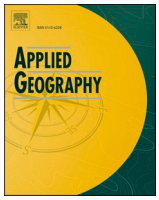
Applied Geography 158 (2023) 103032

Contents lists available at ScienceDirect

Applied Geography

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



Incorporating neighborhoods with explainable artificial intelligence for modeling fine-scale housing prices

Mingxuan Dou, Yanyan Gu *, Hong Fan

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079, China

ARTICLE INFO

Handling Editor: Dr. Y.D. Wei

Keywords:
Housing prices
XGBoost
SHAP
Big data
Nonlinear relationships

ABSTRACT

The hedonic price model (HPM) has been widely used to investigate the association between neighborhoods and housing prices. Empirical studies of HPM assume that mixed land use, accessibility, and housing structures generate a substantial premium in housing prices and follow linear relationships, with less attention paid to their nonlinear effects. To fill this gap, this study integrates transaction records over 57,842 housing units in Shanghai and explainable artificial intelligence methods to examine the nonlinear effects of public service amenities, private service amenities, and street view on housing prices. We identified the global threshold effects and their ranges, as well as the local explanations for the price forecast of each housing unit. Nonlinear analysis showed that all public service amenities and some private service amenities (e.g., entertainment) are positively related to housing prices within a certain range, while shopping and carting services are negatively related. Furthermore, the percentage of green pixels in street view images presented a nearly linear relation to housing prices. Residents in Shanghai have paid a premium of about 2000 yuan/m^2 for a higher green view. This study contributes to a better understanding of the relationships between housing units and their neighborhoods as well as guidance for the scientific and reasonable formulation of housing prices.

1. Introduction

Housing is a fundamental human need and critical to family well-being, however the majority of metropolitan cities are now facing critical housing prices challenges (Bangura & Lee, 2020; Cao et al., 2019). Housing prices are not only influenced by structural and location attributes, but also by social behaviors and the demand of home buyers (Gao et al., 2022; Jia et al., 2022). Due to the tight relationship between housing prices and economic performance, previous studies have performed extensive research on the impact of various features on housing prices, generally using the hedonic pricing model (HPM) as the research method (Liu & Strobl, 2022). To fully understand the mechanisms behind housing prices remains a challenge (Liu et al., 2021; Yuan et al., 2020). Consequently, a comprehensive understanding of how various house features and the environment affect patterns and spatial variation of the housing units' prices can improve sustainable real estate policies and city planning (Taecharungroj, 2021).

Housing prices are related to heterogeneous explanatory features, which can be grouped into three types of determinants: neighborhood, structure, and location characteristics (Chen et al., 2020). Neighborhood characteristics can be defined as the distinguishable qualities of urban amenities and varied urban regions around the residential complex. These could include the presence of amenities such as pleasant views or parks, types and quantities of local public services, and the quality of local environments such as greenery (Feng & Lu, 2013). As for the qualities of the housing unit, location characteristics include distance from housing estates to transport stations or to the central business district (CBD). For example, Yuan et al. (2020) designed a geographically weighted regression model to estimate the effects of transport accessibility. It had been discovered that the estimated coefficients of the housing variables change depending on the spatial characteristics of housing transaction datasets. Yang et al. (2021) showed the association between accessibility to BRT stations and housing prices was highly complicated and nonlinear. Traditionally, structure characteristics include floor level, building age, the greening rate, the floor-area ratio, and property management fees of housing units (Fu et al., 2019). In real estate markets, Xiao et al. (2019) concluded that the influence of floor level varied by the type of building. Feng et al. (2021) found that the total price would rise as the size of the house increased, while the marginal effect of total prices varied by the size of houses.

* Corresponding author. Wuhan University, 129 Luoyu Road, Wuhan, 430079, China.
E-mail address: yyg@whu.edu.cn (Y. Gu).

https://doi.org/10.1016/j.apgeog.2023.103032
Received 30 December 2022; Received in revised form 28 June 2023; Accepted 1 July 2023
Available online 11 July 2023

Generally, HPM emphasized the significance of the neighborhood environment in assessing property values, one essential issue remains: how important are neighborhoods in helping to collectively forecast housing prices and in what ways do these variables affect house prices? Therefore, it is crucial to understand the relationships between housing units and their neighborhoods. Previous studies have assumed a linear or log-linear statistical assumptions when estimating the implicit price of neighborhoods on housing prices (Feng et al., 2021; Jin et al., 2022). Spatial variations in housing prices occur across geographies and can also be influenced by housing neighborhoods. Moreover, previous HPM studies were applied at the regional level or in single-city cases based on aggregate statistics for housing prices at a large geographical scale (Chi et al., 2020). To the best of our knowledge, only a few empirical studies have investigated the nonlinear relationships between housing prices and neighborhoods at individual level (Pérez-Molina, 2022). Inspired by previous research, this study incorporated explainable artificial intelligence (XAI) models with data from online transaction records, POIs, and street view imagery to model housing prices. Neighborhoods characteristics can be described as the availability and accessibility of essential urban infrastructure amenities, such as street greenery and educational facilities (Jim & Chen, 2006). Fine-grained big data helps quantitatively describe neighborhood facilities in detail (Chi et al., 2020; Gu et al., 2023). The nonlinear relationships between neighborhoods and housing prices were investigated and interpreted by using the eXtreme Gradient Boosting (XGBoost) and SHapley Additive exPlanations (SHAP) models. The contributions concerned three main aspects:

(1) Provided a methodological expansion by uncovering potential nonlinear relationships between neighborhoods and housing prices at individual level.
(2) Scrutinized feature importance for housing prices from both global samples and local instances, respectively.
(3) Enriched the experience in real estate market research by combining multi-source data and XAI methods.

The rest of this paper is organized as follows: Section 2 provides a review of the literature on the transition from the HPM to the XAI model. The datasets and methods are introduced in the next section. Nonlinear estimation results and feature interpretation from both global and local perspectives are provided in Section 4. We discussed the model in Section 5. The final section concludes this study.

2. From HPM to XAI model

The hedonic pricing model (HPM), widely used to explore housing prices, views neighborhood, locational, and structural attributes as the most important determinants of housing units' values (Qiu et al., 2022). Housing prices are characterized as hedonic pricing, which are determined not only by the housing's attributes, but also by location-specific variables (Yuan et al., 2020). Therefore, the HPM treats a housing unit as a composite good with a variety of attributes and provide a flexible and efficient method for measuring differences between housing units from the perspective of households (Fu et al., 2019). By integrating human social sensing and urban appearance, the main assumption of the HPM is that potential housing buyers pay not only for the unit itself but also for the surrounding urban environment (Jia et al., 2022; Su et al., 2021). In practice, the HPM is often favored to quantify the effects of explanatory features on housing prices using an ordinary least squares (OLS) estimator or combined with linear regression analysis, which assumes that the housing features are independent with no spatial autocorrelation (Zhang et al., 2021). However, in local regression situations, this assumption could not be satisfied to examine the geographic relationship between housing prices and their potential variables, particularly when spatial variables are involved and there is regional variation in geographic data sets (Cao et al., 2019).

The HPM is not sensitive enough to detect spatial dependency and heterogeneity when analyzing influencing factors in housing prices (Liu & Strobl, 2022). With the development of advanced spatial analyses, the Geographically Weighted Regression (GWR) has been the most preferred localized multiple-regression model, which allows explanatory variables to explain local effects and variations spatially. Recently, the GWR model has been widely applied in housing price studies to explore spatial heterogeneity in the effects of structural amenities, locational context, public service facilities (Liu & Strobl, 2022). For example, Cellmer et al. (2020) assessed the relationship between socio-demographic-environmental factors and average prices using the GWR model and indicated that the determinants of housing prices were spatially differentiated. To determine the potential influencing factors of housing prices, Sisman and Aydinoglu (2022) selected nine significant variables from twenty-eight attributes by using GWR and Multiscale Geographically Weighted Regression (MGWR). They found that results from local models achieved better performance in comparison to global models such as OLS. Hu et al. (2022) proposed a methodology for exploring data relationships between housing prices and their determinants concerning spatial heterogeneity by combining the hierarchical linear model (HLM) and the GWR, namely HLM-GWR. Although the GWR model is widely acknowledged and regarded as an orthodox method in housing price studies, non-linearity in data patterns and multicollinearity between housing variables can seriously impair the performance or generate biased determinants (Hu et al., 2022; Iban, 2022).

In recent years, emerging big data and advanced machine learning (ML) algorithms have provided unprecedented opportunities for researchers to understand the potential determinants of housing prices (Bourassa et al., 2021). Scholars have studied different ML methods, such as XGBoost, to handle linear and nonlinear effects of explanatory variables on housing prices from both categorical and numerical datasets (Soltani et al., 2022). Compared to local regression approaches such as GWR, the local XGBoost is more flexible and makes fewer assumptions when detecting multilevel factor interactions or nonlinear relations (Taecharungroj, 2021). For example, using the Xiamen data, Yang et al. (2021) employed the gradient boosting decision tree (GBDT) to explore the non-linear relationship between BRT and house prices. Their results indicated that ML algorithms have more substantial predictive power than HPM. Despite the excellent fitting performance of ML, interpretation of the results can be extremely challenging, especially in cases of high model complexity. Meanwhile, feature importance cannot decipher the quantitative effect of each feature on the local level for an individual instance, either. To solve this problem, some researchers have started to take advantage of the XAI (explainable artificial intelligence) in housing prices studies. The XAI discusses the explainability of ML algorithms, such as SHAP method (Lundberg et al., 2020) and the Local Interpretable Model-agnostic Explanations (LIME) model (Ribeiro et al., 2016), which are emerging as local approaches to explain the outcome of ML methods across a broad set of domains. The LIME compares the differences between generated data samples and the original data samples, which has the main limitation that the LIME is only reliable if the simple interpretable model closely approximates the black box predictions. The main idea of SHAP is to discover what happens in the ML models (such as XGBoost, random forest, and neural networks) when a feature is missing that meets the requirements of local accuracy, missingness, and consistency. Additionally, XGBoost is well integrated with SHAP, and SHAP values can be efficiently estimated to reveal the spatial variations of the contribution of each determinant to the price of housing units.

3. Methodology

The framework of this study is presented in Fig. 1. The process starts with obtaining the dataset containing transaction records, POIs, and street view imagery for the study area. Then, the various factors that affect the value of the housing unit were determined, including

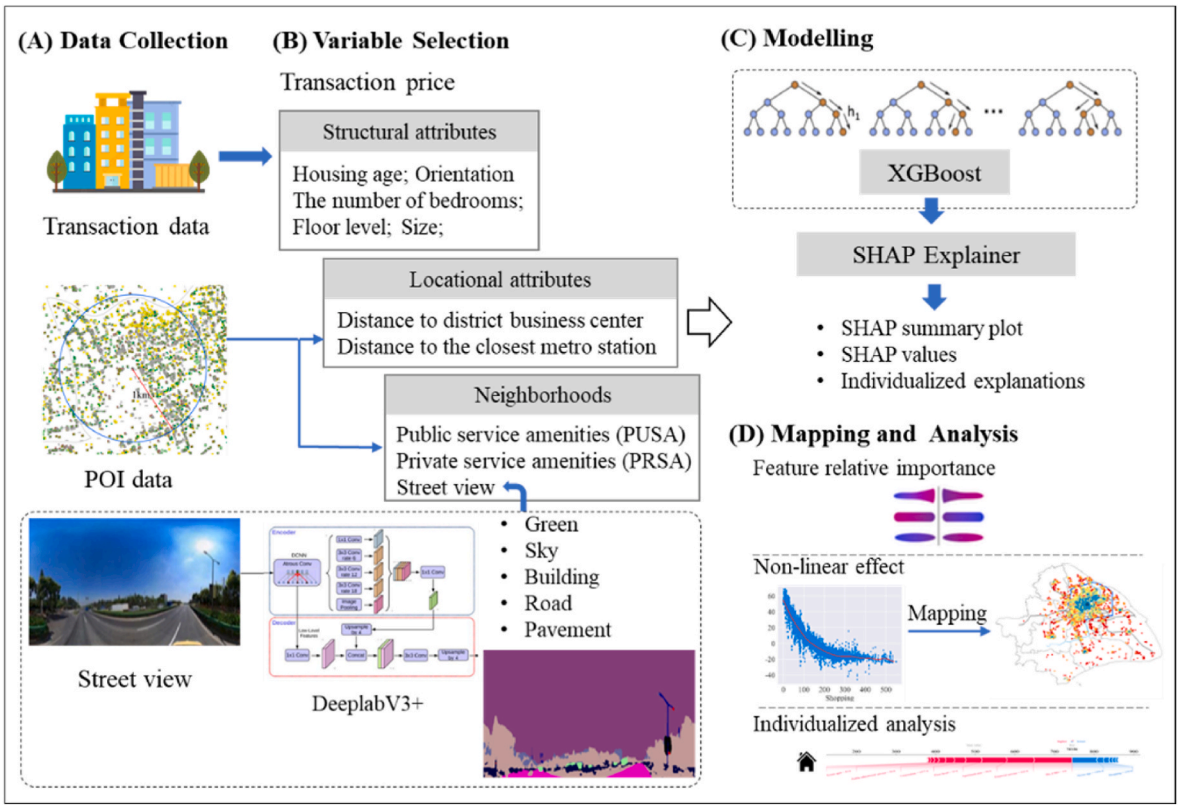


Fig. 1. Overall framework of this study.

neighborhood, locational, and structural attributes. The local regression models (XGBoost) are implemented using the housing prices dataset, and the SHAP method has been introduced to explain the nonlinear effects of public service amenities, private service amenities, and street view on housing prices. Finally, the contributions of neighborhood attributes derived from POI data and street view data were computed and mapped.

3.1. Study area and data

This study has been conducted in Shanghai, which is widely considered to be one of the most economically advanced regions in China. This study involves the full municipality of Shanghai along with its sixteen urban districts (Fig. 2). The central region (667 km^2) of Shanghai has a typical high-density urbanized characteristic (Dou et al., 2021). Like other international megacities, Shanghai has experienced a surge in housing prices. According to the average nominal transactional prices of Shanghai from 2012–2018, the housing prices have skyrocketed by more than 119%, soaring from 25,111 yuan/m^2 to nearly 55,095 yuan/m^2, posing an urgent challenge for planners and policymakers.

In this study, we primarily used the following datasets: 1) transaction data incorporating housing prices and basic structural characteristics; 2) points of interest (POI) data incorporating public service amenities and private service amenities; 3) population density data with 1km × 1 km spatial resolution, which was collected from LandScan global population database (https://landscan.ornl.gov); and 4) street view images representing streetscapes from pedestrians' perspectives. Compared to traditional data sources, these datasets describe neighborhoods of housing units with particular advantages (e.g., fine-grainedness and human-centered scale).

Transaction data of each housing unit: The historical transactional data of houses were collected from Lianjia (https://lianjia.com). Based on the web-crawling program, we obtained the historical

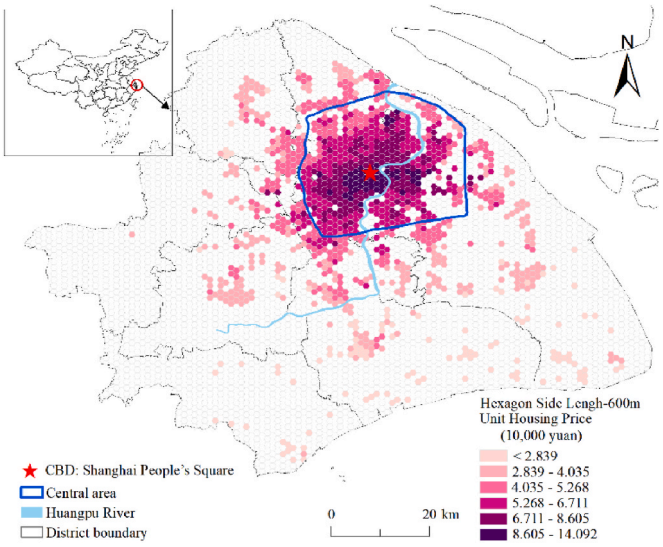


Fig. 2. Spatial distribution of housing prices in Shanghai.

transactional data of each housing unit in November 2018, with attributes including address, number of rooms, apartment size, floor level, orientation, age, total price, and price per square meter. After deduplication, a total of 57,842 transaction records for housing units were reserved in Shanghai.

POI data: POI data can provide spatial and attribute information about a geographic entity, such as name, address, coordinates, and type. The empirical studies show that POI as new geographic data is used to extract housing location and community characteristics (Hu et al., 2019). This study collected POIs in 2018 from an online mapping services company (https://www.amap.com) covering the study area. More

than 1.38 million POIs were obtained within the study area. By merging and deleting secondary industry classification, POI data were divided into subway stations, bus stations, educational facilities, financial facilities, medical facilities, scenic facilities, entertainment facilities, shopping facilities, catering services, and corporation for further analyses.

Street view images: The street view data were crawled during 2018 through the Baidu Street View API interface. Street view images can truly and effectively reflect the visual environment of the surrounding space of urban streets. After data filtering to remove the weight, a total of 2,239,710 pictures of street scenes were collected, covering the main road networks of Shanghai. Georeferenced sample locations were set up every 50 m along the street networks. For each location in the sample, a complete panorama was synthesized with 4096 * 2048 pixels. We used Deeplabv3+ (Chen et al., 2018) to conduct semantic image segmentation. To train the model, we employed an expert labeling company to label 5000 street view images, involving 19 categories. The segmentation accuracy of the DeeplabV3+ reached 98.1% and 95.1% in the training and test sets, respectively. In this study, vegetation, the sky, buildings, roads, and sidewalks were selected to construct the street view index by comprehensively considering the street view elements that may have an impact on the value of housing prices (Fig. 3). The streetscape feature calculation formula for a panorama-generating position is:

$$SVI_{green} = \frac{pixel(green_i)}{pixels_i} \# \quad (1)$$

SVI_{green} represents the percentage of the vegetation in the image. $pixel(green_i)$ and $pixels_i$ represent green pixels and total pixels in image i, respectively. Similarly, SVI_{sky}, $SVI_{building}$, SVI_{road}, $SVI_{pavement}$ were calculated using the same procedure as SVI_{green}.

3.2. Variable description

The selection of explanatory variables is inspired by the literature and the availability of relevant data (Li et al., 2021). Three types of independent variables are selected: structural attributes, location, and neighborhoods. In this study, neighborhood characteristics are measured with a distance threshold of 1 km, corresponding to a walking distance of 15~20 min, which is a reasonable range for most people (Xiao et al., 2019). Most importantly, Chinese cities are now advocating a 15-min walk as the threshold for delineating neighborhoods (Xu et al., 2017). Table 1 summarize the detailed descriptions and statistics of independent variables.

Table 1
Variable descriptions.

Variables	Description	Mean	St.dev.
Dependent variable			
Housing Prices	Transacted price of housing units per m^2(10,000 yuan).	5.81	2.03
Independent variables			
Structural attributes			
Number of bedrooms	Number of bedrooms of a house.	2.51	1.30
Apartment size	Total area of a house (m^2).	89.12	47.14
Floor level	The floor number of a house.	10.30	7.42
Orientation	Dummy variable (1 if southern orientation; 0, otherwise).	-	-
House age	The age of a house unit.	20	9.88
Location			
Distance to CBD	Distance from the unit to the nearest CBD (km).	13.34	9.12
Distance to the station	Distance from the unit to the nearest metro station (km).	1.42	2.29
Neighborhoods			
Population density	Number of populations within 1 km^2.	12464	8430
Public service amenities (PUSA)			
Quality of schools	Dummy variable (1 if high-quality primary of junior high school exists within 1 km; 0, otherwise).	-	-
Public education	Number of public educational service within 1 km.	21.34	19.71
Financial service	Number of financial facilities within 1 km.	73.79	93.17
Medical service	Number of medical facilities within 1 km.	68.68	52.53
Scenic spot	Number of scenic spots within 1 km.	8.91	12.82
Number of bus stops	Number of bus stations within 1 km.	23.91	8.51
Private service amenities (PRSA)			
Entertainment	Number of entertainment amenities within 1 km.	60.49	145.76
Shopping	Number of shopping amenities within 1 km.	136.29	81.19
Catering services	Number of catering amenities within 1 km.	176.11	123.75
Street view (SV)			
Green view	The percentage of green within 100 m.	0.25	0.14
Sky view	The percentage of sky within 100 m.	0.41	0.12
Building view	The percentage of building within 100 m.	0.14	0.09
Road view	The percentage of road within 100 m.	0.08	0.02
Pavement view	The percentage of pavement within 100 m.	0.02	0.01

Note: Medical services include hospitals and clinics. Entertainment facilities refer to bars, gyms, music clubs, galleries, etc. Shopping facilities consist of shopping malls, supermarkets, and other markets. Catering services include Chinese restaurants and foreign restaurants.

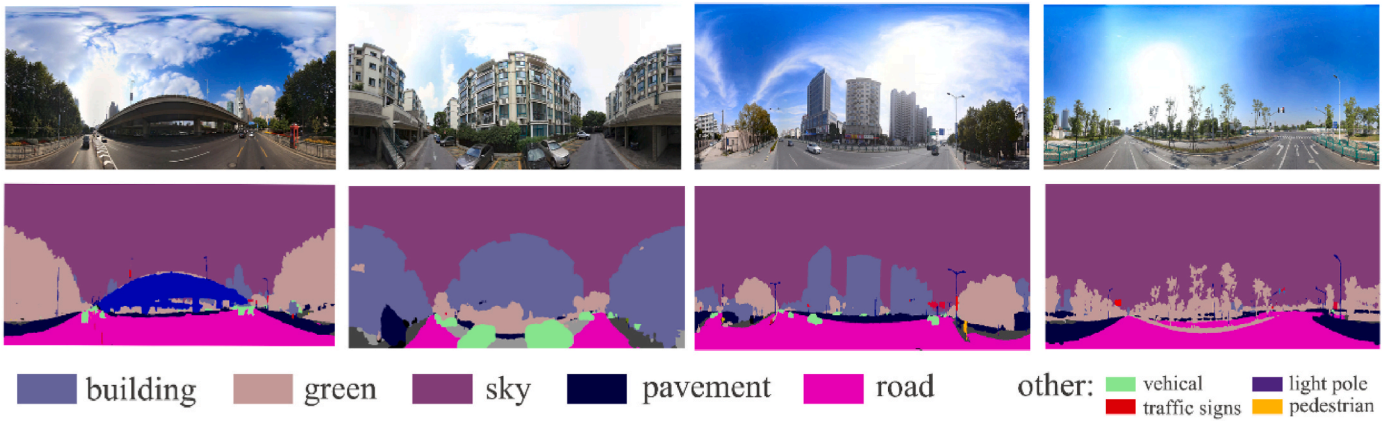


Fig. 3. Examples of DeeplabV3+ results.

3.3. XGBoost model

The XGBoost model is utilized to explore the nonlinear association between neighborhoods and housing prices (Chen & Guestrin, 2016). The XGBoost model is an efficient and powerful machine learning technique that constructs an accurate prediction through a boosting process. XGBoost has been widely used for classification and regression tasks due to its high accuracy, good generalization ability, and strong anti-noise ability in the field of urban planning (Zeng et al., 2022).

The objective function of XGBoost typically has two components, which are known as training loss and regularization, represented by Eq. (2):

$$\mathrm{Obj}(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \# \tag{2}$$

Where y_i is the dependent variable, and $\hat{y}_i$ is the estimated value from the input variables x_i. f_k are the weak learners at the k-th iteration. $\sum_i l(\hat{y}_i, y_i)$ is the training loss from training data and Ω is the regularizing term. The objective goal is to identify a function f_k that best estimates the dependent variables y_i when considering the complexity of the model to avoid overfitting.

As there are multiple parameters in XGBoost, over-fitting occurs when a model begins to learn noises and eventually perceives them as relevant facts. To avoid too much complexity and achieve better model performance, hyperparameters need to be fine-tuned. The most noticeable hyperparameters are the number of iterations k and the maximum depth. Overfitting is more likely with a greater k and a deeper maximum depth, resulting in a loss of generality. Other hyperparameters, such as subsample (e.g., the proportion of observations chosen at random from training samples), learning rate (the step size weights to be updated), colsample_bytree (the subsampling number of columns), and lambda and alpha (i.e., regularization terms on weight scores of leaves). These are the optimized XGBoost hyperparameters value that has the smallest RMSE value, confirming the high capacity of explanatory variables to predict housing prices. Widely used evaluation metrics, such as MAE, RMSE, MAPE, and R-squared, are applied to measure model performance. In addition, 5-fold cross-validation is employed to prevent the overfitting and bias of the model. After parameter optimization, the maximum depth is set to 10, subsample is 0.7, the subsample ratio of columns in the bytree is 0.9, lamda is 0.17, alpha is 9.5, and the learning rate is 0.016. Table 2 presents the model's performance. The cross-validation score is 0.93, which indicates that overfitting is less likely to have occurred in the result. Then, the SHAP model is employed to produce a robust interpretation of the XGBoost prediction results for further analysis.

3.4. SHAP model

SHAP, proposed by Lundberg and Lee (2017) based on game theory, is used for post-facto interpretation of machine learning models. In contrary to the existing importance features in explainable artificial intelligence models, SHAP can determine whether each input feature has a positive or negative contribution. Each observation gets its SHAP value, which can help interpret the model globally as well as locally. Moreover, SHAP values can explain the modeling of local interaction effects that might go unnoticed otherwise. For an individual sample x, this technique can explain the specific prediction $f(x)$ through computing the relative contribution value $\varnothing_i$ of each feature i. Furthermore, the global importance is estimated by averaging the absolute Shapley values per feature throughout the data.

Assume that g is the explanation model, m represents the total number of features in x, and $z \in \{0,1\}^m$ is a simplified feature used to predict an output g(z); $z_i = 1$ means the feature i is utilized in the calculation, otherwise 0; $\varnothing_0$ represents the base value for the entire model, usually the average value of overall predictions; $\varnothing_i \in \mathscr{R}$ is the SHAP value for feature i. The SHAP is defined as follows:

$$\varnothing_i = \sum_{S \in \frac{N}{\{i\}}} \frac{|S|!(m-|S|-1)!}{m!}[f_x(S \cup \{i\}) - f_x(S)] \# \tag{3}$$

Where, x is independent instance, S is subset features used in the model, $f_x(S) = E[f(x)|x_S]$ is the expected prediction value based on subset x_S. $\varnothing_i$ is the Shapley value for each feature i, calculated through the weighted sum of the marginal contribution $[f_x(S \cup \{i\}) - f_x(S)]$ in all sub models containing the feature i. The situation $\varnothing_i > 0$ indicates that feature i has an improvement effect on the final prediction result; otherwise, it indicates that feature i reduces the final prediction result.

In addition, SHAP values utilize the Shapley interaction index from game theory, which allows rewards (i.e., importance) to be allocated not just to individual players, but also between all pairs of players. The difference between the SHAP value and the Shapley interaction index is specified as another analytical indicator, the SHAP main effect value for a prediction. This study derived the contribution of various determinants on housing prices from SHAP value and nonlinear effects of each determinant from the SHAP main values.

4. Results

4.1. Relative importance of variables

The XGBoost model results are shown in Table 3, along with the relative weights and rankings of predictor variables. Fig. 4 provides a graphical representation of the relative significance from both global and local perspectives. In Fig. 4, each row represents a feature, ranked by its global importance on the vertical axis. Each scatter point represents a SHAP value for one feature of each housing unit. The value of SHAP determines where the point is located on the horizontal axis, and the color of the point (ranging from red to blue) shows the relative importance of the feature, from high to low.

XGBoost enables a comparison between the contributions of the three variable categories collectively. The neighborhood characteristics have the highest relative value (45.48%), account for nearly twice over than the structural categories (20.13%) and are slightly higher than the location (34.39%). The results lend credence to the idea that neighborhood amenities have a substantial impact on deciding housing prices (Li et al., 2019).

Further analysis of the neighborhood factors shows that the contribution of public services (23.18%) is greater than that of private services (7.06%). The 11.24% of housing prices could be explained by population density. Among public services variables, the significant price premium is currently being produced by the high land value of financial services (11.73%) as well as the scarcity of scenic spots (3.75%) and public educational services (1.94%). Meanwhile, the street view has a subtle contribution to making predictions of housing prices, especially

Table 2
XGBoost model performance.

R-squared		MAE		RMSE		MAPE	
Training	Test	Training	Test	Training	Test	Training	Test
0.943	0.928	0.405	0.627	0.476	0.590	0.038	0.041

Table 3
Relative importance of independent variables.

Determinants	Variables	Rank	Importance	Total
Structural attributes				20.13%
	Number of bedrooms	18	0.76%	
	Apartment size	6	4.83%	
	Floor level	10	2.80%	
	Orientation	22	0.48%	
	House age	3	11.26%	
Location				34.39%
	Dis_CBD	1	28.49%	
	Dis_metro	5	5.9%	
Neighborhoods				45.48%
POP	Population density	4	11.24%	11.24%
PUSA	Quality of school	9	2.82%	23.18%
	Public educational service	11	2.12%	
	Financial service	2	11.73%	
	Medical service	13	1.69%	
	Scenic spot	8	3.75%	
	Number of bus stations	16	1.07%	
PRSA	Entertainment	12	1.93%	7.06%
	Shopping	7	3.92%	
	Catering services	15	1.21%	
SV	Green view	14	1.26%	4.00%
	Sky view	21	0.58%	
	Building view	17	0.82%	
	Road view	19	0.72%	
	Pavement view	20	0.62%	

the green view.

In the field of real estate studies, the GWR is a more robust model compared to the HPM, which is usually examined by using ordinary least squares (OLS) (Du et al., 2018). Due to the neglect of spatial autocorrelation, the spatial lag model (SLM) and spatial error model (SEM) were developed to solve the issue by incorporating the geographic effects of the dependent variable and error term, respectively (Sisman & Aydinoglu, 2022). Generally, R^2 and RMSE are used to evaluate model performance: R^2 represents the percentage of variation in the dependent variable predicted by the explanatory variables, while RMSE evaluates discrepancies between predicted and observed values (Jin et al., 2022).

Therefore, a larger R^2 with a smaller RMSE means better model performance. We used the Gaussian kernel to calculate the spatial weight matrix and model the spatial interactions between housing prices and determined variables.

The comparison results of R^2 and RMSE for different regression models are illustrated in Table 4. The R^2 value for XGBoost is the highest, compared with the best performance statistics among OLS, SLM, SEM, and GWR. The R^2 values for XGBoost and GWR are 0.912 and 0.551, which means that the XGBoost and GWR models can capture 91.2% and 55.1% of the influence of the variables, respectively. In addition to R^2 values, XGBoost has the lowest RMSE, MAE and MAPE among all the regression methods, indicating a better performance in modeling fine-scale housing prices. In conclusion, the XAI models are a promising way to explore the interactions between neighborhoods and housing prices.

4.2. Associations between the neighborhoods and housing prices

In addition to its relative importance, the SHAP model produces a main effect value for predictor variables. The dependence plot of the SHAP main effect value can clearly visualize the marginal effects of the features on the outcome (Figs. 5 and 6). The X-axes represent feature value, and the Y-axes represent the corresponding SHAP main effect value (i.e., how valuable the variable contributes to the housing prices after other interaction factors are removed). Each blue point represents a sample of housing unit, and the spline-colored red indicates the trends. Each variable exhibits a unique nonlinear association with housing prices.

Fig. 5a represents the nonlinear relationship between population density and housing prices. When the population density is lower than 16,000 people per km^2, the housing price increases rapidly. As it rises

Table 4
Comparison of R^2 and RMSE among regression models.

Criteria	OLS	SLM	SEM	GWR	XGBoost
R^2	0.481	0.412	0.479	0.551	0. 912
RMSE	98,212	79,371	78,023	62,472	45,751
MAE	1.170	0.979	0.972	0.768	0.579
MAPE	1.383	1.157	1.290	0.908	0.039

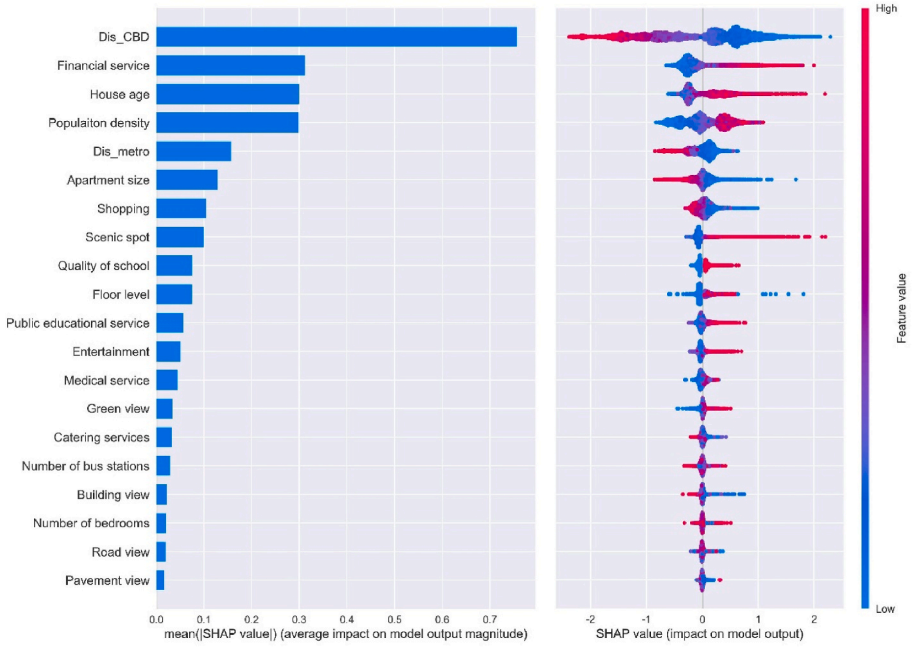


Fig. 4. Global (left) and local (right) feature relative importance by SHAP.

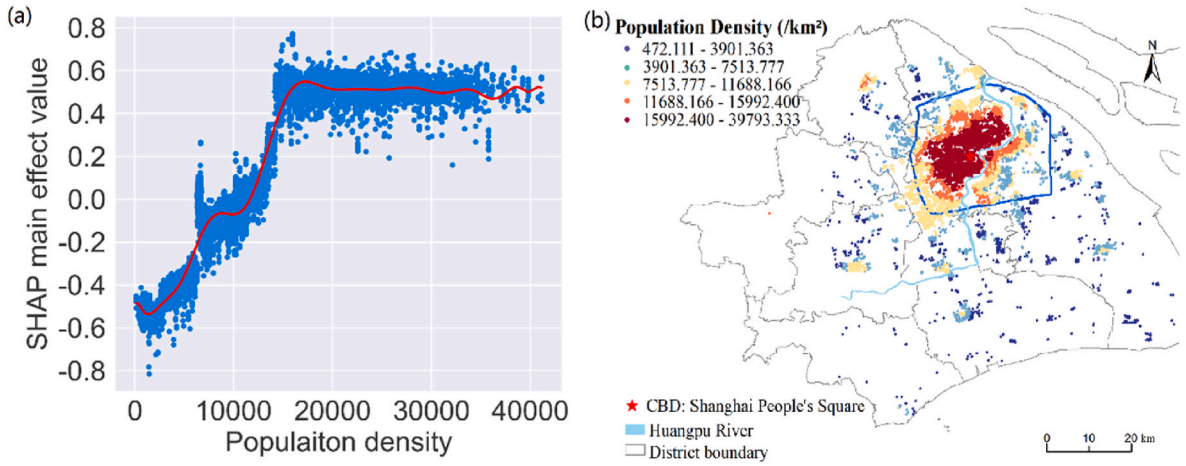


Fig. 5. (a) Nonlinear associations between population density and housing prices; and (b) Spatial distribution of population density per km^2 in Shanghai.

over 16,000 people per km^2, the surrounding neighborhoods no longer benefit from the premium impact of population on housing unit price (Fig. 5b).

Fig. 6 shows that all public service amenities variables have nonlinear and threshold associations with housing prices. For public educational services, within the range of 0~75, the price of housing units increases linearly by about 2600 yuan/m^2. Over this range, the effect is trivial. A similar trend is observed in financial services. When the number of financial services increases from 0 to 300, the price of housing units increases substantially (to about 15,000 yuan/m^2). These indicate that the financial industry often has a higher land economic value, along with a concentrated form in the urban core. As can be seen from the medical service plot, the effect on housing prices may be divided into two categories. Within the range of 0~150, housing prices increase linearly by about 1000 yuan/m^2. When it moves to 150–300, the price of housing units is basically unchanged which indicates the marginal effect diminishes.

Besides the threshold effects, the directions of PUSA influence on housing prices revealed here are mainly consistent with the previous (Yuan et al., 2020). Various findings have shown that public educational and medical services have a positive relationship with housing prices (Qiu et al., 2022). Scenic spots are positively related to housing prices, while bus stations have a negligible effect on housing prices.

Fig. 7 shows the effects of PRSA variables on predicted housing prices. Providing entertainment services would increase the price of housing units by about 3000 yuan/m^2. The results on shopping and catering services are quite unexpected, while they show an obvious negative effect on housing prices. This phenomenon is probably due to increased mobility for urban residents in accessing shopping and catering amenities with the development of e-commerce and the food delivery industry. The spatial distribution of the SHAP main effect value in Fig. 8b is consistent with the evidence from Taipei, which demonstrates that the density of convenience stores is negatively correlated with high property prices, due to residents' higher mobility and possibly noise (Chiang et al., 2015).

The relationships between street view and housing prices are depicted in Fig. 8. Consistent with previous studies (Jia & Zhang, 2021), the green view has a nearly linear positive relationship, increasing

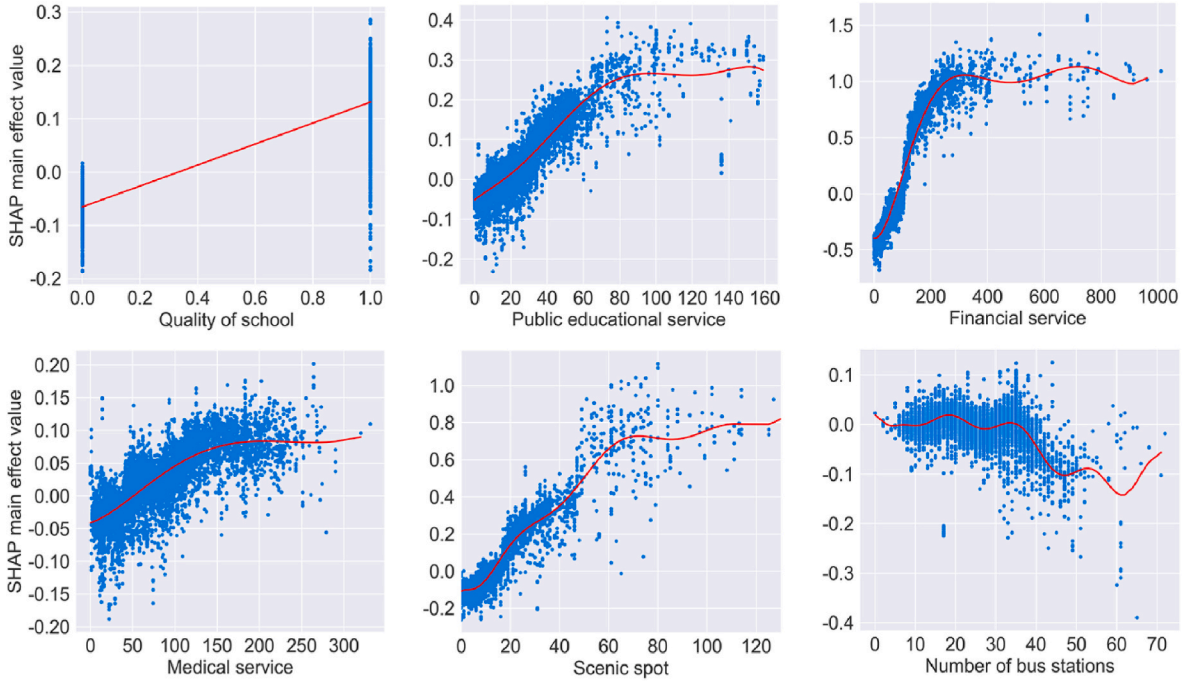


Fig. 6. Nonlinear associations between public service amenities and housing prices.

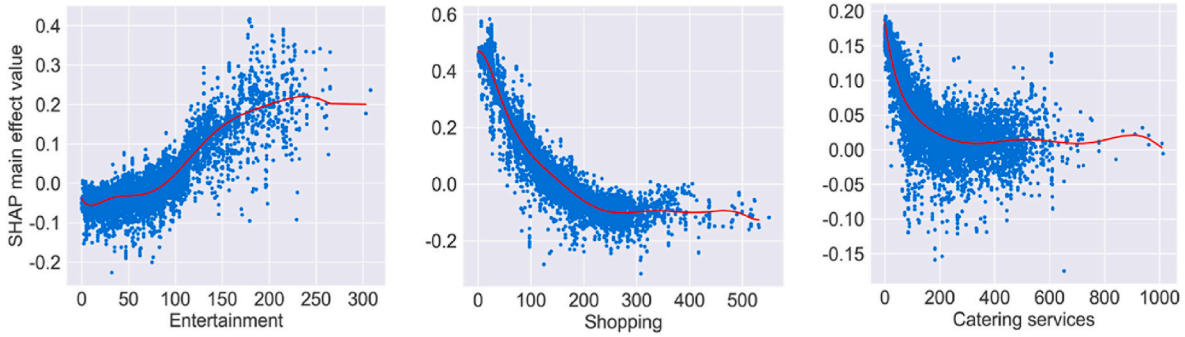


Fig. 7. Nonlinear associations between private service amenities and housing prices.

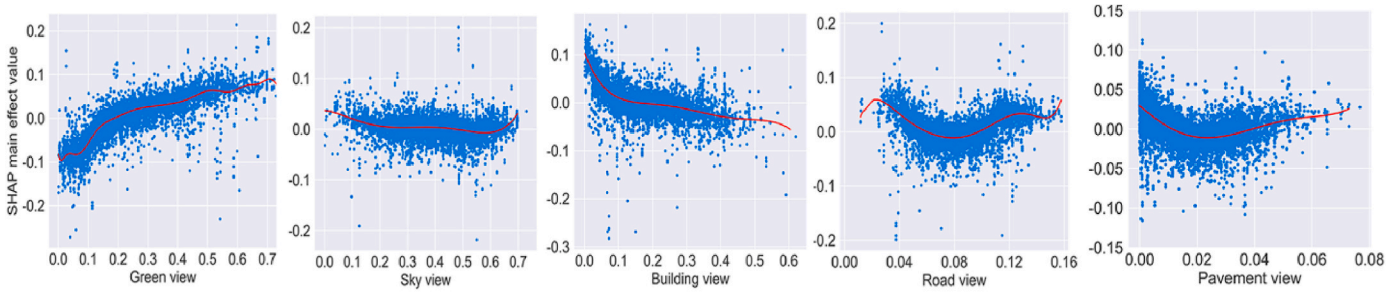


Fig. 8. Nonlinear associations between street view (SV) and housing prices.

housing prices by about 2000/m^2. The spatial distribution of the SHAP main effect value indicates a green view premium is attached to the housing units within the full municipality of Shanghai. On the contrary, the sky view and building view show a negative effect on housing prices, falling by about 500 yuan/m^2 and 1800 yuan/m^2, respectively. For road view and pavement view, relatively imperceptible positive effects can be observed.

Fig. 9 shows the spatial distribution of SHAP main effect values of the top 10 variables. For locational attributes (distance to the CBD and distance to the station), the obvious distance decay effect on housing prices can be seen in Fig. 9a–d. For neighborhood variables, the patterns of spatial variation are distinct: the effects of population density, financial services, and shopping facilities on housing prices vary significantly from the central city to the suburbs (Fig. 9b, c, and 9e).

4.3. SHAP local explanations

To explain how SHAP explains individual characteristics, we randomly sampled four instances from the transaction dataset, as illustrated in Fig. 10. For each instance, the plot shows the direction and strength of the effect of the features on the final predicted result based on the base value (58,100 yuan/m^2), which represents the average prediction value. Fig. 10a, located in the central area of Shanghai and enjoying plenty of amenities, has the highest housing prices among the four cases. Fig. 10b, located in the financial center, benefits from financial service and is close to the CBD. As for Fig. 10c and d, located in the urban periphery, the distance to the CBD and population density clearly have a detrimental effect on housing prices. In contrast, the facilities for shopping show a positive contribution to housing prices.

5. Discussions

5.1. Key findings

In this study, nonlinear effects of neighborhood, locational, and structural factors on housing prices in Shanghai in 2018 were studied using explainable artificial intelligence methods (XGBoost and SHAP) and housing transaction data.

The state-of-the-art machine learning model, XGBoost has been employed to analyze the potential determinants of housing prices, including the number of bedrooms, apartment size, floor level, orientation, house age, distance to CBD and stations, population density, public service amenities, private service amenities, and street view. A comparison of XGBoost with traditional methods is needed to examine whether our proposed methodology improves the accuracy of the regression results. The results demonstrate that among all factors, distance to CBD, population density, financial service, house age, distance to stations, and apartment size significantly affect the housing prices based on the housing transaction data sets in Shanghai in 2018. The percentage of green pixels in a street view image presented a nearly linear relation to housing prices, which indicated that housing owners in Shanghai have paid a premium of about 2000 yuan/m^2 for a higher green view.

5.2. Application of XAI in housing market

Compared to traditional hedonic pricing models, machine learning models developed recently can scrutinize complex non-linear relationships between factors and housing prices with high prediction accuracy. However, the complexity of machine learning makes it often regarded as ‘black boxes’, resulting in insufficient interpretability of the predictions. In response, XAI techniques have emerged to help researchers understand the relative importance of housing characteristics. XAI can be divided into two main categories: global and local explanations. Global explanation methods take a holistic approach and try to make a single ranking of all features for the model. Partial dependence plots (PDPs) and feature importance are two typical representative methods. For example, Yang et al. (2021) applied the PDPs to explain the results of the gradient boosting decision tree (GBDT) model. Ma et al. (2020) employed the Gini index to rank and analyze the feature importance of random forests. In terms of local explanations, two critical methods, i.e., LIME and Shapley values, determine the importance of each feature for each data point. The latter approach measures the average marginal contribution of each feature across all possible combinations of features,

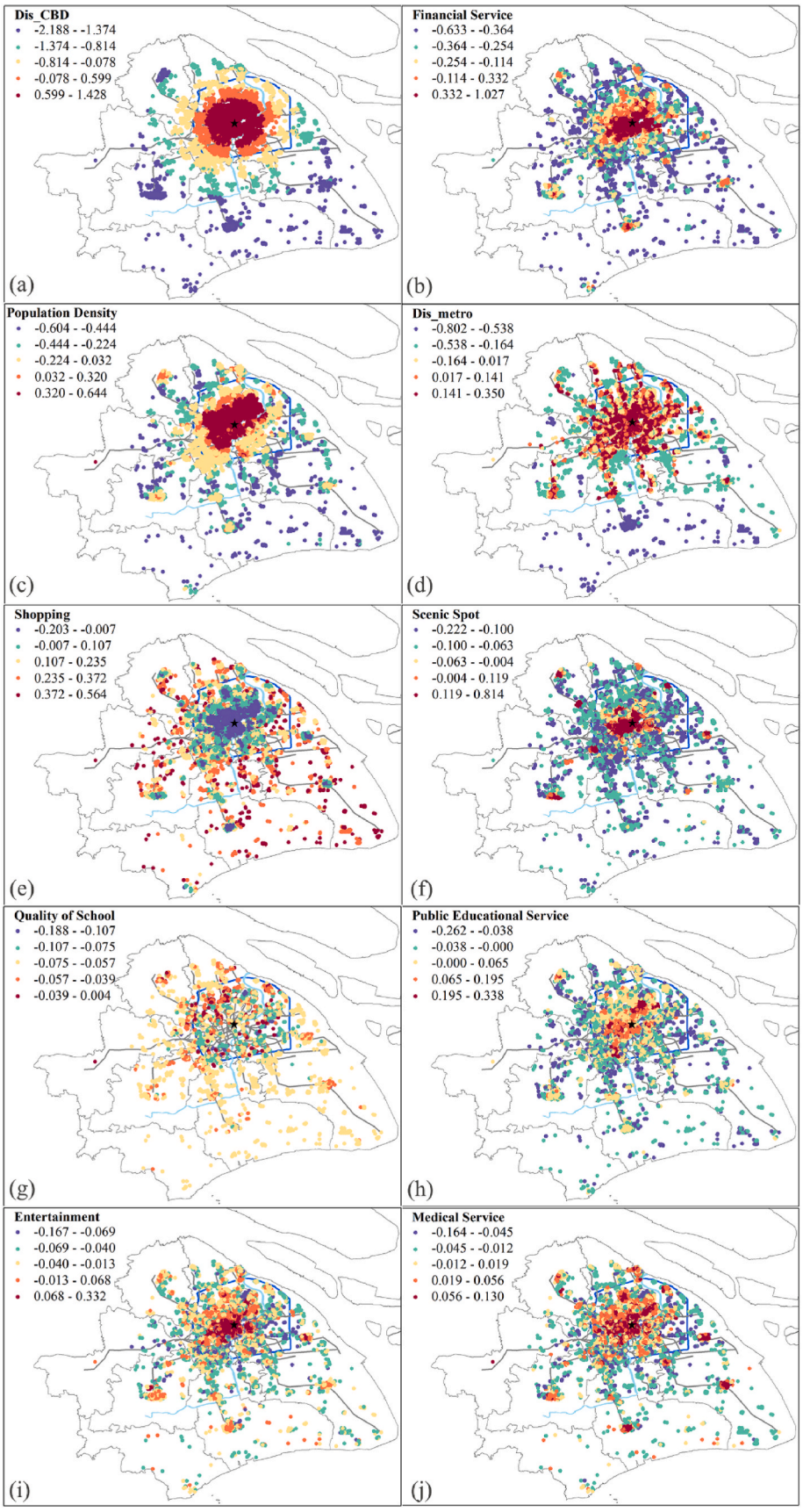


Fig. 9. Spatial distributions of the SHAP main effect value of the top 10 variables.

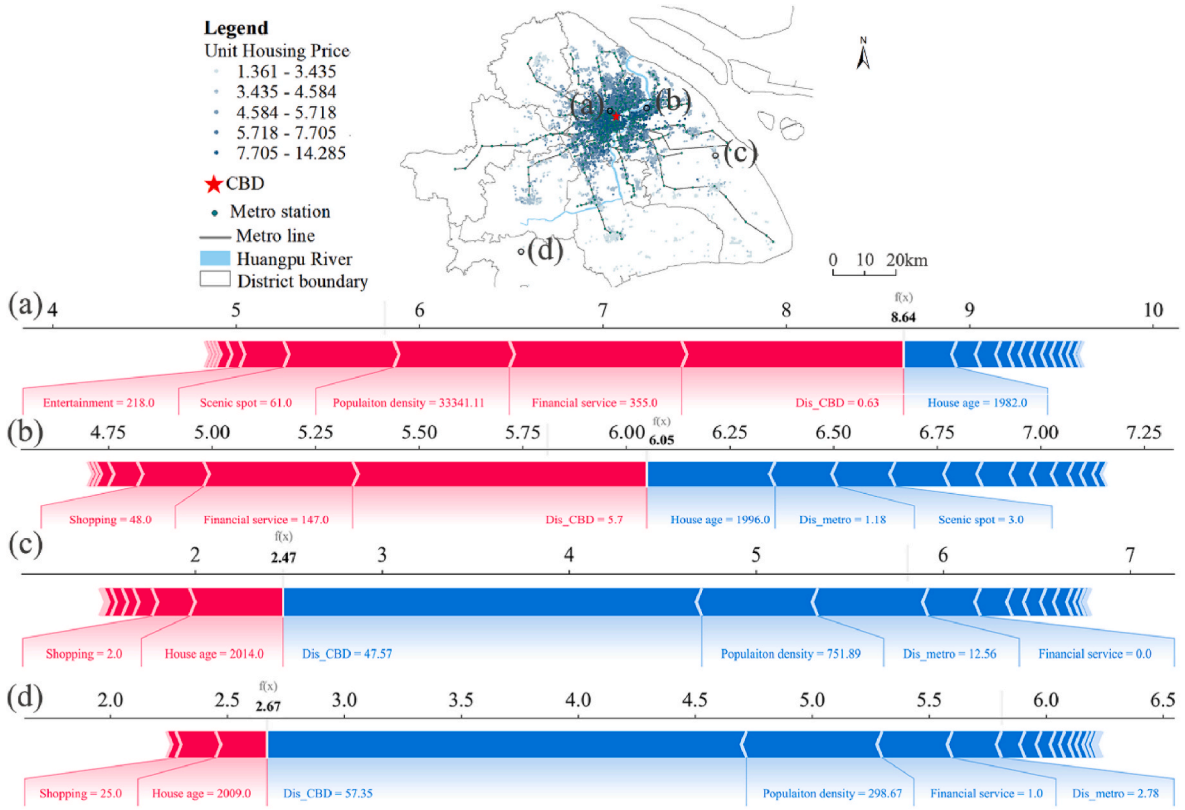


Fig. 10. SHAP individual explanations of housing prices.

which guarantees the accuracy and consistency of the model mathematically. Meanwhile, it can also be used at both a local and global level, which provides global importance and specific individual variability.

This study provides empirical evidence for utilizing the XAI to model housing prices. From a global perspective, the results evaluate and rank the importance of features with desirable model performance. From a local perspective, it reveals the nonlinear effects of public service amenities, private service amenities, and street view on housing prices. Especially fine-scale spatial patterns of SHAP main effect values highlight the heterogeneous spatial effects of different attributes. For example, the location attributes have obvious distance decay, while neighborhood variables exhibit distinct spatial characteristics. The geospatial heterogeneity and association model revealed could shed light on the application potential in identifying spatial sub-market segmentation (Rey-Blanco et al., 2022), and informing better equitable housing policies (Hu et al., 2019).

6. Conclusions

Using fine-scale housing transaction data in Shanghai, this study applied the XGBoost and SHAP models to investigate the nonlinear effects between neighborhoods and housing prices. Neighborhoods factors, including population density, public service amenities, private service amenities, and street view, collectively account for 45.48% of the overall effect on forecasting housing prices. This finding supports the idea that the local neighborhood environment has a significant impact on housing prices. Specifically, PUSA (public service amenities) has a greater influence on housing prices than PRSA (private service amenities), with a contribution of 23.18% and 7.06%, respectively. More importantly, all PUSA and some PRSA (e.g., entertainment) show threshold effects and a positive association with housing prices. Other PRSAs (e.g., shopping and carting services) show a negative association. The results also determined the effective value ranges for the PUSA and PRSA variables. For example, public educational services and entertainment would have the biggest effect on housing prices when the number of public educational services reached 75/km^2. The scenic spots have a significant positive effect when the density is lower than 60/km^2. When the density of financial services is lower than 300/km^2, their contributions to housing prices are linear. These findings provide substantial evidence for understanding the relationship between housing prices and neighborhoods. Furthermore, the street view has a considerable impact on house values. The most important variable of the street view is the green view, with a positive contribution of 0.99%, followed by the building view (0.82%), and the road view (0.72%).

Our study also has limitations. First, considering the geographical heterogeneity, the nonlinear relations between determined variables and housing prices may be different in other megacities. It would be helpful if these findings could be tested in more megacities to see whether they could be applied. Second, the complexity and variety of locational determinants of housing prices implies that neighborhood characteristics may have interactive or synergistic impacts (e.g., floor level and landscapes) on housing prices (Xiao et al., 2019), which this study ignored. These impacts should be examined in future research to better inform urban planning and improve housing affordability.

Credit author statement

Mingxuan Dou: Methodology, Software, Writing - Original Draft. **Yanyan Gu:** Formal analysis, Writing - Review & Editing, Validation. **Hong Fan:** Supervision.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the National Natural Science Foundation

of China (Grant No. 42101464), the China Postdoctoral Science Foundation (BX20220237), Open Research Fund Program of LIESMARS (Grant No. 21I03), and LIESMARS Special Research Funding.

References

Bangura, M., & Lee, C. L. (2020). House price diffusion of housing submarkets in Greater Sydney. *Housing Studies, 35*(6), 1110–1141. https://doi.org/10.1080/02673037.2019.1648772

Bourassa, S. C., Hoesli, M., Merlin, L., & Renne, J. (2021). Big data, accessibility and urban house prices. *Urban Studies, 58*(15), 3176–3195. https://doi.org/10.1177/0042098020982508

Cao, K., Diao, M., & Wu, B. (2019). A big data–based geographically weighted regression model for public housing prices: A case study in Singapore. *Annals of the Association of American Geographers, 109*(1), 173–186. https://doi.org/10.1080/24694452.2018.1470925

Cellmer, R., Cichulska, A., & Bełej, M. (2020). Spatial analysis of housing prices and market activity with the geographically weighted regression. *ISPRS International Journal of Geo-Information, 9*(6), 380. https://doi.org/10.3390/ijgi9060380

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16)* (pp. 785–794). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Chen, L., Yao, X., Liu, Y., Zhu, Y., Chen, W., Zhao, X., … Chi, T. (2020). Measuring impacts of urban environmental elements on housing prices based on multisource data—a case study of Shanghai, China. *ISPRS International Journal of Geo-Information, 9*(2), 106. https://doi.org/10.3390/ijgi9020106

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Lecture notes in computer science: 11211. Computer vision – ECCV 2018*. Cham: Springer. https://doi.org/10.1007/978-3-030-01234-2_49.

Chiang, Y., Peng, T., & Chang, C. (2015). The nonlinear effect of convenience stores on residential property prices: A case study of Taipei, taiwan. *Habitat International, 46*, 82–90. https://doi.org/10.1016/j.habitatint.2014.10.017

Chi, B., Dennett, A., Oléron-Evans, T., & Morphet, R. (2020). Shedding new light on residential property price variation in england: A multi-scale exploration. *Environment and Planning B: Urban Analytics and City Science, 48*(7), 1895–1911. https://doi.org/10.1177/2399808320951212

Dou, M., Wang, Y., & Dong, S. (2021). Integrating network centrality and node-place model to evaluate and classify station areas in Shanghai. *ISPRS International Journal of Geo-Information, 10*(6), 414. https://doi.org/10.3390/ijgi10060414

Du, Q., Wu, C., Ye, X., Ren, F., & Lin, Y. (2018). Evaluating the effects of landscape on housing prices in urban China. *Tijdschrift voor Economische en Sociale Geografie, 109*(4), 525–541. https://doi.org/10.1111/tesg.12308

Feng, H., & Lu, M. (2013). School quality and housing prices: Empirical evidence from a natural experiment in Shanghai, China. *Journal of Housing Economics, 22*(4), 291–307. https://doi.org/10.1016/j.jhe.2013.10.003

Feng, S., Peng, C., Yang, C., & Chen, P. (2021). Non-linear relationships between house size and price. *International Journal of Strategic Property Management, 25*(3), 240–253. https://doi.org/10.3846/ijspm.2021.14607

Fu, X., Jia, T., Zhang, X., Li, S., & Zhang, Y. (2019). Do street-level scene perceptions affect housing prices in Chinese megacities? An analysis using open access datasets and deep learning. *PLoS One, 14*(5), Article e217505. https://doi.org/10.1371/journal.pone.0217505

Gao, G., Bao, Z., Cao, J., Qin, A. K., & Sellis, T. (2022). Location-centered house price prediction: A multi-task learning approach. *ACM Transactions on Intelligent Systems and Technology, 13*(2). https://doi.org/10.1145/3501806

Gu, Y., Shi, R., Zhuang, Y., Li, Q., & Yue, Y. (2023). How to determine city hierarchies and spatial structure of a megaregion? *Geo-spatial Information Science*, 1–13. https://doi.org/10.1080/10095020.2022.2161425

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy, 82*, 657–673. https://doi.org/10.1016/j.landusepol.2018.12.030

Hu, Y., Lu, B., Ge, Y., & Dong, G. (2022). Uncovering spatial heterogeneity in real estate prices via combined hierarchical linear model and geographically weighted regression. *Environment and Planning B: Urban Analytics and City Science, 49*(6), 1715–1740. https://doi.org/10.1177/23998083211063885

Iban, M. C. (2022). An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants. *Habitat International, 128*, Article 102660. https://doi.org/10.1016/j.habitatint.2022.102660

Jia, J., & Zhang, X. (2021). A human-scale investigation into economic benefits of urban green and blue infrastructure based on big data and machine learning: A case study of wuhan. *Journal of Cleaner Production, 316*, Article 128321. https://doi.org/10.1016/j.jclepro.2021.128321

Jia, J., Zhang, X., Huang, C., & Luan, H. (2022). Multiscale analysis of human social sensing of urban appearance and its effects on house price appreciation in Wuhan, China. *Sustainable Cities and Society, 81*, Article 103844. https://doi.org/10.1016/j.scs.2022.103844

Jim, C. Y., & Chen, W. Y. (2006). Impacts of urban environmental elements on residential housing prices in Guangzhou (China). *Landscape and Urban Planning, 78*(4), 422–434. j.landurbplan.2005.12.003.

Jin, T., Cheng, L., Liu, Z., Cao, J., Huang, H., … Witlox, F. (2022). Nonlinear public transit accessibility effects on housing prices: Heterogeneity across price segments. *Transport Policy, 117*, 48–59. https://doi.org/10.1016/j.tranpol.2022.01.004

Li, H., Chen, P., & Grant, R. (2021). Built environment, special economic zone, and housing prices in Shenzhen, China. *Applied Geography, 129*, Article 102429. https://doi.org/10.1016/j.apgeog.2021.102429

Liu, N., & Strobl, J. (2022). Impact of neighborhood features on housing resale prices in Zhuhai (China) based on an (M)GWR model. *Big Earth Data*, 1–24. https://doi.org/10.1080/20964471.2022.2031543

Liu, Y., Yu, S., & Sun, T. (2021). Heterogeneous housing choice and residential mobility under housing reform in China: Evidence from Tianjin. *Applied Geography, 129*, Article 102417. https://doi.org/10.1016/j.apgeog.2021.102417

Li, H., Wei, Y. D., Wu, Y., & Tian, G. (2019). Analyzing housing prices in Shanghai with open data: Amenity, accessibility and urban structure. *Cities, 91*, 165–179. https://doi.org/10.1016/j.cities.2018.11.016

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., … Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems (NIPS'17)* (pp. 4768–4777). Long Beach, California, USA: Curran Associates Inc.. https://doi.org/10.5555/3295222.3295230

Ma, J., Cheng, J. C., Jiang, F., Chen, W., & Zhang, J. (2020). Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. *Land Use Policy, 94*, Article 104537. https://doi.org/10.1016/j.landusepol.2020.104537

Pérez-Molina, E. (2022). Exploring a multilevel approach with spatial effects to model housing price in San José, Costa Rica. *Environment and Planning B: Urban Analytics and City Science, 49*(3), 987–1004. https://doi.org/10.1177/23998083211041122

Qiu, W., Zhang, Z., Liu, X., Li, W., Li, X., Xu, X., … Huang, X. (2022). Subjective or objective measures of street environment, which are more effective in explaining housing prices? *Landscape and Urban Planning, 221*, Article 104358. https://doi.org/10.1016/j.landurbplan.2022.104358

Rey-Blanco, D., Arbués, P., López, F. A., & Páez, A. (2022). Using machine learning to identify spatial market segments. *A reproducible study of major Spanish markets. Environment and Planning B: Urban Analytics and City Science.* , Article 23998083231166952. https://doi.org/10.1177/23998083231166952

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16)* (pp. 1135–1144). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Sisman, S., & Aydinoglu, A. C. (2022). A modelling approach with geographically weighted regression methods for determining geographic variation and influencing factors in housing price: A case in istanbul. *Land Use Policy, 119*, Article 106183. https://doi.org/10.1016/j.landusepol.2022.106183

Soltani, A., Heydari, M., Aghaei, F., & Pettit, C. J. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. *Cities, 131*, Article 103941. https://doi.org/10.1016/j.cities.2022.103941

Su, S., Zhang, J., He, S., Zhang, H., Hu, L., … Kang, M. (2021). Unraveling the impact of TOD on housing rental prices and implications on spatial planning: A comparative analysis of five Chinese megacities. *Habitat International, 107*, Article 102309. https://doi.org/10.1016/j.habitatint.2020.102309

Taecharungroj, V. (2021). Google Maps amenities and condominium prices: Investigating the effects and relationships using machine learning. *Habitat International, 118*, Article 102463. https://doi.org/10.1016/j.habitatint.2021.102463

Xiao, Y., Hui, E. C. M., & Wen, H. (2019). Effects of floor level and landscape proximity on housing price: A hedonic analysis in hangzhou, China. *Habitat International, 87*, 11–26. https://doi.org/10.1016/j.habitatint.2019.03.008

Xu, M., Xin, J., Su, S., Weng, M., & Cai, Z. (2017). Social inequalities of park accessibility in Shenzhen, China: The role of park quality, transport modes, and hierarchical socioeconomic characteristics. *Journal of Transport Geography, 62*, 38–50. https://doi.org/10.1016/j.jtrangeo.2017.05.010

Yang, L., Liang, Y., Zhu, Q., & Chu, X. (2021). Machine learning for inference: Using gradient boosting decision tree to assess non-linear effects of bus rapid transit on house prices. *Annals of GIS, 27*(3), 273–284. https://doi.org/10.1080/19475683.2021.1906746

Yuan, F., Wei, Y. D., & Wu, J. (2020). Amenity effects of urban facilities on housing prices in China: Accessibility, scarcity, and urban spaces. *Cities, 96*, Article 102433. https://doi.org/10.1016/j.cities.2019.102433

Zeng, J., Yue, Y., Gao, Q., Gu, Y., & Ma, C. (2022). Identifying localized amenities for gentrification using a machine learning-based framework. *Applied Geography, 145*, Article 102748. https://doi.org/10.1016/j.apgeog.2022.102748

Zhang, P., Hu, S., Li, W., Zhang, C., Yang, S., … Qu, S. (2021). Modeling fine-scale residential land price distribution: An experimental study using open data and machine learning. *Applied Geography, 129*, Article 102442. https://doi.org/10.1016/j.apgeog.2021.102442