interactive effects of correlated built environment factors on demand for shared mobility, which partial dependency analysis fails to do. Partial dependency analysis will result in biases when investigating the effects of correlated built environment factors. The finding is corroborated by synthetic data and implied by field data.
- The main effects of many built environment factors on demand for DLBS present nonlinear and threshold patterns that can be successfully investigated by accumulated local analysis. Based on the empirical analysis, quantitative results about the effects of different built environment factors are obtained and discussed. The results provide quantitative support (e.g., thresholds in the effect of some built environment factors) and tools for planning DLBS and improving the built environment to facilitate DLBS.
- There are complex interaction effects between different built environment factors on demand for DLBS. The interactions among two built environment factors may change with the values of the two factors, which can be investigated using the proposed method. For instance, education facility density and commercial density have variant interactions. When one of the two factors has a large value, and another has a relatively smaller value, there is a negative interaction effect. Nevertheless, a considerable positive interaction is observed when both factors have large values. Interaction effects of other built environments are discussed as well.

It is worth mentioning that although the proposed methods offer data-driven modeling and interpretation of the effects of different built environment factors, the interpreted relationships are not necessarily causal and, thus, should be explained by combining expertise and practical knowledge. This is a common dilemma for almost all statistical analysis methods. Another potential limitation of the proposed method is that data-driven analysis requires adequate data to attain reliable results. If the sample size of data is not enough or there are noticeable outliers in the used data, the interpreted results may be biased as data-driven methods have a higher risk of overfitting than conventional model-based methods. This may not be an issue for share mobility analysis as current shared mobility systems naturally produce many useful field data, such as transaction data and positioning data. Another point worth mentioning is that if several features are highly correlated (e.g., the correlation coefficient of 0.95), even ALE is not able to distinguish their impacts on the dependent variable as it is not plausible in statistics. In such cases, highly correlated features should be combined into one feature using dimension reduction methods.

The proposed analysis framework is general and can be easily transplanted for analyzing the effects of built environment factors on other shared mobility systems and even urban mobility in general. An exciting direction of further research is to interrogate the potentially distinct effects of a built environment factor on different shared mobility systems, such as car-sharing, e-scooter sharing, and ride-hailing. Furthermore, although this study investigated several different built environment factors, many other factors, such as additional built environment aspects and weather, are not investigated due to data limitations. In this study, we did not specifically analyze the temporal fluctuations of the demand in an analysis zone, which can be further extended in the future. Moreover, our analysis does not distinguish the demand for different trip purposes as we do not have ground-truth information on trip purposes in the data. Notwithstanding, the effects of the built environment on demand for different trip purposes may be divergent. Exploring how the built environment influences demand for different trip purposes is a fascinating research direction. A possible method is to infer trip purpose based on the destination, points of interest around the destination, and the timestamp of each trip. We focus on data-driven and unbiased methods for deciphering nonlinear and interaction effects, so the results from our empirical analysis may have limitations and may merely reflect the case in Shanghai. Using more datasets from different contexts and making comparisons among multiple cities could be one potential avenue of future work.

Credit authorship contribution statement

Kun Gao: Conceptualization, Data curation, Methodology, Formal analysis, Funding acquisition, Writing - original draft, Writing - review & editing. **Ying Yang:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Jorge Gil:** Formal analysis, Validation, Writing - review & editing. **Xiaobo Qu:** Conceptualization, Validation, Supervision, Writing - review & editing.

Declaration of Competing Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

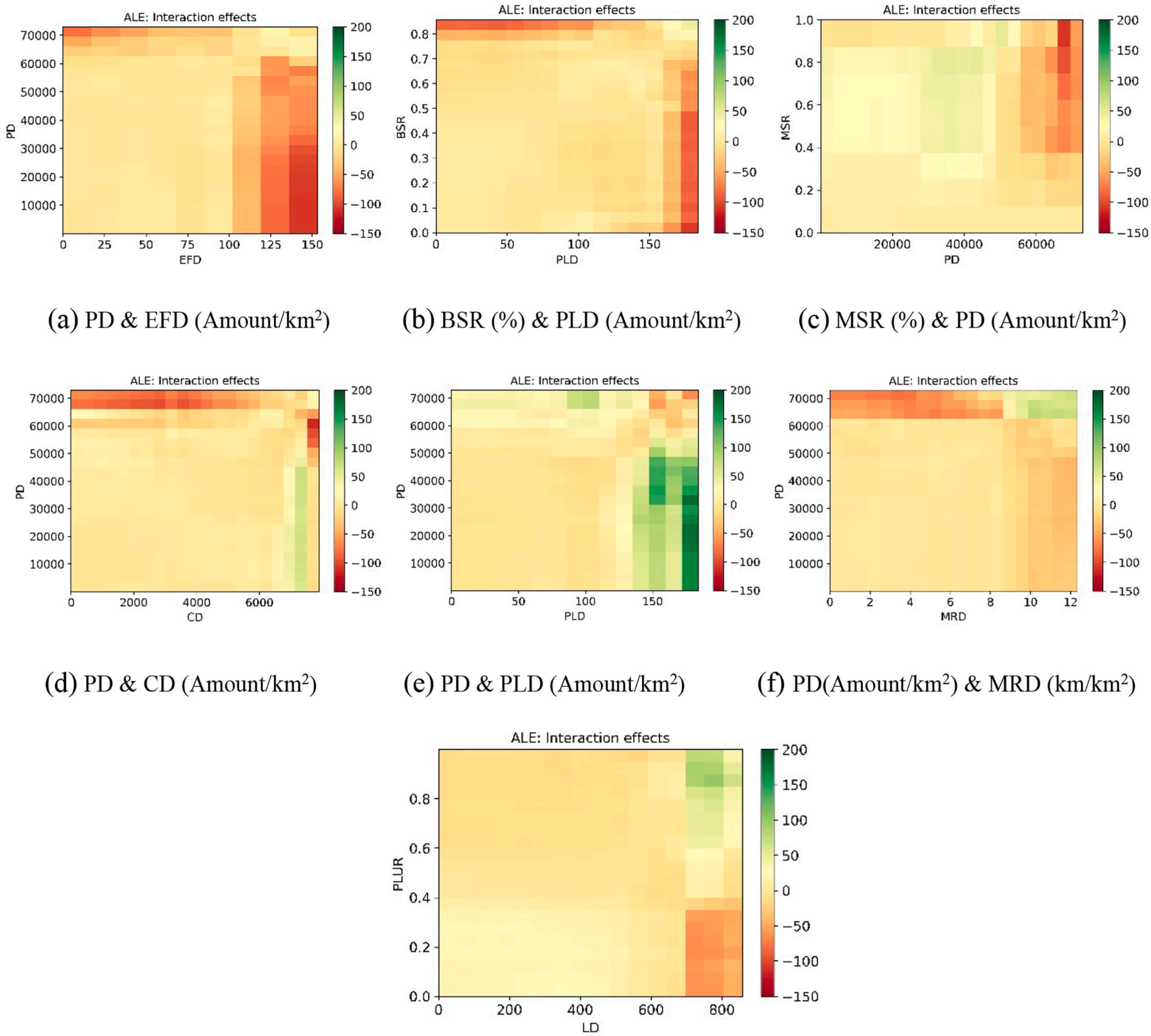
Data availability

Data will be made available on request.

Acknowledgment

The research is supported by the Area of Advance Transport at the Chalmers University of Technology and the Swedish Energy Agency (Project number: P2022-00414).

Appendix A. Appendix



(a) PD & EFD (Amount/km^2)

(b) BSR (%) & PLD (Amount/km^2)

(c) MSR (%) & PD (Amount/km^2)

(d) PD & CD (Amount/km^2)

(e) PD & PLD (Amount/km^2)

(f) PD(Amount/km^2) & MRD (km/km^2)

(g) PLUR (%) & LD (Amount/km^2)

Fig. A.1. Two-way interactions among built environment factors.

References

Amap, 2022. Point of Interest Categories in Chinese Cities.

Apley, D.W., Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. J. Royal Stat. Soc. Ser. B (Statistical Methodology) 82, 1059–1086.

Arias-Molinares, D., Romanillos, G., García-Palomares, J.C., Gutiérrez, J., 2021. Exploring the spatio-temporal dynamics of moped-style scooter sharing services in urban areas. J. Transp. Geogr. 96, 103193.

Attard, M., 2022. Active travel and sustainable transport. Communicat. Transport. Res. 2, 100059.

Awad, M., Khanna, R., 2015. Support vector regression. In: Efficient Learning Machines. Springer, pp. 67–80.

Becker, H., Balac, M., Ciari, F., Axhausen, K.W., 2020. Assessing the welfare impacts of shared mobility and mobility as a service (MaaS). Transp. Res. A Policy Pract. 131, 228–243.
Bi, H., Ye, Z., Zhu, H., 2022. Examining the nonlinear impacts of built environment on ridesourcing usage: focus on the critical urban sub-regions. J. Clean. Prod. 350, 131314.
Breiman, L., 2001. Random forests. In: Machine Learning, 45, pp. 5–32.
Brownson, R.C., Hoehner, C.M., Day, K., Forsyth, A., Sallis, J.F., 2009. Measuring the built environment for physical activity: state of the science. Am. J. Prev. Med. 36. S99-S123.e112.
Cervero, R., Sarmiento, O.L., Jacoby, E., Gomez, L.F., Neiman, A., 2009. Influences of built environments on walking and cycling: lessons from Bogotá. Int. J. Sustain. Transp. 3, 203–226.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Chen, E., Ye, Z., 2021. Identifying the nonlinear relationship between free-floating bike sharing usage and built environment. J. Clean. Prod. 280, 124281.
Cheng, L., Jin, T., Wang, K., Lee, Y., Witlox, F., 2022a. Promoting the integrated use of bikeshare and metro: a focus on the nonlinearity of built environment effects. Multim. Transportat. 1, 100004.
Cheng, L., Wang, K., De Vos, J., Huang, J., Witlox, F., 2022b. Exploring nonlinear built environment effects on the integration of free-floating bike-share and urban rail transport: a quantile regression approach. Transp. Res. A Policy Pract. 162, 175–187.
Cheng, L., Huang, J., Jin, T., Chen, W., Li, A., Witlox, F., 2023. Comparison of station-based and free-floating bikeshare systems as feeder modes to the metro. J. Transp. Geogr. 107, 103545.
Chi, R., Li, H., Shen, D., Hou, Z., Huang, B., 2022. Enhanced P-type control: indirect adaptive learning from set-point updates. IEEE Trans. Autom. Control 68 (3), 1600–1613.
Coretti Sanchez, N., Martinez, I., Alonso Pastor, L., Larson, K., 2022. On the simulation of shared autonomous micro-mobility. Communication in Transportation Research 2, 100065.
De Mol, C., De Vito, E., Rosasco, L., 2009. Elastic-net regularization in learning theory. J. Complex. 25, 201–230.
Ding, C., Cao, X., Næss, P., 2018. Applying gradient boosting decision trees to examine nonlinear effects of the built environment on driving distance in Oslo. Transp. Res. A Policy Pract. 110, 107–117.
Ding, C., Cao, X., Dong, M., Zhang, Y., Yang, J., 2019a. Nonlinear relationships between built environment characteristics and electric-bike ownership in Zhongshan, China. Transp. Res. Part D: Transp. Environ. 75, 286–296.
Ding, C., Cao, X., Liu, C., 2019b. How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify nonlinear thresholds. J. Transp. Geogr. 77, 70–78.
Ding, C., Cao, X., Yu, B., Ju, Y., 2021. Nonlinear associations between zonal built environment attributes and transit commuting mode choice accounting for spatial heterogeneity. Transp. Res. A Policy Pract. 148, 22–35.
Du, Q., Zhou, Y., Huang, Y., Wang, Y., Bai, L., 2022. Spatiotemporal exploration of the nonlinear impacts of accessibility on metro ridership. J. Transp. Geogr. 102, 103380.
Edelenbosch, O.Y., McCollum, D., Van Vuuren, D., Bertram, C., Carrara, S., Daly, H., Fujimori, S., Kitous, A., Kyle, P., Broin, E.Ó., 2017. Decomposing passenger transport futures: comparing results of global integrated assessment models. Transp. Res. Part D: Transp. Environ. 55, 281–293.
Etminani-Ghasrodashti, R., Hamidi, S., 2019. Individuals' demand for ride-hailing services: investigating the combined effects of attitudinal factors, land use, and travel attributes on demand for app-based taxis in Tehran, Iran. Sustainability 11.
Ewing, R., Cervero, R., 2010. Travel and the built environment. J. Am. Plan. Assoc. 76, 265–294.
Gao, F., Li, S., Tan, Z., Wu, Z., Zhang, X., Huang, G., Huang, Z., 2021a. Understanding the modifiable areal unit problem in dockless bike sharing usage and exploring the interactive effects of built environment factors. Int. J. Geogr. Inf. Sci. 35, 1905–1925.
Gao, K., Yang, Y., Li, A., Qu, X., 2021b. Spatial heterogeneity in distance decay of using bike sharing: an empirical large-scale analysis in Shanghai. Transp. Res. Part D: Transp. Environ. 94, 102814.
Gao, K., Yang, Y., Qu, X., 2021c. Diverging effects of subjective prospect values of uncertain time and money. Communicat. Transport. Res. 1, 100007.
Gao, K., Yang, Y., Zhang, T., Li, A., Qu, X., 2021d. Extrapolation-enhanced model for travel decision making: an ensemble machine learning approach considering behavioral theory. Knowl.-Based Syst. 218, 106882.
Giuffrida, N., Pilla, F., Carroll, P., 2023. The social sustainability of cycling: assessing equity in the accessibility of bike-sharing services. J. Transp. Geogr. 106, 103490.
Gomes, H.M., Bifet, A., Read, J., Barddal, J.P., Enembreck, F., Pfharinger, B., Holmes, G., Abdessalem, T., 2017. Adaptive random forests for evolving data stream classification. Mach. Learn. 106, 1469–1495.
Guo, Y., He, S.Y., 2020. Built environment effects on the integration of dockless bike-sharing and the metro. Transp. Res. Part D: Transp. Environ. 83, 102335.
Hasnine, M.S., Dianat, A., Habib, K.N., 2020. Investigating the factors affecting the distance travel and health conditions of e-bike users in Toronto. Transport. Res. Interdiscipl. Perspect. 8, 100265.
Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
Hooker, G., 2004. Diagnostics and Extrapolation in Machine Learning. Stanford university.
Hu, S., Xiong, C., Liu, Z., Zhang, L., 2021. Examining spatiotemporal changing patterns of bike-sharing usage during COVID-19 pandemic. J. Transp. Geogr. 91, 102997.
Huang, G., Qiao, S., Yeh, A.G.-O., 2021. Spatiotemporally heterogeneous willingness to ridesplitting and its relationship with the built environment: a case study in Chengdu, China. Transportat. Res. Part C: Emerg. Technol. 133, 103425.
Huo, J., Yang, H., Li, C., Zheng, R., Yang, L., Wen, Y., 2021. Influence of the built environment on E-scooter sharing ridership: a tale of five cities. J. Transp. Geogr. 93, 103084.
Laporte, G., Meunier, F., Wolfler Calvo, R., 2018. Shared mobility systems: an updated survey. Ann. Oper. Res. 271, 105–126.
Li, S., Lyu, D., Huang, G., Zhang, X., Gao, F., Chen, Y., Liu, X., 2020. Spatially varying impacts of built environment factors on rail transit ridership at station level: a case study in Guangzhou, China. J. Transp. Geogr. 82, 102631.
Li, A., Gao, K., Zhao, P., Qu, X., Axhausen, K.W., 2021a. High-resolution assessment of environmental benefits of dockless bike-sharing systems based on transaction data. J. Clean. Prod. 296, 126423.
Li, S., Zhuang, C., Tan, Z., Gao, F., Lai, Z., Wu, Z., 2021b. Inferring the trip purposes and uncovering spatio-temporal activity patterns from dockless shared bike dataset in Shenzhen, China. J. Transp. Geogr. 91, 102974.
Li, H., Yuan, Z., Novack, T., Huang, W., Zipf, A., 2022. Understanding spatiotemporal trip purposes of urban micro-mobility from the lens of dockless e-scooter sharing. Comput. Environ. Urban. Syst. 96, 101848.
Lin, P., Weng, J., Liang, Q., Alivanistos, D., Ma, S., 2020. Impact of weather conditions and built environment on public Bikesharing trips in Beijing. Netw. Spat. Econ. 20, 1–17.
Liu, F., Gao, F., Yang, L., Han, C., Hao, W., Tang, J., 2022a. Exploring the spatially heterogeneous effect of the built environment on ride-hailing travel demand: a geographically weighted quantile regression model. Travel Behav. Soc. 29, 22–33.
Liu, S., Zhang, X., Zhou, C., Rong, J., Bian, Y., 2022b. Temporal heterogeneous effects of land-use on dockless bike-sharing usage under transit-oriented development context: the case of Beijing. J. Clean. Prod. 380, 134917.
Ma, X., Ji, Y., Yuan, Y., Van Oort, N., Jin, Y., Hoogendoorn, S., 2020. A comparison in travel patterns and determinants of user demand between docked and dockless bike-sharing systems using multi-sourced data. Transp. Res. A Policy Pract. 139, 148–173.
McCollum, D.L., Wilson, C., Bevione, M., Carrara, S., Edelenbosch, O.Y., Emmerling, J., Guivarch, C., Karkatsoulis, P., Keppo, I., Krey, V., 2018. Interaction of consumer preferences and climate policies in the global transition to low-carbon vehicles. Nat. Energy 3, 664–673.
McKenzie, G., 2019. Spatiotemporal comparative analysis of scooter-share and bike-share usage patterns in Washington, DC. J. Transp. Geogr. 78, 19–28.
Molnar, C., 2020. Interpretable Machine Learning.
Ortúzar, J.d.D., 2021. Future transportation: sustainability, complexity and individualization of choices. Communicat. Transport. Res. 1, 100010.
Pérez-Fernández, O., García-Palomares, J.C., 2021. Parking places to moped-style scooter sharing services using GIS location-allocation models and GPS data. ISPRS Int. J. Geo Inf. 10, 230.
Radzimski, A., Dzięcielski, M., 2021. Exploring the relationship between bike-sharing and public transport in Poznań, Poland. Transp. Res. A Policy Pract. 145, 189–202.
Roman, R.C., Precup, R.E., Petriu, E.M., 2021. Hybrid data-driven fuzzy active disturbance rejection control for tower crane systems. Eur. J. Control. 58, 373–387.
Schüle, S.A., Bolte, G., 2015. Interactive and independent associations between the socioeconomic and objective built environment on the neighbourhood level and individual health: a systematic review of multilevel studies. PLoS One 10, e0123456.
Shao, Q., Zhang, W., Cao, X., Yang, J., 2022. Nonlinear and interaction effects of land use and motorcycles/E-bikes on car ownership. Transp. Res. Part D: Transp. Environ. 102, 103115.
Tang, J., Deng, C., Huang, G.-B., 2015. Extreme learning machine for multilayer perceptron. In: IEEE Transactions on Neural Networks and Learning Systems, 27, pp. 809–821.
Torrisi, V., Ignaccolo, M., Inturri, G., Tesoriere, G., Campisi, T., 2021. Exploring the factors affecting bike-sharing demand: evidence from student perceptions, usage patterns and adoption barriers. Transport. Res. Procedia 52, 573–580.
Tu, M., Li, W., Orfila, O., Li, Y., Gruyer, D., 2021. Exploring nonlinear effects of the built environment on ridesplitting: evidence from Chengdu. Transp. Res. Part D: Transp. Environ. 93, 102776.
Wagner, F., Milojevic-Dupont, N., Franken, L., Zekar, A., Thies, B., Koch, N., Creutzig, F., 2022. Using explainable machine learning to understand how urban form shapes sustainable mobility. Transp. Res. Part D: Transp. Environ. 111, 103442.
Wang, Y., Zhan, Z., Mi, Y., Sobhani, A., Zhou, H., 2022. Nonlinear effects of factors on dockless bike-sharing usage considering grid-based spatiotemporal heterogeneity. Transp. Res. Part D: Transp. Environ. 104, 103194.
Xu, Y., Yan, X., Liu, X., Zhao, X., 2021. Identifying key factors associated with ridesplitting adoption rate and modeling their nonlinear relationships. Transp. Res. A Policy Pract. 144, 170–188.
Yang, H., Zhang, Y., Zhong, L., Zhang, X., Ling, Z., 2020. Exploring spatial variation of bike sharing trip production and attraction: a study based on Chicago's divvy system. Appl. Geogr. 115, 102130.
Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models. Travel Behav. Soc. 20, 22–35.
Zhuang, C., Li, S., Tan, Z., Gao, F., Wu, Z., 2022. Nonlinear and threshold effects of traffic condition and built environment on dockless bike sharing at street level. J. Transp. Geogr. 102, 103375.