Economics of Transportation 31 (2022) 100272

Contents lists available at ScienceDirect

Economics of Transportation

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

Exploring the nonlinear impact of air pollution on housing prices: A machine learning approach

Guojian Zou [a,b], Ziliang Lai [a,b], Ye Li [a,b], Xinghua Liu [a,b], Wenxiang Li [c,*]

[a] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai, 201804, PR China
[b] College of Transportation Engineering, Tongji University, Shanghai 201804, PR China
[c] Business School, University of Shanghai for Science and Technology, 516 Jungong Road, Shanghai 200093, PR China

ARTICLE INFO

Keywords:
Air pollution
Housing prices
Gradient boosting decision tree
Machine learning
Nonlinear impact

ABSTRACT

Air pollution has profoundly impacted residents' lifestyles as well as their willingness to pay for real estate. Exploring the relationship between air pollution and housing prices has become increasingly prominent. Current research on housing prices mainly uses the hedonic pricing model and the spatial econometric model, which are both linear methods. However, it is difficult to use these methods to model the nonlinear relationship between housing price and its determinants. In addition, most of the existing studies neglect the effects of multiple pollutants on housing prices. To fill these gaps, this study uses a machine learning approach, the gradient boosting decision tree (GBDT) model to analyze the nonlinear impacts of air pollution and the built environment on housing prices in Shanghai. The experimental results show that the GBDT can better fit the nonlinear relationship between housing prices and various explanatory variables compared with traditional linear models. Furthermore, the relative importance rankings of the built environment and air pollution variables are analyzed based on the GBDT model. It indicates that built environment variables contribute 97.21% of the influences on housing prices, whereas the contribution of air pollution variables is 2.79%. Although the impact of air pollution is relatively small, the marginal willingness of residents to pay for clean air is significant. With an improvement of 1 μg/m^3 in the average concentrations of $PM_{2.5}$ and NO_2, the average housing price increases by 155.93 Yuan/m^2 and 278.03 Yuan/m^2, respectively. Therefore, this study can improve our understanding of the nonlinear impact of air pollution on housing prices and provide a basis for formulating and revising policies related to housing prices.

1. Introduction

With the rapid development of China's economy, air pollution has become an increasingly prominent problem and a great challenge that the Chinese government needs to face (Fong et al., 2020). In the four decades since the reform and expansion stage, China's urbanization rate has increased rapidly. In 2017, China's urbanization level exceeded the world average by 3.7 percent (Chuanglin, 2018). However, the continuous increases in the energy demand, industrial expansion and private car ownership in megacities such as Beijing, Shanghai, and Guangzhou have led to the serious deterioration of air quality. According to the Bulletin on the State of China's Ecology and the Environment released by the Ministry of Ecology and the Environment of China in 2019, only 157 of 337 cities at or above the prefectural level met the air quality standard in China, and 180 cities, accounting for 53.4% of all considered cities, exceeded the standard. The reasons for the frequent occurrence of extreme weather events, such as smog, in Chinese cities are comprehensive, but such events generally stem from problems related to urban development, such as extensive urban development, an unbalanced energy structure, low energy efficiency, and inefficient environmental governance (Yan et al., 2018).

Air pollution has a negative impact on human health, and it has adverse effects on property values. Ebenstein, Fan et al. noted that the concentration of PM_{10} in northern Chinese cities is significantly higher than that in the south because of the use of coal-fired heating in northern China, and the cost of energy switching would be enormous; however, the average life expectancy of residents in the north is 3.1 years that of residents in the south (Ebenstein et al., 2017). Moreover, with a focus on establishing a moderately prosperous society, living standards, quality of life, and the environment have improved, and these factors, especially air pollution, have attracted increasing research attention (Hao and Zheng, 2017). Therefore, to assess the

* Corresponding author.
E-mail addresses: 2010768@tongji.edu.cn (G. Zou), 2033402@tongji.edu.cn (Z. Lai), JamesLI@tongji.edu.cn (Y. Li), 1910893@tongji.edu.cn (X. Liu), liwx@usst.edu.cn (W. Li).

https://doi.org/10.1016/j.ecotra.2022.100272
Received 19 November 2021; Received in revised form 14 May 2022; Accepted 21 June 2022

controlling costs required to achieve effective urban air governance, we need to quantify the economic value of urban air quality.

We choose housing price as a quantitative indicator of the economic value of clean air. Housing is not only a fundamental guarantee for human life but also a family's fixed asset, and housing prices can affect the macro-economy through several channels (Huang et al., 2009). In China, with the development of real estate marketization in 1998, residents began to increasingly consider location and environmental factors when choosing houses. According to the spatial equilibrium model of urban economy, housing prices can reflect residents' preference for environmental factors, including clean air (Lan et al., 2020; Albouy, 2016). Improving air quality will increase housing's value, and this growth will be transmitted to the real estate market in the form of rising housing prices or rent (Lai et al., 2021). In addition, residents' expectations for air quality also affect current housing prices (Tra, 2010). If real estate owners expect the air quality around their houses to rise in the future, they will more highly estimate a current home's price, which will cause residents to pay more for the expected clean air. According to current relevant research results, air pollution has a significant negative impact on urban housing prices (Harrison Jr. and Rubinfeld, 1978; Smith and Huang, 1993; Koblyakova et al., 2021). Studies on the impact of pollutants on housing prices have helped increase people's awareness of environmental protection, reduced the cost of improving air quality and supported the government's formulation of public policies on air pollution (Zou, 2019). To date, these topics have received considerable attention in the research community and have been recognized as key challenges, especially concerning marginal willingness to pay (MWTP). By measuring residents' MWTP for clean air, we can also calculate the welfare cost of air pollution, or equally calculate the welfare benefit of clean air. It is worth noting that the goal of our research is to assess the relationship between input variables and target housing prices. Therefore, this research can be defined as a regression prediction analysis.

In many existing works on the relationship between air pollution concentrations and housing prices, linear regression prediction methods have been widely used; these methods consider historical air pollutant data and built environment information to forecast and analyze housing prices. The most common linear regression prediction methods include the ordinary least squares (OLS) regression analysis method and the spatial econometric method, which considers spatial heterogeneity (Zou, 2019; Casetti, 1972; Fotheringham et al., 2001; Kestens et al., 2006; Bitter et al., 2007; Chen and Jin, 2019; Mei et al., 2020; Chen et al., 2018; Sun and Yang, 2020; Trojanek et al., 2017; Cohen and Coughlin, 2008; Suryowati et al., 2021; Liu et al., 2018; Li et al., 2019b; Gilderbloom and Meares, 2020; Zhou et al., 2020; Brécard et al., 2018; Le Boennec and Salladarré, 2017; Wen et al., 2017). However, most current housing price research considers the impact of infrastructure on housing prices and lacks consideration of the impact of air pollutants, including $PM_{2.5}$ and NO_2, on housing prices. Second, the existing methods for simulating the relationship between impact factors and housing prices are generally based on linear regression; therefore, it is difficult to accurately fit the nonlinear relationship between impact factors and housing prices. Based on the above problems, we must urgently develop a new method that can effectively fit the nonlinear relationship between impact factors and housing prices and quantify the MWTP for clean air in cities.

Traditionally, the impact of air pollution on housing prices can be regarded as a regression problem, that is, involving the modeling and analysis of relevant historical data, including data on pollutants and the built environment. Many studies have shown that in the process of housing price changes, the interactions and effects of related factors are complex (Kumar and Imam, 2013; Alvanchi et al., 2020; Huang and Yi, 2010). These complex interactions and effects must be taken into account for housing price prediction. Moreover, when studying the impact of pollutants on housing prices, it is important to fully consider the features of pollutants and to correlate pollutants with other influencing factors.

Machine learning algorithms have been widely used in regression problems in recent years and can effectively fit the relationship between independent variables and dependent variables. Tree-based boosting learning algorithms, especially gradient boosting decision trees (GBDTs), which combine multiple weak learning regression trees, have been widely used in solving classification and regression tasks. Compared with traditional regression prediction models, the GBDT is more stable and interpretable and shows better predictive performance by combining multiple weak learning regression trees (Zhang and Haghani, 2015; Tu et al., 2021).

The main goals of this study are to explore the nonlinear impacts of air pollution on housing prices and identify the relative importance of explanatory variables. This study can provide scientific support for the development of region-specific air quality improvement programs and policies. The main contributions of this study are as follows.

(1) A housing prices prediction model is established based on a machine learning approach, the GBDT model, considering various independent variables. It can capture the complex nonlinear relationships between independent variables and dependent variables, compared with conventional regression models.
(2) The GBDT model is interpretable. The relative importance of independent variables can be ranked based on the GBDT model. Besides, the nonlinear relationships between independent variables and housing prices can be visualized based on partial dependence plots of the GBDT model.
(3) Given the trained GBDT model, the marginal willingness of residents to pay for clean air in each census tract of Shanghai is measured. Then, the economic loss caused by air pollution in Shanghai can be estimated.

The remainder of this paper is organized as follows. In Section 2, we summarize the previous related studies. Section 3 describes the data and variables used in this study. The GBDT model working process is described in Section 4. Section 5 presents the experiments and the results in detail. Finally, we summarize the main findings, present policy implications, and propose future research directions in Section 6.

2. Related work

Recent studies on the complex relationship between air pollutants and housing prices can be divided into the traditional hedonic pricing method and the spatial econometric method (Zou, 2019; Casetti, 1972; Fotheringham et al., 2001; Kestens et al., 2006; Bitter et al., 2007; Chen and Jin, 2019; Mei et al., 2020; Chen et al., 2018; Sun and Yang, 2020; Trojanek et al., 2017; Cohen and Coughlin, 2008; Suryowati et al., 2021; Liu et al., 2018; Li et al., 2019b; Gilderbloom and Meares, 2020; Zhou et al., 2020; Brécard et al., 2018; Le Boennec and Salladarré, 2017; Wen et al., 2017; Lai et al., 2021; Yang et al., 2022).

The traditional hedonic pricing method holds that real estate is composed of many different characteristics, and that real estate price is determined by the utility that all of these characteristics give to people (Yang et al., 2022, 2020b). Due to each feature's different quantity and combination modes, real estate prices vary (Zou, 2019; Chen et al., 2018). In this context, the changes in environmental quality parameters, such as air pollution, can easily lead to fluctuations in people's expected house prices. Ordinary least squares (OLS) regression is widely used to solve the hedonic pricing model in many studies (Zou, 2019; Chen and Jin, 2019; Mei et al., 2020; Chen et al., 2018; Sun and Yang, 2020; Gilderbloom and Meares, 2020; Brécard et al., 2018). This approach truncates the dependent variable and leads to estimation errors. Meanwhile, the basic assumption of the OLS model is that housing price data in different regions are spatially independent and static. However, due to the variations among different regions, housing prices and the factors that influence these prices exhibit considerable spatial heterogeneity. For example, residents who live near factories pay more attention to air pollution, while those living in the city center pay more attention to location factors such as transportation, shopping, medical treatment and education available nearby (Lai et al., 2021). Thus, the OLS method is not applicable because it ignores spatial changes (Zou, 2019). Therefore, to measure the impact of pollutants on housing prices in various census tracts, an accurate estimation model is required.

Recently, when studying the impact of pollutants on the value of housing prices in different areas, some studies have considered the problem of spatial heterogeneity. Notably, some researchers have used spatial econometric models to explore housing prices (Trojanek et al., 2017; Cohen and Coughlin, 2008; Suryowati et al., 2021; Liu et al., 2018; Zhou et al., 2020; Le Boennec and Salladarré, 2017; Wen et al., 2017). There are several commonly used methods for assessing the impact of pollutants on housing prices based on spatial econometric approaches: the spatial autoregressive model (SAR) (Trojanek et al., 2017; Cohen and Coughlin, 2008; Suryowati et al., 2021), the spatial Durbin model (SDM) (Suryowati et al., 2021; Liu et al., 2018; Zhou et al., 2020), the spatial error model (SEM) (Liu et al., 2018; Le Boennec and Salladarré, 2017), moving window regression (MWR) (Li et al., 2019b), and geographically weighted regression (GWR) (Zou, 2019; Li et al., 2019b; Wen et al., 2017). These methods have been improved and upgraded, and their performance is better than OLS regression (Zou, 2019; Casetti, 1972; Fotheringham et al., 2001; Kestens et al., 2006; Bitter et al., 2007; Trojanek et al., 2017; Cohen and Coughlin, 2008; Suryowati et al., 2021; Liu et al., 2018; Zhou et al., 2020; Le Boennec and Salladarré, 2017; Wen et al., 2017; Yang et al., 2021; Li et al., 2021b). The GWR model is a statistical regression model that was proposed by British scholars to consider the spatial heterogeneity of variables, and it is a type of spatial econometric model (Casetti, 1972; Fotheringham et al., 2001; Kestens et al., 2006; Bitter et al., 2007). It is generally believed that the GWR model is an improvement to traditional spatial regression, and various studies have shown that the GWR model is the best approach for exploring the spatial heterogeneity of housing prices and the corresponding influential factors. In addition, GWR is superior to SAR, SDM, SEM, and MWR in terms of generalization ability and prediction accuracy (Zou, 2019; Bitter et al., 2007; Yang et al., 2020a). There is a highly nonlinear relationship between the built environment, pollutants and housing prices, and the nonlinear relationships among data types are particularly important for assessing the relationship between pollutants and housing prices. However, GWR is a spatial regression method, and it is difficult to simulate the nonlinear relationships between independent and dependent variables: in this case, the nonlinear relationship between the built environment, pollutants and housing prices.

To explore the complex nonlinear relationship between the built environment, pollutants and housing prices and to identify the importance of variables, this study proposes a tree-based boosting learning algorithm, the GBDT model, to predict housing prices. The boosting learning algorithm is one of the most common and efficient algorithms in machine learning, and it can combine multiple weak learning models to improve the prediction performance of a model (Chung, 2013). Unlike other ML methods, the boosting learning algorithm can reflect the interaction between input variables and prediction models and can identify the relative importance of input variables (Zhang and Haghani, 2015). The GBDT approach is a highly interpretable, stable, relatively novel and accurate method in the field of machine learning that can effectively consider different types of input variables and identify the impact of the built environment and pollutants on housing prices in a given census tract; thus, the GBDT model outperforms classical methods.

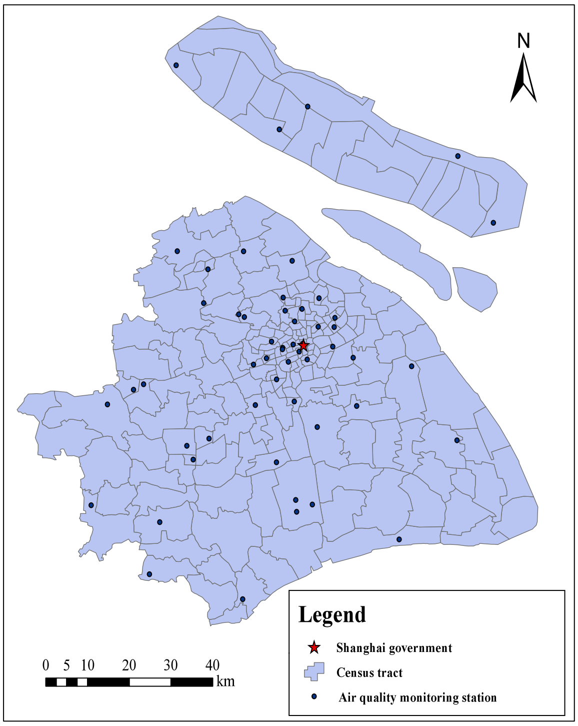

Fig. 1. Study area and distribution of air quality monitoring stations.

3. Data

3.1. Study area

Since the reform and opening up that began more than 40 years ago, China's economy has experienced unprecedented growth. People increasingly worry that massive environmental costs, including air and water pollution, have been paid to achieve economic "miracles" (Matus et al., 2012). Shanghai, as China's economic, financial, trade, science and technology center, is also the most populous city in China. As of 2020, the city had 16 districts with a total area of 6340.5 square kilometers, a built-up area of 1237.85 square kilometers, a permanent population of 24,281,400, an urban population of 21,391,900, and an urbanization rate of 88.10%. Housing prices in Shanghai have been rising annually, with the sale price of commercial housing increasing from an average of 3866 Yuan per square meter in 2001 to 23,804 Yuan per square meter in 2017 (Chen et al., 2017). As one of the largest cities in China, Shanghai can reflect the MWTP of people in the eastern and southern coastal areas of China. Thus, we choose Shanghai as the study area. In addition, to make our study more detailed, Shanghai is divided into 218 census tracts based on the sixth census of administrative divisions, as shown in Fig. 1.

We focus on the air quality around residential areas. The data of air quality monitoring stations, including $PM_{2.5}$, were released by the Chinese government only after 2012 (Fan et al., 2019). Fig. 2 shows the daily mean concentrations of SO_2, NO_2, and PM_{10} from 2001 to 2019 in Shanghai. Overall, the air quality in Shanghai has improved over time, mainly due to the active governance of the Chinese government. Specifically, in 2019, the daily average concentration of SO_2 in Shanghai was 7 $\mu g/m^3$ (attaining the Grade I national standards for air quality); additionally, that of NO_2 was 42 $\mu g/m^3$ (exceeding the Grade II national standard of 2 $\mu g/m^3$), and that of PM_{10} was 45 $\mu g/m^3$ (attaining the Grade II national standards for air quality).

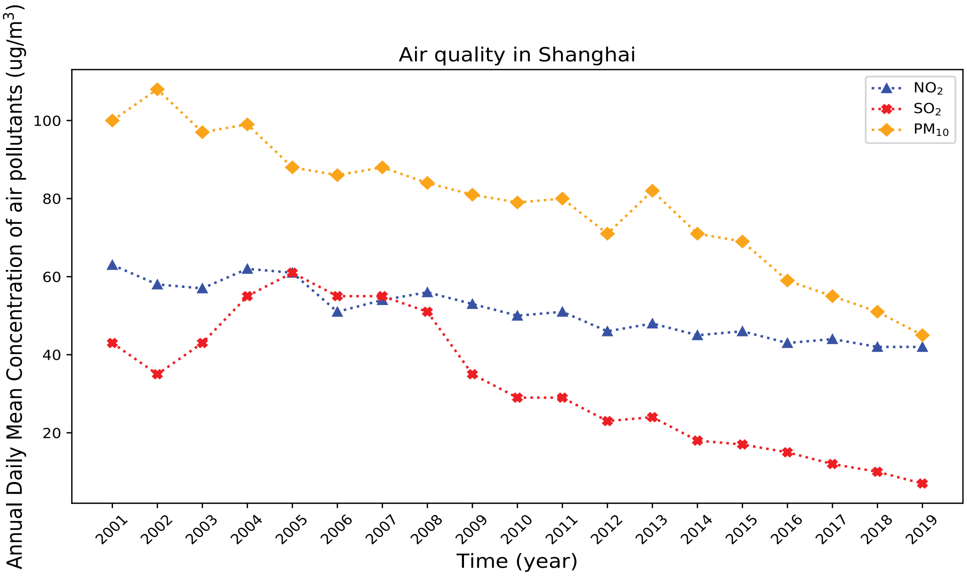

Fig. 2. Air quality in Shanghai from 2001–2019.

3.2. *Data description and processing*

The data used in this paper include housing prices, air quality, and built environment data from Shanghai and cover the period from January to December 2018. Housing price information from 27,608 residential areas in Shanghai was used, and these data were obtained from a platform for real estate transactions in Shanghai (http://cd.lianjia.com). The specific housing price data include the address, average housing prices, construction time, number of housing units, and latitude and longitude coordinates of each residential area.

The air quality data were from the Shanghai Municipal Bureau of Ecology and the Environment. We collected data from 59 air quality monitoring sites in Shanghai (Fig. 1). These data included the hourly mean concentrations of pollutants, monitoring time, and current air pollution levels. Air quality data can be used to calculate the annual average air pollution level in different census tracts in Shanghai.

The built environment can be reflected by the 5Ds, namely, density, design, diversity, distance to transit and destination accessibility, and the selected indicators include the bus stop density, land use mix, proportion of educational institutions, road network density, park accessibility, living facility accessibility, etc (Li et al., 2021a). Population, urban road network and point of interest (POI) data constitute the built environment data. The population data were obtained from the sixth census of China. The urban road network data for Shanghai were acquired from Open Street Map (OSM). POI data were obtained from Baidu, one of the largest Chinese web map service applications. The POI data included 14 types of locations, including schools, restaurants, hospitals, parks, supermarkets, transportation facilities, etc. (Li et al., 2017). The number of POIs was 603,085, and the basic information included the name, type, location, latitude, and longitude of each POI.

The minimum research units in this study include 218 census tracts in Shanghai, so the housing prices, air quality, and built environment data needed to be processed. As shown in Fig. 3, the processing method mainly involved four steps: data cleaning, data analysis, data extraction, and variable calculation.

3.3. *Variable descriptions*

According to related research (Chen et al., 2018; Li et al., 2019a), we consider various factors that may affect housing prices and then determine the dependent and independent variables to retain in this paper. We describe all the selected variables in detail and give the statistical analysis results for each variable, as shown in Table 1. Six air quality components are considered: CO, O_3, SO_2, NO_2, PM_{10} and $PM_{2.5}$. The built environment variables include 17 indicators related to the 5Ds. In particular, land-use diversity is measured by the entropy index, and the corresponding values range between 0.0 and 1.0, where 0.0 shows that the land use is singular and homogeneous and 1.0 suggests that all land types are relatively uniform (Cervero and Kockelman, 1997). In addition, we selected the Shanghai government as the center of Shanghai to compute the straight-line distance from the city center to each district. This area is the geographic center, transportation hub, and heart of Shanghai. Furthermore, to eliminate the influence of dimension differences among variables, the min–max normalization method was used to standardize the data, which converts the ranges of all variables to [0.0, 1.0].

4. Methods

4.1. *Methodology*

4.1.1. *Gradient boosting decision trees*

In our study, the high-performing GBDT method is incorporated into the model to predict housing prices. We assume that $g(x)$ is an approximation function that fits y based on the input variable x, and uses the squared difference function as a loss function for evaluating the error between the observed value and the predicted value (Ke et al., 2017):

$$Loss\left(g\left(x\right),y\right)=\left(g\left(x\right)-y\right)^2 \tag{1}$$

We define the number of splits of each regression tree as K, and each tree divides the input space into K regions $\{R_{1n},\dots,\ R_{Kn}\}$ that do not coincide with each other. In this case, each regression tree predicts a constant value c_{kn} for the region R_{kn}, and the final output of each regression tree is obtained by the following cumulative equation (Chung, 2013):

$$t_n\left(x\right)=\sum_{k=1}^{K}c_{kn}I\left(x\in R_{kn}\right),\ where\ I=1\ if\ x\in R_{kn};\ I=0\ otherwise \tag{2}$$

Using the training set $\{x_i,y_i\}_1^D$, the GBDT needs to iteratively construct N different individual regression trees $\{t_1\left(x\right),\dots,\ t_N\left(x\right)\}$. When constructing each tree, the approximation function $g_n\left(x\right)$ needs to be

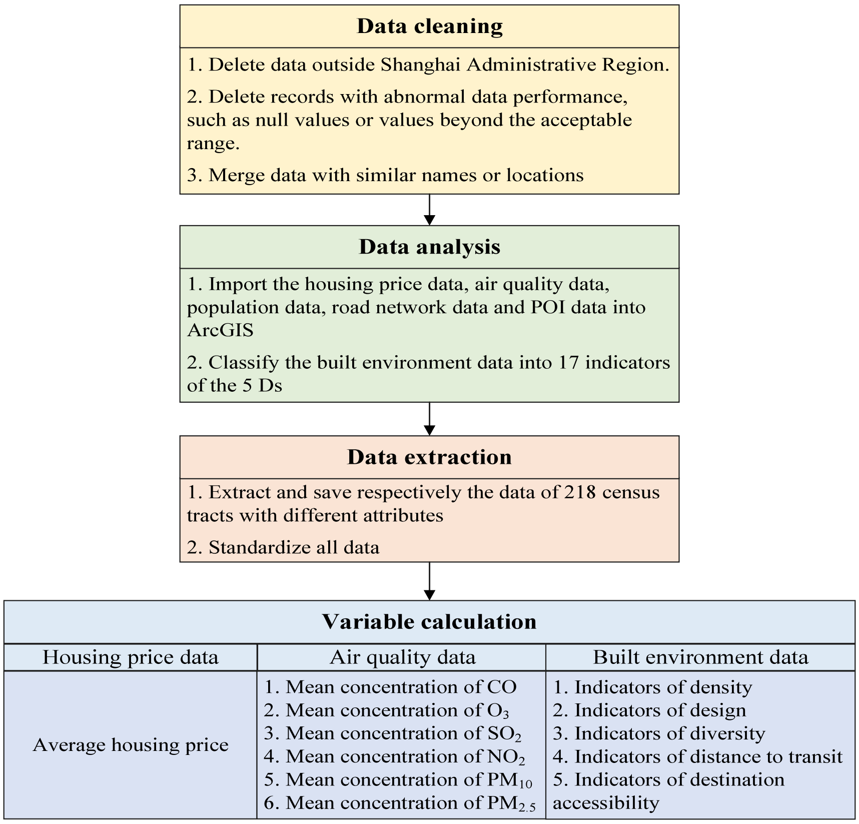
Data cleaning
1. Delete data outside Shanghai Administrative Region.
2. Delete records with abnormal data performance, such as null values or values beyond the acceptable range.
3. Merge data with similar names or locations
Data analysis
1. Import the housing price data, air quality data, population data, road network data and POI data into ArcGIS
2. Classify the built environment data into 17 indicators of the 5 Ds
Data extraction
1. Extract and save respectively the data of 218 census tracts with different attributes
2. Standardize all data
Variable calculation
Housing price data
Air quality data
Built environment data
Average housing price
1. Mean concentration of CO
2. Mean concentration of O3
3. Mean concentration of SO2
4. Mean concentration of NO2
5. Mean concentration of PM10
6. Mean concentration of PM2.5
1. Indicators of density
2. Indicators of design
3. Indicators of diversity
4. Indicators of distance to transit
5. Indicators of destination accessibility

Fig. 3. Diagram of data processing. The data preprocessing process is shown from top to bottom, and includes four steps: data cleaning, data analysis, data extraction, and variable calculation.

Table 1
Variable definitions and statistics.

Variables	Description	Mean	St. dev.	Min	Max
Dependent variable					
Housing price	Average housing price (Yuan/m^2)	50,090.00	24,235.00	11,061.00	111,382.00
Air pollutant					
CO	Mean concentration of CO (μg/m^3)	0.59	0.04	0.45	0.71
O_3	Mean concentration of O_3 (μg/m^3)	53.07	4.93	42.20	68.67
SO_2	Mean concentration of SO_2 (μg/m^3)	6.67	0.55	5.23	8.34
NO_2	Mean concentration of NO_2 (μg/m^3)	68.33	10.02	64.14	72.80
PM_{10}	Mean concentration of PM_{10} (μg/m^3)	59.19	2.09	54.23	64.28
$PM_{2.5}$	Mean concentration of $PM_{2.5}$ (μg/m^3)	43.19	1.37	38.86	47.84
Built environment					
PopD	Population density (number/km^2)	15 314.00	15 971.00	68.06	68 129.00
MetroSD	Metro station density (number/km^2)	0.30	0.40	0.00	1.96
MetroLD	Metro line density (km/km^2)	1.00	1.34	0.00	1.96
BusD	Bus stop density (numbers/km^2)	5.46	3.99	0.19	21.98
RoadD	Road network density (km/km^2)	8.96	6.24	0.34	39.11
PLD	Parking lot density (number/km^2)	29.52	34.78	0.00	150.25
Edu	Percentage of educational service POIs (%)	0.32	0.47	0.00	3.63
Leis	Percentage of leisure POIs (%)	0.41	0.34	0.01	2.17
Stad	Percentage of stadium POIs (%)	0.51	0.42	0.00	2.20
Med	Percentage of medical institution POIs (%)	1.26	0.33	0.00	2.37
Park	Percentage of park POIs (%)	0.77	0.44	0.00	3.36
Tour	Percentage of tourist attraction POIs (%)	1.28	0.96	0.00	11.31
Super	Percentage of shopping mall and supermarket POIs (%)	0.51	0.36	0.00	1.90
Tbh	Percentage of telecom business POIs (%)	1.01	0.38	0.00	2.58
Res	Percentage of restaurant POIs (%)	0.57	0.36	0.00	2.20
Landuse	Entropy index of the land use mix (–)	0.23	0.17	0.06	0.95
Discent	Straight-line distance from city center (km)	19.78	15.70	0.18	58.64

updated again, where $g_n(x)$ and the gradient descent step length ρ_n can be described as follows (Ding et al., 2016):

$$g_n(x) = g_{n-1}(x) + \rho_n \sum_{k=1}^{K} c_{kn} I\left(x \in R_{kn}\right) \tag{3}$$

$$\rho_n = \underset{\rho}{argmin} \sum_{i=1}^{D} Loss\left(g_{n-1}\left(x_i\right) + \rho \sum_{k=1}^{K} c_{kn} I\left(x \in R_{kn}\right), y_i\right) \tag{4}$$

Therefore, each region R_{kn} obtains optimal best individual α_{kn}, so c_{kn} can be (Zhang and Haghani, 2015). Eq. (3) can be updated as follows:

$$g_n(x) = g_{n-1}(x) + \sum_{k=1}^{K} \alpha_{kn} I\left(x \in R_{kn}\right) \tag{5}$$

and the optimal α_{kn} can be obtained as follows:

$$\alpha_{kn} = \underset{\alpha}{argmin} \sum_{x_i \in R_{kn}} Loss\left(g_{n-1}\left(x_i\right) + \alpha, y_i\right) = \underset{\alpha}{argmin} \sum_{x_i \in R_{kn}} \left(\alpha - \widetilde{y}_i\right)^2 \tag{6}$$

where

$$\widetilde{y}_i = -\left[\frac{\partial Loss\left(g\left(x_i\right), y_i\right)}{\partial g\left(x_i\right)}\right]_{g(x)=g_{n-1}(x)} \tag{7}$$

The GBDT builds the model by continuously adding weak regression trees to fit the residuals and updates the model parameters by minimizing the output value of the loss function. To improve the predictive performance of the model and prevent the occurrence of overfitting, the GBDT applies a shrinkage strategy (Schonlau, 2005). Shrinkage, also known as the learning rate, is used to measure the contribution of each regression tree by introducing the weight factor η $(0 < \ \eta \ \leq 1)$, as shown below:

$$g_n(x) = g_{n-1}(x) + \eta \cdot \sum_{k=1}^{K} \alpha_{kn} I\left(x \in R_{kn}\right), \ where \ 0 < \ \eta \ \leq 1 \tag{8}$$

where the smaller the value of η is, the greater the shrinkage. The GBDT uses a shrinking strategy to avoid over-fitting problems by reducing the weight of each regression tree. A smaller shrinkage value is beneficial to minimize the loss function; however, the entire GBDT model may need to add a large number of regression trees. Another hyperparameter, the complexity of the tree, refers to the number of splits K, which is used to fit each regression tree; it represents the depth of the input space interaction of the regression tree. Increasing the complexity of the tree can help capture more complex interactions among input variables and improve the strength and performance of the GBDT. According to the shrinkage value and the complexity of the regression tree, we can determine the number of best regression trees by evaluating the degree of fit between the GBDT model and the data set (Schonlau, 2005). The optimal performance of the GBDT depends on selecting the optimal combination of the learning rate and tree complexity.

4.1.2. Relative importance of influential factors

For regression tasks, input variables rarely have the same correlation with the prediction target, and the weights of the influence of input variables on the model output are usually substantially different (Friedman and Meulman, 2003). Therefore, when predicting the response, it is particularly important to determine the relative importance or weight of each independent variable. However, accuracy and interpretability, as the two fundamental objectives of regression analysis algorithms, are not always consistent (Guelman, 2012). In contrast to general modeling approaches, such as hidden Markov models (HMMs), naive Bayes models (NBM), and deep learning networks, the GBDT method can identify and calculate the influence of input variables on the prediction target.

For a single regression tree T, (Loh, 2011) proposed an approximate method to measure the relative importance of the input variable x_j in the prediction task, as shown below:

$$\mathrm{W}_j^2(T) = \sum_{v=1}^{K-1} \hat{\tau}_t^2 \mathrm{W}\left(v = x_j\right) \tag{9}$$

where the calculation is performed on the non-terminal node v of the K-terminal node tree T, x_j is the variable associated with the split node v, and $\hat{\tau}_v^2$ represents the corresponding empirical improvement in the form of squared error using the variable x_j as the result of the non-terminal split node v. We obtain the set of regression trees $\{T_n\}_1^N$ through the gradient boosting method, so we can generalize Eq. (9) by calculating the average of all additive trees:

$$\mathrm{W}_j^2 = \frac{1}{N} \sum_{n=1}^{N} \mathrm{W}_j^2\left(T_n\right) \tag{10}$$

4.2. External factor module

In real-world research, housing prices have different inherent patterns and regularities, but they are easily affected by other factors from different fields. For instance, housing prices are affected by pollutants and the built environment; e.g., high $PM_{2.5}$ or NO_2 levels may cause changes in housing prices, and convenient transportation will influence housing prices.

To consider the influence of these external factors, we propose a commonly used solution. First, we combine all the external factors considered into a vector (e.g., $PM_{2.5}$, CO, population density, bus stop density, metro station density, etc.). Second, we solve the existing problem related to the high linear correlations among explanatory factors. Then, we use the nonlinear GBDT model as a feature extractor to learn the relationship between external information and housing prices.

4.3. Analysis of multicollinearity

Multicollinearity refers to the existence of a high degree of linear correlation among explanatory variables, resulting in the model regression results being nonideal or difficult to obtain. To solve this problem, we adopted the variance inflation factor (VIF), which is an indicator of the degree of multicollinearity of a model (Alin, 2010; Vu et al., 2015). Specifically, to a certain extent, the fitting effect of the model is closely related to the VIF of the explanatory variable. In general, when the VIF value of an explanatory variable is greater than 7.5, multicollinearity with other explanatory variables exists, and this variable should be removed from the model (Wu et al., 2019). Therefore, VIF was used in this study to identify multicollinearity. The VIF_j of an input variable x_j is based on the linear relationship between the input variable x_j and the other input variables $\{x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m\}$, as expressed in (11) (Alin, 2010).

$$\mathrm{VIF}_j = \frac{1}{\left(1 - \mathrm{R}_j^2\right)} \tag{11}$$

where R_j^2 is the coefficient of determination of the regression of x_j for all other independent variables $\{x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m\}$ in the dataset, and m represents the number of independent variables.

4.4. Evaluation metrics

To compare the results of the GBDT model to those of regression-based methods, we introduce the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) to evaluate model performance.

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=1}^{M}\left(y_i - \hat{y}_i\right)^2}{M}} \tag{12}$$

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{M} \sum_{i=1}^{M} \left|y_i - \hat{y}_i\right| \tag{13}$$

$$\mathrm{R}^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{M}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{M}\left(y_i - \bar{y}\right)^2} \tag{14}$$

The RMSE and MAE results reflect the difference between the predicted values of housing prices and the corresponding manually verified actual values in the dataset. R^2 is a statistic that provides some information about the goodness of fit of a model. High R^2 and low RMSE and MAE values indicate more accurate prediction performance. In the above equations, y_i is the observed value, $\hat{y}_i$ denotes the predicted value, $\bar{y}$ represents the average of the observed values, and M is the dataset size.

4.5. MWTP

In this study, $MWTP_j$ represents the house price residents are willing to pay more when the target variable value x_j increases by 1 unit. For example, to increase the $PM_{2.5}$ of 1 μg/m^3, residents are willing to pay −155.93 Yuan/m^2 more for the house price. The MWTP equation is as follows:

$$MWTP_j = \frac{1}{M}\sum_{i=1}^{M}\left(g\left(x'_i\right) - g\left(x_i\right)\right), \left(x'_{i,j} = x_{i,j} + 1\right) \quad (15)$$

where i denotes the census tract index, j denotes the target variable index, x_i represents the input built environment and pollutant variables $\{x_{i,1}, x_{i,2}, \ldots, x_{i,j}, \ldots, x_{i,m}\}$ of census tract i, m denotes the number of variables, and $g(x)$ represents the GBDT function.

5. Experimental results and discussion

5.1. Settings

We use multiple cross-validations on the dataset so that the GBDT obtains the best hyperparameters and robust results. The parameter selection process can be described in three steps as follows. First, we set three parameter sets, including the number of trees, shrinkage values, and tree complexity sets. Second, one parameter is selected from each parameter set as the initial parameter of GBDT. This study uses the k-fold cross-validation method (Fushiki, 2011); $k - 1$ groups of samples are selected as the training set ($k = 10$), the remaining 1 group of samples is used as the validation set, the cycle is repeated k times, and the average error on the validation set is recorded. For the third step, we repeat the second step until three parameter sets are traversed, and we then select the parameter combination with the minor average error on the validation set as the final GBDT parameters. We pre-specify the hyperparameter range of the GBDT model: numbers of trees (200, 150, 100, 50, and 10), shrinkage values (0.005, 0.05, 0.01, 0.1, and 0.5) and tree complexities (1, 2, 3, 5, and 7) based on all the results of previous experiments. After many experiments, we found that the GBDT performed best with the following settings: set the number of trees, shrinkage value and tree complexity were set as 100, 0.1 and 2, respectively.

5.2. Explanatory variable selection

First, the Toolbox tool in ArcGIS was used to perform multicollinear tests on all the independent variables in Table 1, and the variables with VIF values greater than 7.5 were removed. Then, we adopted stepwise OLS regression to test the impact of characteristic variables on house prices and selected the variables with significant performance. Finally, after the stepwise regression, only 10 characteristic variables remained: the mean concentration of $PM_{2.5}$, the mean concentration of NO_2, metro station density, road network density, parking lot density, the percentage of educational service POIs, the percentage of stadium POIs, the percentage of medical institution POIs, the percentage of park POIs, and straight-line distance from the city center.

The results of OLS regression for these characteristic variables are shown in Table 2. Notably, $PM_{2.5}$ and NO_2 have a significant negative impact on housing prices, and this result is consistent with the findings in Refs. (Chen and Chen, 2017) and (Chen and Lee, 2020); the coefficients of other built environment variables are generally consistent with the theoretical and expected values. Specifically, the farther a residential property is from the city center, the lower the housing price is. The greater the numbers of subway stations, schools, and other infrastructure types in a given area are, the higher the housing price is. In addition, the VIF values of these 10 characteristic variables are all relatively small, indicating that the collinearity among the variables is very weak.

Table 2
OLS regression results.

Variables	Coefficient	t	P-value	VIF
Intercept	0.599	10.702	<0.000001***	/
NO_2	−0.126	−4.216	<0.000001***	1.859
$PM_{2.5}$	−0.077	−2.749	0.0065***	1.649
MetroSD	0.314	5.067	<0.000001***	2.762
RoadD	0.709	8.576	<0.000001***	3.039
PLD	0.120	2.149	0.0328**	2.141
Edu	0.151	1.982	0.0489**	2.712
Stad	0.202	2.541	0.0118**	5.288
Med	0.131	2.109	0.0347**	3.405
Park	0.215	2.474	0.0204**	1.984
Discent	−0.509	−12.807	<0.000001***	3.237
Area coverage		218		
Sample period		2018–2019		

Notes: See Table 1 for variable definitions. The standard deviations are given in parentheses. ***, **, and * denote the $p < 0.01$, $p < 0.05$, and $p < 0.1$ statistical significance levels, respectively.

Table 3
Performance of the GBDT model and comparison models.

Metrics	OLS	GWR	GBDT	Percentage improvement
R^2	0.803	0.876	0.975	+11.30%
MAE	0.085	0.080	0.017	−78.75%
RMSE	0.107	0.099	0.028	−71.72%

5.3. Performance of the GBDT model

Housing prices in different census tracts of a city generally exhibit discrete irregularities, as shown in Fig. 4 that air pollution in Shanghai is heterogeneous and is mainly concentrated in the outer suburbs near the heavy industrial areas and in the city center with heavy traffic. For example, the highest average concentrations of SO_2 and PM_{10} were found in Baoshan, Jinshan, Yangpu and Jiading districts because these areas contain large numbers of polluting factories, such as electricity, steel, and automobile factories. In addition, large trucks driving in these areas during the day contribute to air pollution. Due to the large population and heavy traffic, the areas with the highest average concentrations of NO_2 and $PM_{2.5}$ are mainly located in the inner-ring area and the region directly to its north. However, unlike the distribution of air pollutant concentrations, the density gradually decreases from the city center to the suburbs for the built environment, including parking lots, parks, educational services, and metro stations. So it is difficult to use a linear model to predict housing prices in different census tracts.

According to current research, housing prices are affected by the built environment and air pollutant factors. In this study, we fully considered the nonlinear relationship between explanatory variables and housing prices. We select four explanatory variables as representatives to show the relationship between explanatory variables and housing prices. The four variables are Straight-line distance from the city center, Road network density, NO_2, and $PM_{2.5}$. As Fig. 5 shows, the relationship between housing prices and explanatory variables is nonlinear, which is extremely important for studying pollutants' impact on housing prices. Therefore, these nonlinear relations must be considered in modeling applications.

This paper uses a GBDT model to fit the nonlinear relationship between explanatory variables and housing prices. To clearly show the advantages of the GBDT model, Table 3 lists the RMSE, MAE, and R^2 values for each model based on the whole training set.

As shown in Table 3, the R^2, MAE and RMSE values for GBDT model predictions are 0.975, 0.017 and 0.028, respectively. The corresponding values for GWR, a traditional spatial linear regression model, are 0.876, 0.080 and 0.099. Thus, with the proposed method, R^2 is improved by 11.30%, and the MAE and RMSE decrease by 78.75% and 71.72%, respectively. The experimental results verify that the

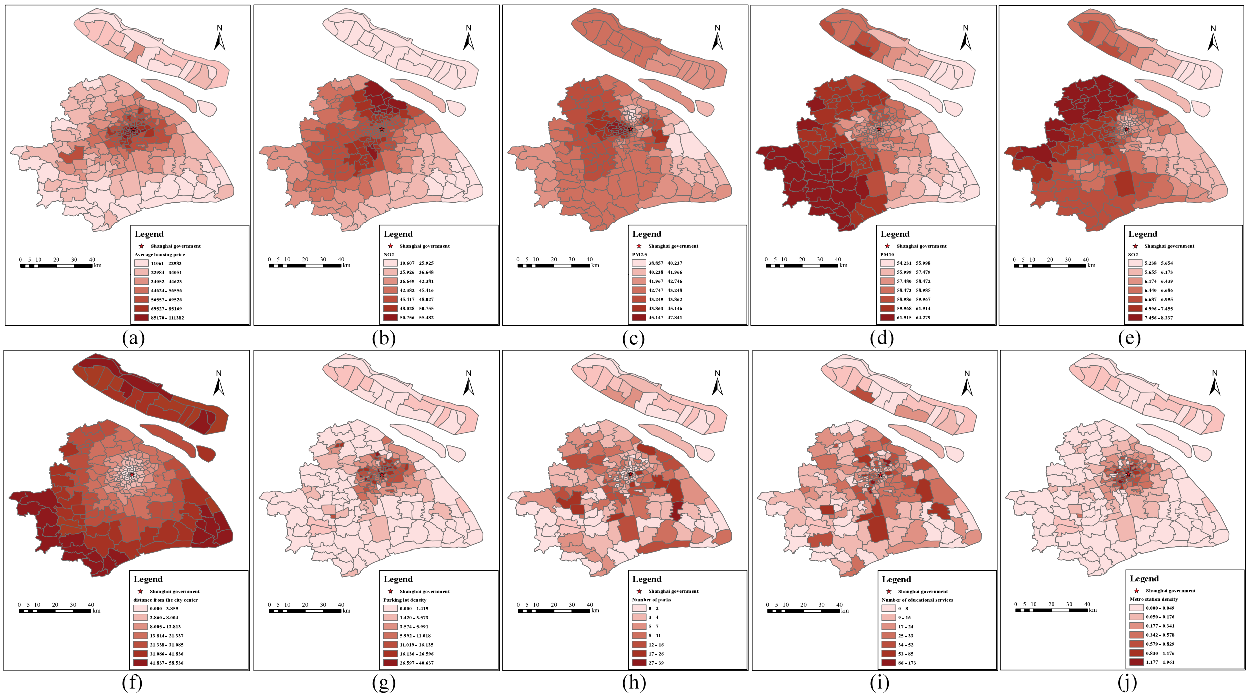

Fig. 4. Distributions of some explanatory variables in Shanghai: (a) average housing price; (b) NO_2; (c) $PM_{2.5}$; (d) PM_{10}; (e) SO_2; (f) distance from the city center; (g) parking lot density; (h) number of parks; (i) number of educational services; (j) metro station density.

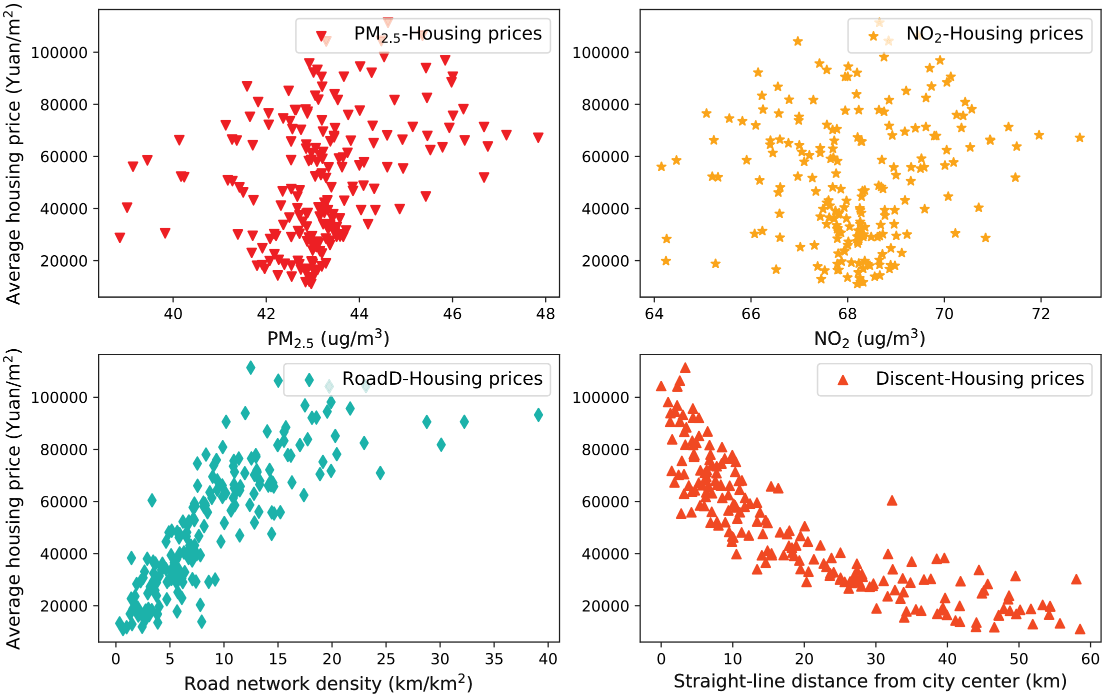

Fig. 5. Fitting trends of explanatory variables and housing prices. The *x*-axis value represents the variable value, and the *y*-axis value represents the average housing price.

GBDT model is more suitable for analyzing nonlinear effects than the OLS and GWR models; additionally, the relationship between input variables and target output variables can easily be identified with the GBDT method. Finally, by controlling the input variables, the proposed approach provides a solid theoretical basis to quantify the MWTP for clean air in Shanghai.

5.4. Relative importance of explanatory variables

In the task of housing price prediction, the relative importance of explanatory variables is extremely important to the accuracy and interpretability of the model. Fig. 6 shows the importance and ranking of the explanatory variables. It should be noted that the sum of all

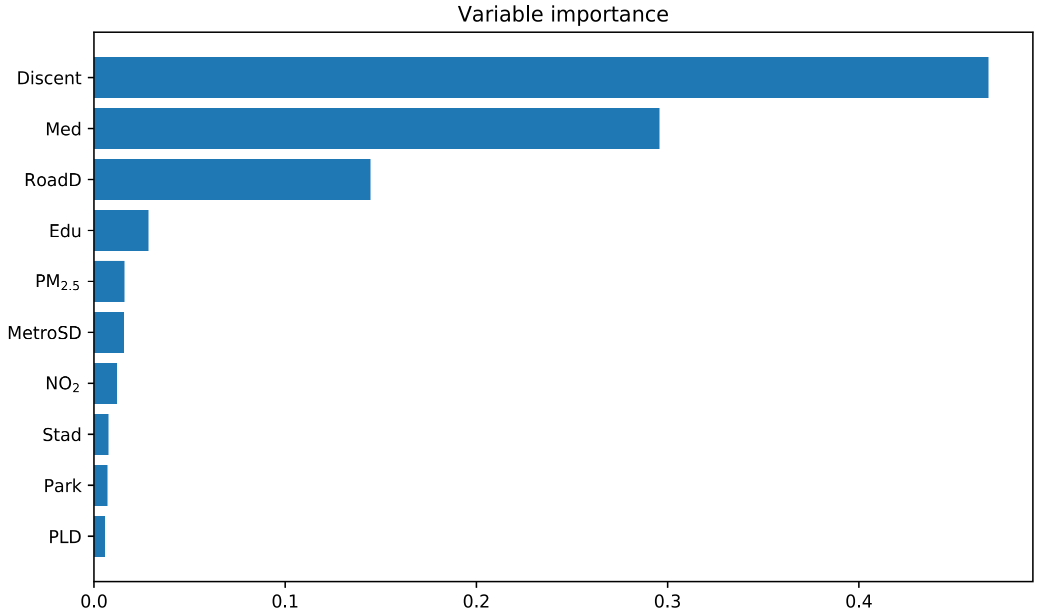

Fig. 6. The importance rankings of all the explanatory variables.

Table 4
The relative importance rankings of all the explanatory variables.

Explanatory variables	Relative importance	Ranking
Air pollutant variables		
$PM_{2.5}$	1.60%	5
NO_2	1.19%	7
Sum	2.79%	–
Built environment variables		
Discent	46.76%	1
Med	29.57%	2
RoadD	14.45%	3
Edu	2.85%	4
MetroSD	1.60%	6
Stad	0.76%	8
Park	0.69%	9
PLD	0.56%	10
Sum	97.21%	–

explanatory variables is equal to 100%, because they are measured in a relative way.

As shown in Table 4, built environment factors contribute 97.21% of the influence on housing prices, whereas the contribution of air pollutant factors is 2.79%. Among the built environment factors, straight-line distance from the city center, percentage of medical institution POIs, road network density, percentage of educational service POIs, and metro station density are the top-five most important explanatory variables, and they contribute to 46.76%, 29.57%, 14.45%, 2.85% and 1.60% of the overall influence on housing prices, respectively. The three least-important factors are the percentage of stadium POIs, percentage of park POIs and parking lot density, with contributions of 0.76%, 0.69% and 0.56% to the influence on housing prices, respectively.

For the air pollutant factors, $PM_{2.5}$ and NO_2 are the two important explanatory pollutant variables, and they contribute to 1.60% and 1.19% of the influence on housing prices, respectively. Compared with that of built environment factors, the average importance of $PM_{2.5}$ and NO_2 is much higher than the average importance of the metro station density, percentage of stadium POIs, percentage of park POIs and parking lot density. Based on the above results, we can present the following findings:

(1) In general, the built environment is the most important in housing price prediction;
(2) Compared with some factors related to the built environment, including the metro station density, percentage of stadium POIs, percentage of park POIs and parking lot density, the importance of pollutants is higher;
(3) The impact of pollutants on housing price prediction cannot be ignored and needs to be taken seriously in future urban governance;
(4) Compared with traditional methods, including GWR, the GBDT model has the advantage of not only extracting the nonlinear features from input data but also efficiently calculating the importance of the input variables used in prediction tasks.

5.5. Non-linear effects of explanatory variables

To intuitively understand the relationship between input variables and predictors, we use Partial dependence plots (PDP) to aid in-depth analysis, as shown in Fig. 7. We use partial dependence plots to demonstrate how explanatory variables are associated with housing prices, controlling for all other independent variables when one variable is studied. A smoothed curve (shown in red) is drawn to reduce the noise and better describe the relationship, following the work of (Tao et al., 2020). The PDP shows the dependency between the input feature of interest and the housing price and controls the value of all other input features (Hu et al., 2020). PDP can intuitively tell us the interaction between housing prices and input features of interest, which include linear or nonlinear. We analyze the built environment variables and air pollution variables separately as follows:

Built environment variables The nonlinear relationship between built environment variables and housing prices is quite different as Fig. 7 shows. The three-built environment variables—Metro station density, Percentage of educational service POIs, and Parking lot density—have a negligible effect on housing prices, and their values are approximately zero. Secondly, the three-built environment variables—Percentage of stadium POIs, Percentage of medical institution POIs, and Percentage of park POIs—have realized a nonlinear relationship with housing prices, but the effect is relatively small. Finally, the two-built environment variables—Straight-line distance from the city center and Road network density—have the most significant effect on housing prices, showing a solid nonlinear relationship. As the value of the variable increases, housing prices fluctuate considerably.

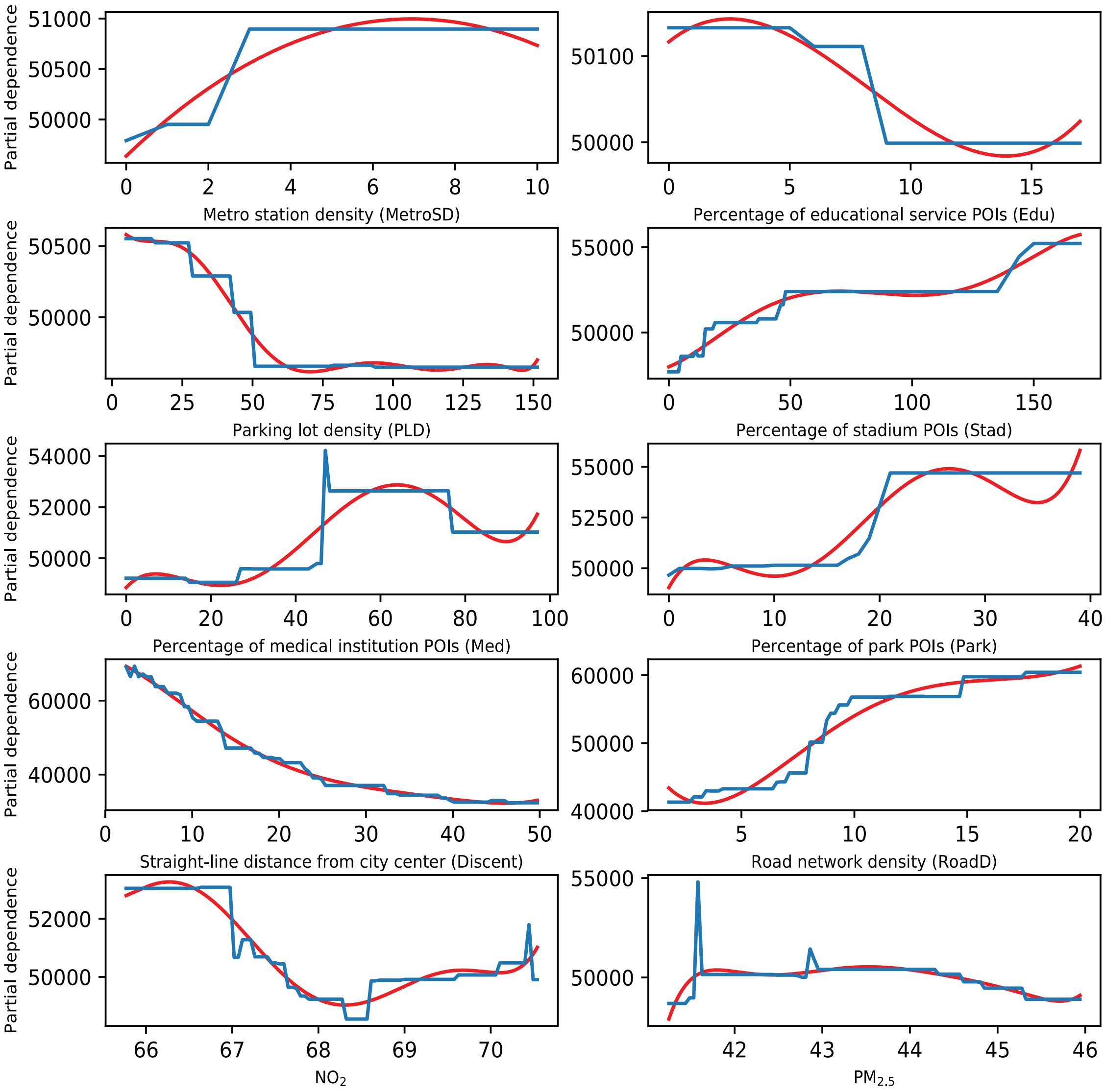

Fig. 7. The nonlinear effects of building environment variables and air pollutants variables on housing prices. The *x*-axis value represents the variable value, and the *y*-axis value represents the average housing price.

Air pollutant variables The nonlinear relationship between the concentration of air pollutants and housing prices is easily overlooked in related studies, and the value of studying the effects of pollutants on housing prices is obvious. Fig. 7 shows a nonlinear relationship between the two air pollutant variables of NO_2 and $PM_{2.5}$ and housing prices, and the overall impact on housing prices shows a negative trend. Moreover, when the concentration of NO_2 is greater than 67 μg/m^3, the house price drops sharply; similarly, when the concentration of $PM_{2.5}$ is 41.7 μg/m^3, the house price also drops sharply. Secondly, compared with $PM_{2.5}$, the overall impact of NO_2 on housing prices fluctuates more intensely, and in the case of controlled variables, the impact of NO_2 on housing prices is more significant. Therefore, studying the nonlinear relationship between air pollution and housing prices is particularly important for future urban planning as well as pollution prevention and control.

According to the above results, air pollution reduces housing prices in Shanghai, and a non-linear relationship occurs between pollution and housing prices. In our future urban planning, the Shanghai government must consider the built environment's impact on overall housing prices as well as strengthen air pollution supervision.

5.6. *Measuring residents' MWTP for clean air in Shanghai*

How much is the specific value of residents' MWTP for clean air in Shanghai? To quantify the value of residents' MWTP for clean air, we control the ranges of input variables in the trained GBDT model. When studying each variable's impact on housing prices, we control other input variables as constants, and the target variable's value is increased by 1 unit; then, we observe the changes in the dependent variable: housing prices. The experimental results are shown in Fig. 8. According to our experimental results, the absolute impact of air pollution on the Shanghai housing market is that for every increase of 1 μg/m^3 in the average concentration of $PM_{2.5}$ and NO_2, house prices will fall by 155.93 Yuan/m^2 and 278.03 Yuan/m^2, respectively (the average housing prices are 49,934 Yuan/m^2 and 49,812 Yuan/m^2, respectively).

Based on the results of these experimental analyses, we try to calculate the actual economic loss caused by air pollution in Shanghai from the perspective of the changes in the total asset value of the entire Shanghai real housing market. In 2018, the total building area of the housing stock in Shanghai was 594.6 million m^2 (Shanghai Blue Book—Shanghai Social Development Report in 2019).[1] If the

[1] https://www.pishu.com.cn.

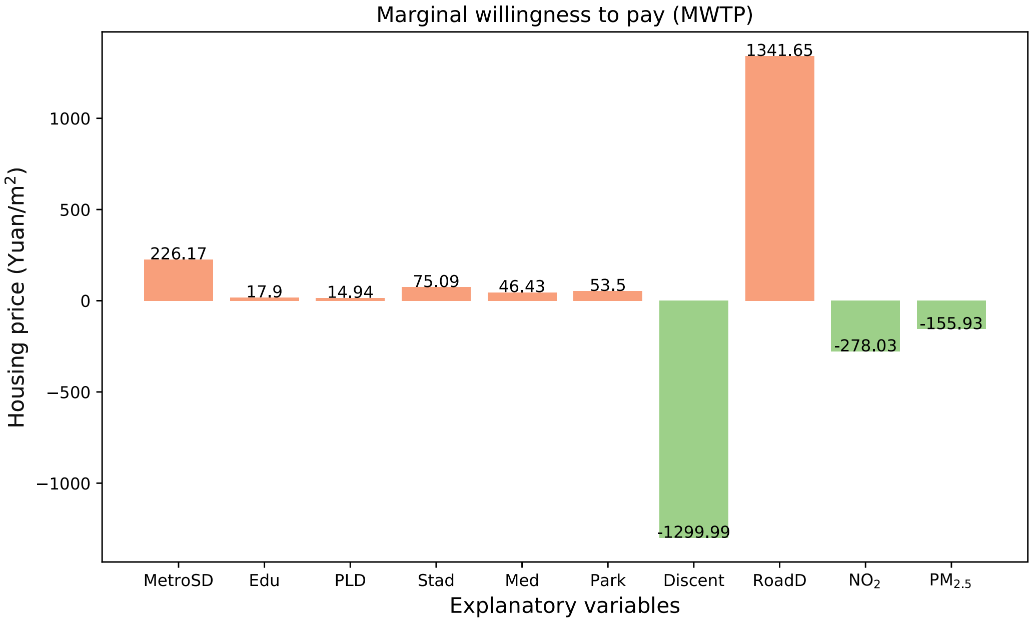

Fig. 8. The effect of a unit increase in explanatory variables on the average housing price in Shanghai. The red bar graph indicates that the housing price gain is positive, and the blue bar graph indicates the opposite.

mean concentrations of $PM_{2.5}$ and NO_2 increased by 1 μg/m^3, the estimated losses of the whole Shanghai real estate market would be 92.716 billion Yuan and 165.317 billion Yuan, respectively (in 2018, the average housing price in Shanghai was 50,090 Yuan/m^2, and the mean concentrations of $PM_{2.5}$ and NO_2 were 43.19 μg/m^3 and 42.33 μg/m^3, respectively).

According to the above analysis, our research results can prove that Shanghai residents are more sensitive to air pollution when choosing a place to live, so the marginal impact of air pollution on the real housing market cannot be ignored.

6. Conclusions

This study aims to explore the nonlinear impacts of air pollution and the built environment on housing prices in Shanghai. First, we use OLS and multicollinearity test methods to select explanatory variables related to housing prices. Second, the GBDT model is used to fit the nonlinear relationship between the explanatory variables and housing prices. The model performance is compared with that of the traditional OLS and GWR models. Then, the relative importance rankings and partial dependence plots of explanatory variables are analyzed based on the GBDT model. Finally, the marginal willingness of residents to pay for clean air and the actual economic losses caused by air pollution are measured based on the prediction of the GBDT model. The main findings are as follows:

(1) The GBDT model can accurately predict housing prices considering the nonlinear effects of air pollution and the built environment. Compared with the traditional linear regression models, the GBDT model has better performance and interpretability. The optimal R^2, MAE, and RMSE values obtained by the GBDT model are 10.320, 7.087, and 0.975, respectively.

(2) The built environment variables are the most important influencing factors of housing prices contributing 97.21% of the influences on housing prices. $PM_{2.5}$ and NO_2 are the two important pollutants that also significantly influence housing prices, accounting for 1.60% and 1.19% of the influences on housing prices, respectively.

(3) Air pollution has significant negative impacts on housing prices in Shanghai. On average, an increase of 1 μg/m^3 in the concentrations of $PM_{2.5}$ and NO_2 is associated with 0.3% and 0.6% reductions in the housing prices, respectively, corresponding to 155.93 Yuan/m^2 and 278.03 Yuan/m^2, respectively. In this case, the estimated losses of the entire real estate market in Shanghai would be 92.716 billion Yuan and 165.317 billion Yuan, respectively.

(4) Nonlinear thresholds are observed in the partial dependence plots of air pollution variables. Specifically, the housing prices drop sharply when the concentrations of NO_2 and $PM_{2.5}$ are greater than 67 μg/m^3 and 41.7 μg/m^3, respectively.

The experimental results above also provide ideas and directions for future urban planning and environmental pollution control. Some policy implications are listed as follows:

* Since air pollution can affect housing prices related to regional land values, a land value capture mechanism can be proposed to finance sustainable transportation and housing to improve urban air quality.
* Given the nonlinear relationship between housing prices and air pollution, differentiated pollution control measures should be designed for different districts with different air quality.
* Considering the increasing marginal willingness of residents to pay for clean air, real estate developers are responsible to develop more green buildings and help to improve the air quality around the buildings.

There are some limitations in the current study, which should be further addressed in the future. First, this study only uses the housing price data in Shanghai for a specific year. Given more data in the future, a more comprehensive study should be conducted to validate the applicability of the methods and findings for different cities over multiple years. Second, this study mainly analyzes the influences of air pollution and the built environment on housing prices. Other influencing factors, such as district GDP and house characteristics, should also be considered in future studies. Finally, this study neglects the spatial heterogeneity of explanatory variables' impacts. In future work, we will combine the GBDT model with the econometric model to explore more complex relationships between housing prices and explanatory variables.

CRediT authorship contribution statement

Guojian Zou: Data curation, Writing – original draft, Visualization, Investigation, Writing – review & editing. **Ziliang Lai:** Conceptualization, Methodology. **Ye Li:** Supervision. **Xinghua Liu:** Software, Validation. **Wenxiang Li:** Conceptualization, Methodology, Validation, Writing – review & editing.

Data availability statements

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

This study was sponsored by the National Natural Science Foundation of China [Grant Nos.: 71961137006, 52002244]; Shanghai Chenguang Program, China [Grant No.: 20CG55]; and the Shanghai Pujiang Program, China [Grant No.: 2020PJC083].

References

Albouy, D., 2016. What are cities worth? Land rents, local productivity, and the total value of amenities. Rev. Econ. Stat. 98 (3), 477–487.

Alin, A., 2010. Multicollinearity. Wiley Interdiscip. Rev. Comput. Stat. 2 (3), 370–374.

Alvanchi, A., Rahimi, M., Mousavi, M., Alikhani, H., 2020. Construction schedule, an influential factor on air pollution in urban infrastructure projects. J. Cleaner Prod. 255, 120222.

Bitter, C., Mulligan, G.F., Dall'erba, S., 2007. Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. J. Geogr. Syst. 9 (1), 7–27.

Brécard, D., Le Boennec, R., Salladarré, F., 2018. Accessibility, local pollution and housing prices. Evidence from nantes Métropole, France. Econ. Statist./Econ. Statist. (500), 97–115.

Casetti, E., 1972. Generating models by the expansion method: applications to geographical research. Geogr. Anal. 4 (1), 81–91.

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

Chen, D., Chen, S., 2017. Particulate air pollution and real estate valuation: Evidence from 286 Chinese prefecture-level cities over 2004–2013. Energy Policy 109, 884–897.

Chen, J., Hao, Q., Yoon, C., 2018. Measuring the welfare cost of air pollution in Shanghai: evidence from the housing market. J. Environ. Plan. Manage. 61 (10), 1744–1757.

Chen, S., Jin, H., 2019. Pricing for the clean air: Evidence from Chinese housing market. J. Cleaner Prod. 206, 297–306.

Chen, Y., Lee, C.-C., 2020. The impact of real estate investment on air quality: evidence from China. Environ. Sci. Pollut. Res. 27 (18), 22989–23001.

Chen, X., Zhang, L., Zhang, J., Li, S., 2017. China city statistical yearbook 2011.

Chuanglin, F., 2018. Important progress and prospects of China's urbanization and urban agglomeration in the past 40 years of reform and opening-up. Econ. Geogr. 38 (9), 1–9.

Chung, Y.-S., 2013. Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. Accid. Anal. Prev. 61, 107–118.

Cohen, J.P., Coughlin, C.C., 2008. Spatial hedonic models of airport noise, proximity, and housing prices. J. Reg. Sci. 48 (5), 859–878.

Ding, C., Wu, X., Yu, G., Wang, Y., 2016. A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data. Transp. Res. C 72, 225–238.

Ebenstein, A., Fan, M., Greenstone, M., He, G., Zhou, M., 2017. New evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River Policy. Proc. Natl. Acad. Sci. 114 (39), 10384–10389.

Fan, Q., Yang, S., Liu, S., 2019. Asymmetrically spatial effects of urban scale and agglomeration on haze pollution in China. Int. J. Environ. Res. Public Health 16 (24), 4936.

Fong, I.H., Li, T., Fong, S., Wong, R.K., Tallón-Ballesteros, A.J., 2020. Predicting concentration levels of air pollutants by transfer learning and recurrent neural network. Knowl.-Based Syst. 192, 105622.

Fotheringham, A.S., Charlton, M.E., Brunsdon, C., 2001. Spatial variations in school performance: a local analysis using geographically weighted regression. Geogr. Environ. Model. 5 (1), 43–66.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22 (9), 1365–1381.

Fushiki, T., 2011. Estimation of prediction error by using K-fold cross-validation. Stat. Comput. 21 (2), 137–146.

Gilderbloom, J.H., Meares, W.L., 2020. How inter-city rents are shaped by health considerations of pollution and walkability: A study of 146 mid-sized cities. J. Urban Aff. 1–17.

Guelman, L., 2012. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst. Appl. 39 (3), 3659–3667.

Hao, Y., Zheng, S., 2017. Would environmental pollution affect home prices? An empirical study based on China's key cities. Environ. Sci. Pollut. Res. 24 (31), 24545–24561.

Harrison Jr., D., Rubinfeld, D.L., 1978. Hedonic housing prices and the demand for clean air. J. Environ. Econ. Manage. 5 (1), 81–102.

Hu, S., Xie, K., Shan, X., Lin, H., Chen, X., 2020. Modeling usage frequencies and vehicle preferences in a large-scale electric vehicle sharing system. IEEE Intell. Transp. Syst. Mag. 1–10.

Huang, Z., Du, X., Wu, C., 2009. Review of research on housing price and macro-economy. China Land Sci. 23 (1), 62–69.

Huang, Y., Yi, C., 2010. Consumption and tenure choice of multiple homes in transitional urban China. Int. J. Hous. Policy 10 (2), 105–131.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, 3146–3154.

Kestens, Y., Thériault, M., Des Rosiers, F., 2006. Heterogeneity in hedonic modelling of house prices: looking at buyers' household profiles. J. Geogr. Syst. 8 (1), 61–96.

Koblyakova, A., Fleishman, L., Furman, O., 2021. Accuracy of households' dwelling valuations, housing demand and mortgage decisions: Israeli case. J. Real Estate Financ. Econ. 1–27.

Kumar, P., Imam, B., 2013. Footprints of air pollution and changing environment on the sustainability of built infrastructure. Sci. Total Environ. 444, 85–101.

Lai, Z., Liu, X., Li, W., Li, Y., Zou, G., Tu, M., 2021. Exploring the spatial heterogeneity of residents' marginal willingness to pay for clean air in shanghai. Front. Public Health 9.

Lan, F., Lv, J., Chen, J., Zhang, X., Zhao, Z., Pui, D.Y., 2020. Willingness to pay for staying away from haze: Evidence from a quasi-natural experiment in Xi'an. J. Environ. Manag. 262, 110301.

Le Boennec, R., Salladarré, F., 2017. The impact of air pollution and noise on the real estate market. The case of the 2013 European Green Capital: Nantes, France. Ecol. Econom. 138, 82–89.

Li, W., Chen, S., Dong, J., Wu, J., 2021a. Exploring the spatial variations of transfer distances between dockless bike-sharing systems and metros. J. Transp. Geogr. 92, 103032.

Li, W., Li, Y., Fan, J., Deng, H., 2017. Siting of carsharing stations based on spatial multi-criteria evaluation: A case study of Shanghai EVCARD. Sustainability 9 (1), 152.

Li, W., Pu, Z., Li, Y., Ban, X.J., 2019a. Characterization of ridesplitting based on observed data: A case study of Chengdu, China. Transp. Res. C 100, 330–353.

Li, W., Pu, Z., Li, Y., Tu, M., 2021b. How does ridesplitting reduce emissions from ridesourcing? A spatiotemporal analysis in Chengdu, China. Transp. Res. D: Transp. Environ. 95, 102885.

Li, H., Wei, Y.D., Wu, Y., Tian, G., 2019b. Analyzing housing prices in shanghai with open data: Amenity, accessibility and urban structure. Cities 91, 165–179.

Liu, R., Yu, C., Liu, C., Jiang, J., Xu, J., 2018. Impacts of haze on housing prices: an empirical analysis based on data from Chengdu (China). Int. J. Environ. Res. Public Health 15 (6), 1161.

Loh, W.-Y., 2011. Classification and regression trees. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 1 (1), 14–23.

Matus, K., Nam, K.-M., Selin, N.E., Lamsal, L.N., Reilly, J.M., Paltsev, S., 2012. Health damages from air pollution in China. Global Environ. Change 22 (1), 55–66.

Mei, Y., Gao, L., Zhang, J., Wang, J., 2020. Valuing urban air quality: a hedonic price analysis in Beijing, China. Environ. Sci. Pollut. Res. 27 (2), 1373–1385.

Schonlau, M., 2005. Boosted regression (boosting): An introductory tutorial and a Stata plugin. Stata J. 5 (3), 330–354.

Smith, V.K., Huang, J.C., 1993. Hedonic models and air pollution: twenty-five years and counting. Environ. Resour. Econ. 3 (4), 381–394.

Sun, B., Yang, S., 2020. Asymmetric and spatial non-stationary effects of particulate air pollution on urban housing prices in Chinese cities. Int. J. Environ. Res. Public Health 17 (20), 7443.

Suryowati, K., Bekti, R., Fajiriyah, R., Siswoyo, E., 2021. The effect of regional characteristics and relationship among locations in air pollution using spatial autoregressive (SAR) and spatial durbin models (SDM). J. Phys.: Conf. Ser. 1776 (1), 012051, IOP Publishing.

Tao, T., Wang, J., Cao, X., 2020. Exploring the non-linear associations between spatial attributes and walking distance to transit. J. Transp. Geogr. 82, 102560.

Tra, C.I., 2010. A discrete choice equilibrium approach to valuing large environmental changes. J. Public Econ. 94 (1–2), 183–196.

Trojanek, R., Tanas, J., Raslanas, S., Banaitis, A., 2017. The impact of aircraft noise on housing prices in Poznan. Sustainability 9 (11), 2088.

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. D: Transp. Environ. 93, 102776.

Vu, D.H., Muttaqi, K.M., Agalgaonkar, A.P., 2015. A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. Appl. Energy 140, 385–394.

Wen, H., Jin, Y., Zhang, L., 2017. Spatial heterogeneity in implicit housing prices: evidence from Hangzhou, China. Int. J. Strateg. Prop. Manage. 21 (1), 15–28.

Wu, Y., Song, Y., Yu, T., 2019. Spatial differences in China's population aging and influencing factors: The perspectives of spatial dependence and spatial heterogeneity. Sustainability 11 (21), 5959.

Yan, Y., Yao, L., Xu, T., Zhao, M., 2018. Public preference and policy evaluation of air pollution control: Taking the haze governance in Xi'an as an example. Resour. Environ. Arid Areas 32 (04), 19–25.

Yang, L., Ao, Y., Ke, J., Lu, Y., Liang, Y., 2021. To walk or not to walk? Examining non-linear effects of streetscape greenery on walking propensity of older adults. J. Transp. Geogr. 94, 103099.

Yang, L., Chau, K., Szeto, W., Cui, X., Wang, X., 2020a. Accessibility to transit, by transit, and property prices: Spatially varying relationships. Transp. Res. D: Transp. Environ. 85, 102387.

Yang, L., Chen, Y., Xu, N., Zhao, R., Chau, K., Hong, S., 2020b. Place-varying impacts of urban rail transit on property prices in Shenzhen, China: Insights for value capture. Sustainable Cities Soc. 58, 102140.

Yang, L., Liang, Y., He, B., Lu, Y., Gou, Z., 2022. Covid-19 effects on property markets: The pandemic decreases the implicit price of metro accessibility. Tunn. Undergr. Space Technol. 104528.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transp. Res. C 58, 308–324.

Zhou, Q., Zhang, X., Chen, J., Zhang, Y., 2020. Do double-edged swords cut both ways? Housing inequality and haze pollution in Chinese cities. Sci. Total Environ. 719, 137404.

Zou, Y., 2019. Air pollution and housing prices across Chinese cities. J. Urban Plann. Dev. 145 (4), 04019012.