land

MDPI

Article

Understanding the Effects of Influential Factors on Housing Prices by Combining Extreme Gradient Boosting and a Hedonic Price Model (XGBoost-HPM)

Sheng Li [1,2], Yi Jiang [2], Shuisong Ke [2], Ke Nie [1,3] and Chao Wu [4,5,*]

1 Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518034, China; Leesheng@whu.edu.cn (S.L.); nieke@whu.edu.cn (K.N.)
2 Shenzhen Municipal Planning & Land Real Estate Information Centre, Shenzhen 518034, China; jiangbirdman@163.com (Y.J.); gtj_ke@163.com (S.K.)
3 Shenzhen Research Center of Digital City Engineering, Shenzhen 518034, China
4 School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
5 Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
* Correspondence: chaowu@njupt.edu.cn; Tel.: +86-1582-727-5160

Abstract: The characteristics of housing and location conditions are the main drivers of spatial differences in housing prices, which is a topic attracting high interest in both real estate and geography research. One of the most popular models, the hedonic price model (HPM), has limitations in identifying nonlinear relationships and distinguishing the importance of influential factors. Therefore, extreme gradient boosting (XGBoost), a popular machine learning technology, and the HPM were combined to analyse the comprehensive effects of influential factors on housing prices. XGBoost was employed to identify the importance order of factors and HPM was adopted to reveal the value of the original non-market priced influential factors. The results showed that combining the two models can lead to good performance and increase understanding of the spatial variations in housing prices. Our work found that (1) the five most important variables for Shenzhen housing prices were distance to city centre, green view index, population density, property management fee and economic level; (2) space quality at the human scale had important effects on housing prices; and (3) some traditional factors, especially variables related to education, should be modified according to the development of the real estate market. The results showed that the demonstrated multisource geo-tagged data fusion framework, which integrated XGBoost and HPM, is practical and supports a comprehensive understanding of the relationships between housing prices and influential factors. The findings in this article provide essential implications for informing equitable housing policies and designing liveable neighbourhoods.

Keywords: housing prices; variations; XGBoost; HPM; Shenzhen

Citation: Li, S.; Jiang, Y.; Ke, S.; Nie, K.; Wu, C. Understanding the Effects of Influential Factors on Housing Prices by Combining Extreme Gradient Boosting and a Hedonic Price Model (XGBoost-HPM). *Land* **2021**, *10*, 533. https://doi.org/10.3390/land10050533

Academic Editors: Shiliang Su, Shenjing He and Monika Kuffer

Received: 7 April 2021
Accepted: 11 May 2021
Published: 18 May 2021

1. Introduction

In the last decade, housing prices have become one of the top issues in economic development and for determining whether urban residents can live a better life [1–3]. The rapid growth of housing prices and spatial differentiation greatly concern managers, scholars, developers and residents [4–6]. As cities continue to expand and renew, the non-uniformity of spatial reconstruction and resource allocation is becoming increasingly prominent, which accelerates the spatial variations in housing prices within cities [5,7–10]. Understanding the mechanisms influencing spatial variations in housing prices is essential to formulate scientific housing policies, divide submarkets, optimize urban spatial layouts, allocate public infrastructure and equalize spatial resources [11,12]. Greater efforts

and improvements to previous studies are required to deeply understand the complex relationships between housing prices and influential factors.

First, although the hedonic price model (HPM) has been widely applied to housing prices and can identify the economic value of influential factors well [5,13,14], the traditional HPM has been criticized for some limitations, including: (1) a poor ability to reduce the impact of collinearity; (2) the assumption of linear relationships between influential factors and housing prices; and (3) a lack of robustness in the results [15–18]. The above limitations of the HPM might directly reduce the accuracy of housing price modelling and muddle our overall understanding of the influential factors of housing prices; thus, housing prices modelling should be improved by applying new data sources, methods and technologies [5,19].

Second, despite some efforts [20,21], studies using eye-level data to evaluate the neighbourhood environment and explore the effects of factors at the human scale on housing prices are rather limited. Related literature has shown that spatial differentiation in housing prices is the result of the effects of internal and external factors [11,22–29]. Specifically, the internal factors of housing prices include housing type, housing age, community environment, hardware facilities, property services, etc. The external factors can be classified into location conditions, traffic accessibility, landscape environment, living facilities, spatial quality, etc. Many studies have proven that street view images and semantic segmentation methods can be employed to derive human perceptual evaluations of the local environment [30,31]. In fact, people's perceptions of streetscapes, such as openness, security, walkability, development intensity, and greenness, can significantly affect their willingness to pay housing prices [32–34]. However, these factors at the human scale are rarely discussed in real estate studies.

Increasingly available big geo-data and machine learning demonstrate considerable potential to alleviate some of the aforementioned limitations. First, a combination of extreme gradient boosting (XGBoost) and the traditional HPM was applied in this article with the aim of identifying the spatial variations in housing prices more accurately. XGBoost is superior in choosing influential factors that have a significant effect on model accuracy, which can reduce the risk of model overfitting and identify the importance of variables [20,35]. The HPM was used to evaluate the qualified effects of influential factors on housing prices. Second, the wide application of street view images and semantic segmentation algorithms proves that as a new and available data source, street view data can be used for urban local environmental assessment on the human scale. An increasing number of studies suggest that it is necessary to use street view data to extract human-scale influential factors to study spatial variations in housing prices [20,36].

To date, limited effort has been exerted to integrate machine learning technologies and regression models to estimate spatial variations in housing prices [37], although these were proven effective in modelling geographic phenomena and spatial distributions in the fields of PM2.5, traffic fatalities and tropical cyclone intensity [38–40]. Therefore, this article proposes a multisource geo-tagged data fusion framework to estimate spatial variations in housing prices by extending the influential factors to the human scale and applying a combination of XGBoost and the HPM. Taking Shenzhen, China as the study area, this article analyses the main factors and combination characteristics that result in spatial differentiation in housing prices. We aim to contribute to the related literature in the following two ways: (1) an analytic framework combining XGBoost and the HPM is proposed to estimate spatial variations in housing prices; and (2) this article constructs a multidimensional and multilevel system of housing price influential factors using multi big geo-data, including building, point of interest (POI), road network, land-use, and street view data. In particular, based on street view data, a series of metrics for measuring spatial quality are proposed to reflect the physical and social structure characteristics of the urban local environment under fine spatial granularity. The research results of this paper are expected to provide references for constructing a theory of urban housing price

differentiation, promoting the healthy and stable development of the real estate market and constructing a liveable neighbourhood environment.

2. Analytic Framework

2.1. Overall Analytic Framework for Understanding Housing Prices

Figure 1 presents the overall methodological framework for understanding spatial variations in housing prices; it comprises data collection, factor quantification, modelling and results mapping and analysis. The first step was the collection of multi-source geo-datasets, including housing prices, community information, land-use data, POI data and building data. Second, we quantified the dependent variable (i.e., average home price) and independent variables. Third, the models, including XGBoost and the HPM, were implemented in turn to identify the spatially varying effects of influential factors on housing prices. Finally, the results were visualized and discussed from various perspectives, including the importance degree of influential factors and quantified relationships between housing prices and influential factors. In this article, housing price modelling adopting XGBoost and the HPM was built and implemented using Python 3.6.

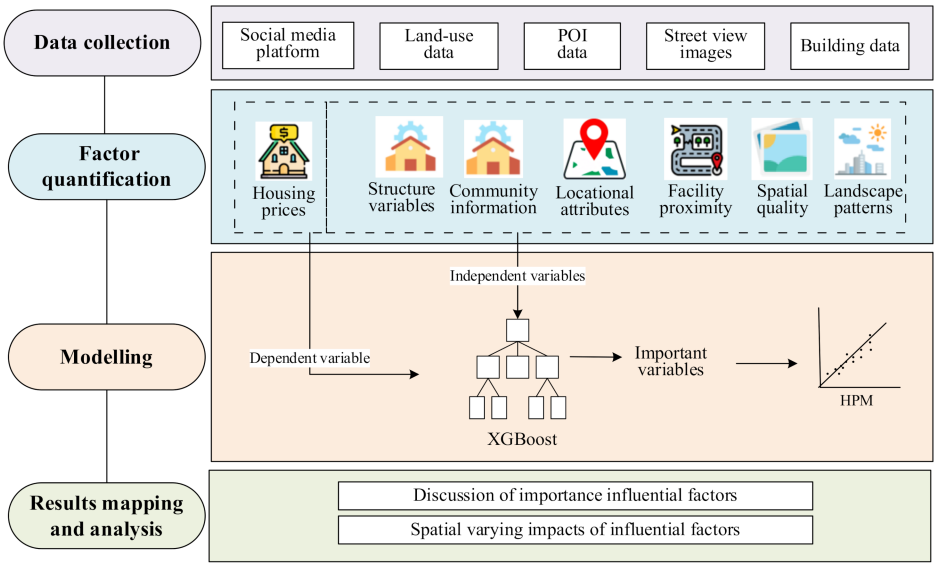

Figure 1. Overall methodological framework.

2.2. Data Collection

Our work was conducted in Shenzhen, one of the largest and fastest growing cities in China, which is located on the southeast coast of Guangdong Province. As a special economic zone, Shenzhen has experienced a series of highly urbanized development projects, urban renewal, space contraction and industrial upgrading, which have also provided many favourable policies to appeal to young people nationwide. The two main reasons for selecting Shenzhen as the study area are as follows: first, the immense housing demand, the shortage of residential land supply, and the tension between people and available land have inevitably led to the high cost of land and recent increases in housing prices in Shenzhen. To maintain a stable and healthy real estate market, it is necessary to understand the spatial variations and the influential factors of housing prices in Shenzhen. The results of our work can also provide some suggestions for the real estate development of similar cities. Second, there are abundant existing studies on housing prices in Shenzhen [14,41], which can be used as a reference and for comparison with our work.

In this article, housing price data for gated commercial communities were collected from one of the leading real estate information service platforms in China, the Anjuke website (https://shenzhen.anjuke.com/, accessed on 16 January 2021). Based on a massive

user dataset, Anjuke has launched research reports on real estate data, housing rentals, etc., analysing market trends and user behaviours, providing guidance, helping home buyers find apartments and providing references for developers to help them make decisions. Moreover, gated commercial community information, including the spatial location, green ratio, property management fee and plot ratio, can also be acquired from the Anjuke website. Therefore, through large-scale real estate data collection, a database of real housing information was constructed that can ensure the reliability and authenticity of housing information. This paper collected data on 12,137 housing units in a total of 3186 gated commercial communities. A 1 km buffer zone was constructed to designate the neighbourhood of a residential area [42,43], and all of the location-related characteristics of communities were thus aggregated within a 1 km buffer. The study area, Shenzhen, and the locations of the communities are shown in Figure 2.

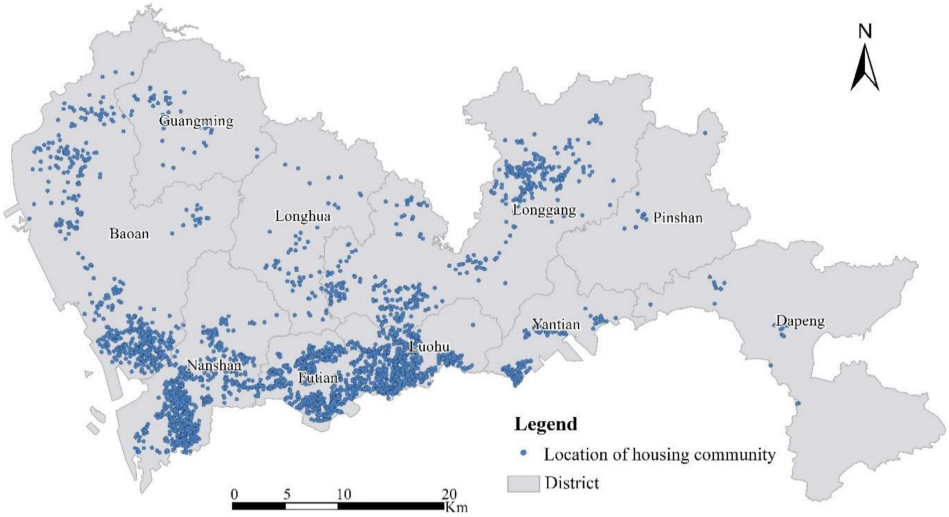

Figure 2. The study area: Shenzhen and the locations of the communities.

The POI data were obtained from an online mapping services provider called Auto map (https://lbs.amap.com/, accessed on 16 January 2021). POI data refer to geographic entities that are of interest to users within a specific spatial scope and that can be abstracted as points, such as schools, hospitals, parks, and supermarkets. The road network data were acquired from Open Street Map (OSM, https://www.openstreetmap.org/, accessed on 16 January 2021), which is a free, open source, editable map service created by the internet public. Economic development levels were extracted from Luojia-1 Night-time Light Imagery. Urban land-use data for 2019 (including residential, commercial, industrial, transportation, public management, and service) were acquired by land change investigation, which was provided by the Shenzhen Municipal Bureau of Planning and Natural Resources. The building data were acquired from the official department mentioned above to ensure the authority and reliability of the datasets.

Notably, street view images were used to measure the space quality at the human scale. At present, network map service providers, represented by Google Maps, Tencent Maps and Baidu Maps, can provide street view services. The street view images used in this article were obtained from Baidu Maps (http://quanjing.baidu.com/#/, accessed on 16 January 2021). Based on road network data obtained from OSM, the sampling points of the street view images were generated using 50 m intervals along the road network to ensure the continuity of the street landscapes and avoid data redundancy. To reflect the human perspective, four angles of 0°, 90°, 180° and 270° were selected for street view data collection to achieve comprehensive inclusion of the visual environment around each viewpoint. The 1 km buffer and examples of the data used are shown in Figure 3. An example of street view images from the four angles for a sample location is shown in Figure 4.

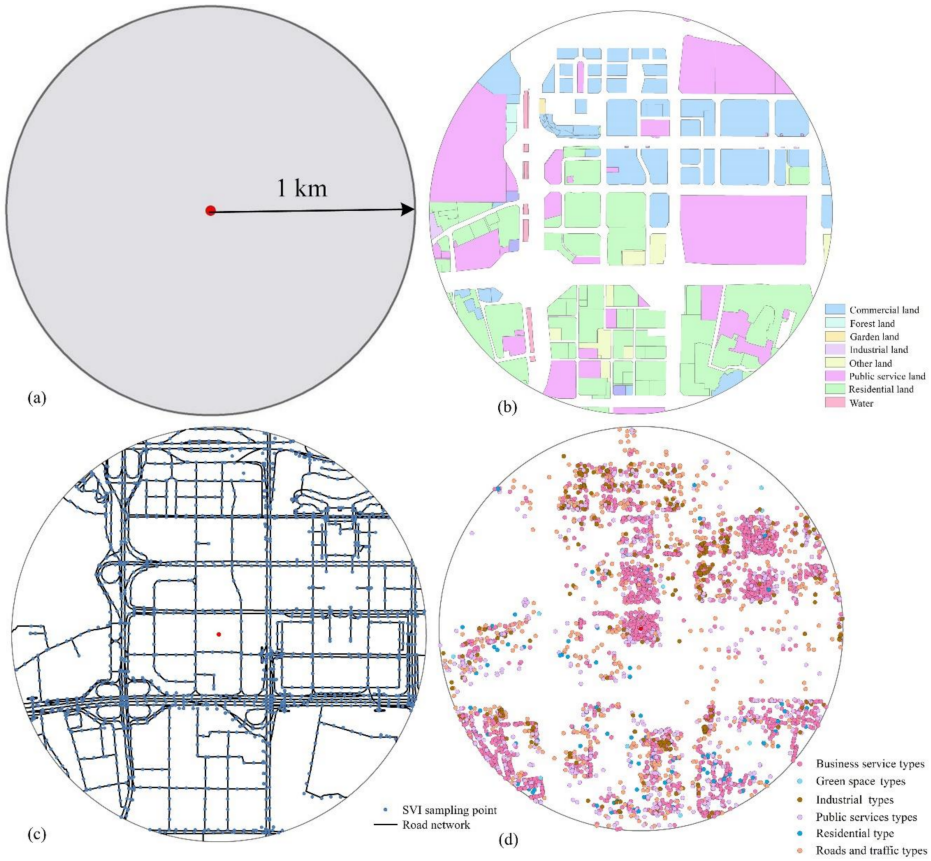

Figure 3. Examples of the data used to calculate influential factors for housing prices: (**a**) community location and 1 km buffer area; (**b**) land-use data in buffer area of sample; (**c**) road network and street view sampling points in buffer area; and (**d**) POI data in buffer area of sample.

Figure 4. Example of a street view image in a given sample location from four angles.

2.3. Factor Quantification

The acquisition of massive data information usually relies on network crawling tools. The average price of each residence was obtained from the social media platform Anjuke. To ensure that the housing price data conformed to the assumptions of the regression model, this paper used the HPM in semi-logarithmic form. That is, the distribution of housing prices was adjusted to a logarithmic scale. According to previous studies, housing prices are affected by many factors, and different influential factors result in spatial variations in housing prices. Therefore, a number of influential factors for housing prices were quantified from the perspectives of structure, community, facilities, location, space quality and landscape patterns [20,23,41].

2.3.1. Structural Information

Based on the information provided by the real estate information service platform Anjuke, the number of rooms and the number of halls were used to represent the shape and size of the residence. Its area and which floor it was on both have a direct impact on the lives of residents, affecting factors such as comfort, views and light exposure.

2.3.2. Community Information

This article selected the construction year of the community, plot ratio, green rate and property management fee to represent community information. The plot ratio indicator is an important index reflecting the intensity of land development and the efficiency of land use. The higher the residential plot ratio is, the greater the development intensity and the higher the land utilization rate. However, plot ratios that are too high affect both the quality of the urban landscape and the living environment [44]. The green rate of a community is one of the most direct measurements of green, ecological and healthy communities. Although the property management fee is not the only standard used to measure the quality of property management services, many property management companies in residential areas charge higher fees to provide better services. The main reason is that high-quality services require more manpower and material resources; property management fees will thus have an impact on the quality of property services to some extent [45].

2.3.3. Locational Attributes

This paper mainly reflects the locational attributes of residential areas by considering two perspectives: traffic conditions and socio-economic activities. The evaluation of traffic conditions mainly depended on the bus stations, subway stations and road networks near the community, including the number of subway entrances and bus stations in the neighbourhood, the distance to the nearest subway entrance and bus stations, and the road density in the neighbourhood. The factors used for evaluating socio-economic activities included the distance to the city centre (Shenzhen Civic Centre in Futian district), the population density of the 1 km buffer, and the economic development level represented by night-time light imagery [46].

2.3.4. Facility Proximity

Facilities, including educational facilities, medical facilities, recreational facilities and commercial facilities, have an important influence on the convenience and quality of residents' daily lives. The proximity of educational facilities was measured by the number of kindergartens in a 1 km buffer, the number of primary and secondary schools, the number of high schools, the distance to the nearest kindergarten, and the distance to the nearest primary and secondary schools. Medical facilities were measured by the number of AAA class hospitals (3A hospitals are some of the highest-level hospitals) in the 1 km buffer. The proximity of recreational facilities was quantified using the number of community parks in the 1 km buffer and distance to the nearest community park. Finally, the proximity of commercial facilities was determined by using factors including the number of supermarkets in the 1 km buffer, the distance to the nearest supermarket, and the distance to the nearest farmers' market.

2.3.5. Space Quality

Space quality is a concept formed to evaluate space by reflecting the comprehensive demand of the population for urban space. As an intuitive representation of streets, street view images can be segmented into streetscape scenes by semantic segmentation algorithms, which can be applied to measure the quality of urban space. This study adopted SegNet, an efficient pixel-level semantic segmentation algorithm, to segment a street view [47]. SegNet is a full consolidation neural network consisting of encoders and decoders. Finally, each pixel is classified by the softmax layer. One of the most obvious innovations of SegNet is the sampling method of its decoder to low resolution. SegNet can segment a street view image into 12 types of streetscapes. The green view index, sky openness (openness) and walkability, as measured by scene elements, were adopted to measure the space quality of the neighbourhood. In addition, the average normalized difference vegetation index (NDVI) of the 1 km buffer was included to measure space quality. A good visual environment provides places for residents to rest and engage in leisure activities.

2.3.6. Landscape Patterns

Landscape patterns can be used to reflect both ecological and socio-economic functions, which both affect housing prices [41,48]. This article adopted landscape metrics based on land-use data to describe the landscape patterns of neighbourhoods. Three selected metrics measured at the landscape scale were Shannon's diversity index (SHDI), patch density and patch area. Specifically, Shannon's diversity index quantifies diversity by describing the uncertainty that occurs among individuals in a population. The higher the uncertainty is, the higher the diversity, and vice versa. Patch density is the ratio of patch number to patch area, which can be used to represent landscape fragmentation. Patch area is used to represent the area of patch and is a shape metric. The three selected factors are independent from one another and do not have a strong correlation, which can reveal the multidimensional features of landscape patterns and remove the risk of information redundancy. In addition, the three factors reflect the characteristics of landscape patterns from different perspectives, thereby offering a comprehensive view. Overall general descriptive statistics for the selected influential factors are shown in Table 1.

Table 1. Descriptive statistics of the influential factors of housing price.

Variable	Minimum	Maximum	Mean	SD
Rooms in residence	1.000	8.000	3.000	0.921
Halls in residence	0.000	5.000	2.000	0.398
Area of residence	18.000	276.000	94.537	38.597
Floor of residence	1.000	33.000	16.844	10.770
Construction year of community	1980.000	2020.000	2008.000	8.000
Plot ratio	0.350	20.970	3.273	1.849
Greening rate	0.200	0.800	36.200	10.696
Property management fee	0.100	23.800	2.964	1.426
Number of subway entrances in 1 km buffer	0.000	7.000	1.331	1.317
Number of bus stations in 1 km buffer	2.000	88.000	35.364	16.858
Distance to the nearest subway entrance	0.053	18.731	1.439	2.563
Distance to the nearest bus station	0.002	0.889	0.194	0.137
Road density of 1 km buffer	0.000	3.976	1.375	0.683
Distance to city centre	0.593	44.576	18.869	10.523
Number of kindergartens in 1 km buffer	0.000	14.000	3.000	2.416
Number of primary schools in 1 km buffer	0.000	9.000	1.000	1.370
Number of middle schools in 1 km buffer	0.000	5.000	1.000	0.999
Distance to the nearest kindergarten	0.014	4.939	0.587	0.532
Distance to the nearest primary school	0.048	10.618	1.063	1.060
Distance to the nearest middle school	0.047	9.753	1.284	1.518
Number of 3A hospitals in 1 km buffer	0.000	2.000	0.140	0.408
NDVI	0.118	0.875	0.511	0.125
Number of community parks in 1 km buffer	0.000	30.000	3.734	4.994
Distance to the nearest community park	0.000	2.642	0.578	0.459
Number of supermarkets in 1 km buffer	0.000	186.000	39.812	33.303
Distance to the nearest farmers' market	0.004	1.747	0.278	0.267
Economic level of 1 km buffer	2852.282	211,382.154	44,575.631	20,014.800
POI-based mixed use of 1 km buffer	0.754	1.646	1.127	0.139
Population density of 1 km buffer	220.000	59,513.000	13,912.000	12,238.880
Patch density	0.495	99.000	23.433	18.483
Patch area	1.010	202.005	9.862	15.629
Shannon's diversity index	0.001	1.681	0.745	0.347
Green view index	0.047	0.629	0.287	0.090
Walkability index	0.000	0.016	0.005	0.002
Openness index	0.158	0.701	0.366	0.092

2.4. Modelling Methods

2.4.1. XGBoost-Based Identification of Important Independent Variables

XGBoost is a boosting integration model that combines the gradient lift algorithm and decision trees, specifically using several preferred weak learners (i.e., decision trees) to complete the learning task [35,49,50]. Instead of using the search method, XGBoost

directly utilizes the first and second derivative values of the loss function and improves the performance of the algorithm through techniques, such as pre-ordering and node number of bits. After introducing the regularization term, the XGBoost model chooses a simple model with good performance. The regularization item is used to suppress weak learner overfitting in each iteration and does not participate in the integration of the final model. In each iteration, the objective function is expanded by Taylor's formula, as in Equation (1):

$$L^{(t)} \simeq \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i{}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2}h_i f_t(x_i)^2\right] + \Omega(f_t) \quad (1)$$

where t denotes the tth interaction, i is the ith sample, and y_i is the real value of the ith sample; $\hat{y}_i{}^{(t-1)}$ represents the predictive value of the $(t-1)$th iteration; g_i and h_i are the first and second derivatives, respectively; and $\Omega(f_t)$ is the regularization item. The complexity of the tree is shown in Equation (2):

$$\Omega = \gamma^{T_t} + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \quad (2)$$

where T_t is the number of leaf nodes in the round t iteration, and ω_j represents the weight of the jth leaf node.

In the process of constructing the decision tree, sorting the values of the features to determine the optimal split point is the most extensive step. The greatest advantage of XGBoost is that the data features are sorted before the training and then stored as blocks. As a result, the existing blocks can be used for subsequent iterations, which greatly reduces the amount of computation required. XGBoost yields values measuring feature importance that can identify how important each feature is in its feature set. In this article, XGBoost was used to analyse the feature contributions and calculate the importance of the influential factors, which can deliver effective inputs of independent variables for HPM modelling.

2.4.2. Hedonic Price Model-Based Exploration of the Effects of Influential Factors

The HPM is the most commonly used method for studying the market values of the influential factors of housing prices, which are non-market prices. In 1974, Rosen [51] first applied the HPM to real estate research and it has been widely used since to study the marginal price of factors influencing housing prices. The essence of the HPM is ordinary linear regression, which can reveal the marginal price of housing property. The HPM has a variety of forms, such as logarithmic form, semi-logarithmic form and exponential form. The semi-logarithmic equation can effectively solve or reduce the heteroscedasticity problem, and its form is simple. Therefore, this paper modelled housing prices and their influential factors in semi-logarithmic form. The HPM model in semi-logarithmic form is as follows:

$$\mathrm{Log}P_i = \beta_0 + \sum\beta_j S_{ij} + \sum\beta_k C_{ik} + \sum\beta_l L_{il} + \sum\beta_m F_{im} + \sum\beta_n Q_{in} + \varepsilon \quad (3)$$

where P_i is the home price, and $\mathrm{Log}P_i$ is the logarithm base 10 of the home price. S_{ij}, C_{ik}, L_{il}, F_{im} and Q_{in} represent the structural, community, locational, facility and space quality-related factors. β_0 is the intercept item. β_j, β_k, β_l, β_m and β_n are estimated coefficients for the structural, community, locational, facility and space quality-related factors. That is, they can be understood as semi-elastic prices.

3. Results

3.1. The Results of XGBoost and the HPM

To understand the significant inequality evident in the housing prices of Shenzhen, it was necessary to study the influencing mechanisms behind those prices. First, XGBoost was used to explore the importance of different factors of housing prices by revealing non-linear relationships between influential factors and housing prices. The performance

of XGBoost is summarized in Table 2. It generally presented a good performance, with an R-square of 0.944. Other indicators, including root mean squared error (RMSE) and its percentage (%RMSE), mean absolute error (MAE) and its percentage (%MAE) and accuracy (P), were also used to evaluate the performance of the XGBoost model. Specifically, the validation indicators of the prediction results showed that the RMSE was 0.057, the %RMSE was 1.203%, the MAE was 0.039, and the %MAE was 0.825%. The HPM was then applied to study the effects of influential factors on housing prices. The values of the variance inflation factor (VIF) were all less than 10, which indicates that there is no obvious multicollinearity between the variables. The R-square and RMSE in the HPM were 0.600 and 0.153, respectively, which indicates that almost 60% of the variation in housing prices can be explained. The statistical information of XGBoost and the HPM are shown in Table 2.

Table 2. Performance of the individual algorithms.

	XGBoost	**HPM**
RMSE	0.057	0.153
%RMSE	1.203%	3.232%
MAE	0.039	0.117
%MAE	0.825%	2.502%
P	0.992	0.975
R^2	0.944	0.600

3.2. The Importance of the Different Factors of Housing Prices Based on XGBoost

XGBoost is capable of measuring feature importance using their weights. Figure 5 shows the relative importance of the influential factors based on the XGBoost results. As shown in Figure 5, the top five variables were distance to city centre, green view index, population density, property management fee and economic level. The top five variables can predominantly explain the spatial variability of housing prices. The variable of distance to the city centre had the largest effect on home prices, and its effect was much larger than that of other variables. First, the economic development of Futian and Nanshan were very rapid. Traffic conditions are also becoming more developed and health care and education resources have become more advanced, which can attract many people to aggregate in these areas. Therefore, there is high and urgent demand for homes, which has caused a short supply in the property market. Second, the level of urbanization is higher, and land resources are scarcer. Land prices are high, which naturally drives high housing prices. Compared with other variables for green space, such as the greening ratio, NDVI and park accessibility, the green view index showed a much greater impact on housing prices. The green view index systematically, and in detail, records the greening quality at the urban street level from the perspective of pedestrians. The green view index provides the spatial distribution of greenness at the street level, which can directly and accurately reflect information on the facade of green space. This result also showed that street greening, which is more common and accessible to residents, should not be ignored in planning. The effect of population density on housing prices is also concentrated in supply and demand.

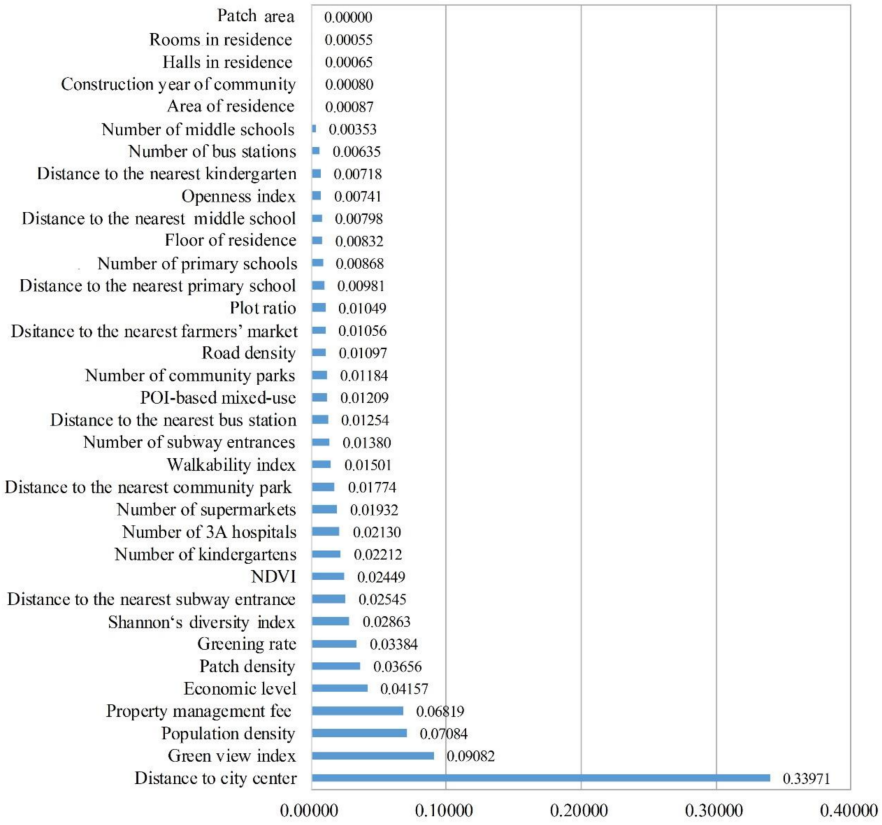

Figure 5. A list of the relative importance of influential factors on housing prices.

Following the top five variables, the landscape index, represented by patch density and Shannon's diversity index, also played a role in housing prices. A better urban ecological landscape can enhance the comfort of residents and improve the value and competitiveness of the city. This result also verified the effects of the urban ecological landscape on housing prices, consistent with a previous study [41]. Subway entrances, kindergartens, 3A hospitals, supermarkets and parks are facilities that are very important to the daily life of residents and that provide them with convenience [17]. The distance to schools had less of an effect on home prices than expected. An increasing number of studies have pointed out that the quality of public school districts in which housing is located has becoming increasingly important. Finally, the structure of a residence and its age had a relatively small effect on its price. Buyers paid more attention to the floor location of a home, as this is directly related to its lighting, ventilation and views.

3.3. Effects of Influential Factors on Housing Prices: HPM Results

A summary report of the HPM was obtained by inputting the attributes into ordinary least squares (Table 3). Some variables warrant attention according to Table 3. First, the structural factors of a residence did not have a significant effect on housing prices, except for the floor on which the home was located. This is consistent with the XGBoost results, which suggests that these structural variables were less important for home prices in Shenzhen. The floor of the home had a positive effect on its prices, and higher floors were generally perceived to have better views and lighting than lower floors. Second, for the variables related to community information, the plot ratio, greening rate and property management fee all had significant positive effects on housing prices. Notably, the plot ratio variable was related to building density, building height, building spacing and number of residents, which directly affected the residential quality of the community. Generally, the lower the plot ratio, the higher the residential quality of the community. One of the main reasons for the positive relationship between the plot ratio and housing prices was that,

given a background of high housing prices and tight land supply, high plot ratio residential communities have become mainstream.

Table 3. The estimated coefficients of the HPM.

Variable	Coefficient	Std Error	*t*	Probability
Constant	4.452	0.375	11.927	0.000 ***
Rooms in residence	−0.003	0.003	−0.309	0.757
Halls in residence	0.001	0.004	0.131	0.895
Area of residence	0.001	0.000	0.074	0.941
Floor of residence	0.009	0.007	1.461	0.044 **
Construction year of community	−0.004	0.000	−0.658	0.511
Plot ratio	0.060	0.001	8.660	0.000 ***
Greening rate	0.146	0.000	22.510	0.000 ***
Property management fee	0.155	0.001	22.756	0.000 ***
Number of subway entrances	0.177	0.002	16.635	0.000 ***
Number of bus stations	0.063	0.000	4.702	0.000 ***
Distance to the nearest subway entrance	0.001	0.000	0.040	0.968
Distance to the nearest bus station	0.052	0.000	6.233	0.000 ***
Road density	0.031	0.003	3.824	0.000 ***
Distance to city centre	−0.203	0.000	−17.450	0.000 ***
Number of kindergartens	−0.053	0.001	−6.244	0.000 ***
Number of primary schools	0.082	0.002	9.578	0.000 ***
Number of middle schools	−0.051	0.002	−6.587	0.000 ***
Distance to the nearest kindergarten	−0.160	0.000	−16.006	0.000 ***
Distance to the nearest primary school	0.117	0.000	9.312	0.000 ***
Distance to the nearest middle school	−0.045	0.000	−4.227	0.000 ***
Number of 3A hospitals	0.020	0.022	2.505	0.012 **
NDVI	−0.228	0.000	−20.403	0.000 ***
Number of community parks	0.070	0.000	6.833	0.000 ***
Distance to the nearest community park	−0.096	0.000	−12.486	0.000 ***
Number of supermarkets	−0.047	0.000	−3.713	0.000 ***
Distance to the nearest farmers' market	0.040	0.000	4.874	0.000 ***
Economic level	0.044	0.015	5.054	0.000 ***
POI-based mixed-use	0.194	0.000	22.972	0.000 ***
Population density	0.040	0.000	4.524	0.000 ***
Patch density	−0.090	0.000	−7.323	0.000 ***
Patch area	0.003	0.010	0.367	0.714
Shannon's diversity index	−0.160	0.035	−11.442	0.000 ***
Green view index	0.339	1.483	25.717	0.000 ***
Walkability index	−0.138	0.050	−10.217	0.000 ***
Openness index	0.028	0.005	1.456	0.145

Note: *** and ** represent the significance level of 1% and 5% respectively.

The factors for locational attributes and facility proximity, except for distance to the nearest subway entrance, all showed statistically significant effects on housing prices. Locational attributes and facility proximity were essential attributes of housing prices. For example, residences near city centres were favoured by families due to well-established cultural and athletic facilities, clusters of medical treatment and health organizations, and the presence of entertainment and amusement functions, which can be highly convenient and accessible for families. In addition, many high-rise office buildings cluster in Futian, encouraging a large number of high-income groups to aggregate there, thus effectively stimulating housing prices. The effects of some variables related to educational resources, such as the number of kindergartens, the number of middle schools and the distance to the nearest primary school, were unexpected. The main reason was that the housing price in school districts was promoted by educational resources, while the housing price in high-quality school districts was not affected by the distance from the school.

In this paper, a POI-based mixed-use variable and Shannon's diversity index were used to represent the degree of functional mix in a neighbourhood. However, the two

variables had very different effects on home prices. Specifically, the POI-based mixed-use variable had a significant positive effect on housing prices. Shannon's diversity index showed significant negative effects on housing prices for the following reasons: (1) there were obvious differences in the classifications used for POI and those for land use; (2) buyers focused much more on facilities, and POI data can better reflect interactions with people than land use. Finally, this paper adopted the street view image to determine the space quality of each neighbourhood. There were positive relationships between the green view index and housing prices. Compared with the factors of the number of community parks and distance to the nearest community park, the green view index had a larger effect on housing prices. Many studies show that eye-level photographs have important effects on residents' health and living environment [52]. Green view, as measured by street view images, can accurately assess people's daily exposure to greenery.

4. Discussion

4.1. Implications for Housing Policy

Identifying the relationships between housing prices and influential factors is crucial for policy decision making to optimize the urban infrastructure layout, develop a healthy and equitable real estate market, and design liveable neighbourhoods. First, the relationships between housing prices and factors are complex, and it is necessary to study them by combining advanced technologies with classical methods. Therefore, this paper adopted one of the most popular machine learning methods, XGBoost, and the classical housing price method, the HPM, to reveal the ranked importance of factors for housing prices and their quantified effects on housing prices. Notably, XGBoost has shown potential in providing insights for real estate studies and the appraisal of the real estate price. According to the research framework of this article, real estate-related management departments can monitor or assess urban property prices. On the one hand, a machine learning algorithm can be used to rapidly map the relative importance of factors that affect housing prices. On the other hand, using traditional characteristics shows factors affecting the price model of market value. The influential factors of housing prices should be continuously expanded, supplemented, and improved as new data and methods become available. For example, we used street view images and semantic segmentation to extract factors at the human scale. Street view images, as a new type of data covering urban areas and the general landscape, can fully represent the physical appearance of city space. These new data and methods have great potential and applicability for study because they reflect the living environment of residential districts from a human perspective. Finally, our work demonstrated that current urban infrastructure is not perfect, as indicated by features such as differences in school resources, subway distribution, etc. These imperfections are the main reasons for the difference in housing prices in cities. Therefore, in future urban design and planning, we should pay attention to the factors with important influence to promote the sustainable development of the real estate market and urban design. Overall, this paper provided a theoretical reference for the systematic, scientific and comprehensive evaluation of the factors influencing urban housing prices.

4.2. Limitations and Prospects

Although this paper applied new methods to explore the complex relationships between housing prices and influential factors, and further improves the system of influential factors for housing prices, we also recognize several limitations of this work that should be addressed in future studies. First, temporal variations in housing prices should also be considered. Housing price appreciation rates and influence mechanisms are equally important for buyers, property developers and governments [20]. Second, the proportions of the scenes extracted from street view images were not sufficient to evaluate space quality. Residents' perception of the neighbourhood environment, such as safety, liveliness, depression and wealth, should be evaluated based on street view images [53]. Finally, this paper proposed a framework integrating machine learning and a hedonic price model that

was only applied in Shenzhen, China. In the future, more overseas and domestic cities can be studied to improve the generalizability and reproducibility of the research framework presented in this paper.

5. Conclusions

A deep and comprehensive understanding of the spatial variations in housing prices and their influential factors is critical for housing price control, public facility construction and urban planning [54]. We presented a multisource geo-tagged data fusion framework integrating XGBoost and the HPM to study the complex relationships between housing prices and influential factors. Specifically, XGBoost and the HPM can reveal different aspects of the complex relationships between housing prices and influential factors by ranking the influential factors by importance and determining their quantified effects. The XGBoost results identified the five most important variables for Shenzhen housing prices as distance to city centre, green view index, population density, property management fee and economic level. The HPM results proved that green view index is a good objective indicator for measuring street-level greenery, which has significant and positive effects on housing prices. Understanding the variations in housing prices and their influential factors can better inform sustainable city planning and urban and housing policy making. In addition, some new factors at the human scale extracted from street view data were added to enrich the system of factors affecting housing prices. Our work also demonstrated that urban big data, machine learning and spatial statistical methods provide us with new data sources and enable interdisciplinary approaches to understanding the distribution of housing prices.

Author Contributions: Conceptualization, S.L. and C.W.; methodology, C.W. and K.N.; software, Y.J. and S.K.; validation, Y.J. and S.K.; data curation, K.N.; writing—original draft preparation, S.L. and C.W.; writing—review and editing, Y.J., S.K. and K.N.; visualization, C.W.; project administration, S.L.; funding acquisition, C.W. and K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (41901326 and 41801322), the Natural Science Foundation of Jiangsu Province (BK20190742), the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources (KF-2019-04-078), the Natural Science Research of Jiangsu Higher Education Institutions of China (Grant number 19KJB170009) and the Open Research Fund Program of Key Laboratory of Digital Mapping and Land Information Application Engineering (NASGZRZYBWD201906).

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yao, Y.; Zhang, J.; Hong, Y.; Liang, H.; He, J. Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Trans. GIS* **2018**, *22*, 561–581. [CrossRef]
2. Wen, H.; Goodman, A.C. Relationship between urban land price and housing price: Evidence from 21 provincial capitals in China. *Habitat Int.* **2013**, *40*, 9–17. [CrossRef]
3. Wu, C.; Ren, F.; Hu, W.; Du, Q. Multiscale geographically and temporally weighted regression: Exploring the spatiotemporal determinants of housing prices. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 489–511. [CrossRef]
4. Li, H.; Wei, Y.D.; Wu, Y.; Tian, G. Analyzing housing prices in Shanghai with open data: Amenity, accessibility and urban structure. *Cities* **2019**, *91*, 165–179. [CrossRef]
5. Wang, Y.; Wang, S.; Li, G.; Zhang, H.; Jin, L.; Su, Y.; Wu, K. Identifying the determinants of housing prices in China using spatial regression and the geographical detector technique. *Appl. Geogr.* **2017**, *79*, 26–36. [CrossRef]
6. Wu, C.; Ye, X.; Ren, F.; Du, Q. Modified Data-Driven Framework for Housing Market Segmentation. *J. Urban Plan. Dev.* **2018**, *144*, 04018036.
7. Hernandez, D. Uneven mobilities, uneven opportunities: Social distribution of public transport accessibility to jobs and education in Montevideo. *J. Transp. Geogr.* **2018**, *67*, 119–125. [CrossRef]
8. Hannum, E.; Meiyan, W. Geography and educational inequality in China. *China Econ. Rev.* **2006**, *17*, 253–265. [CrossRef]

9. Hu, L.; He, S.; Luo, Y.; Su, S.; Xin, J.; Weng, M. A social-media-based approach to assessing the effectiveness of equitable housing policy in mitigating education accessibility induced social inequalities in Shanghai, China. *Land Use Policy* **2020**, *94*, 104513. [CrossRef]
10. Wang, S.; Wang, J.; Wang, Y. Effect of land prices on the spatial differentiation of housing prices: Evidence from cross-county analyses in China. *J. Geogr. Sci.* **2018**, *28*, 725–740. [CrossRef]
11. Fik, T.J.; Ling, D.C.; Mulligan, G.F. Modeling spatial variation in housing prices: A variable interaction approach. *Real Estate Econ.* **2003**, *31*, 623–646. [CrossRef]
12. Pavlov, A.D. Space-varying regression coefficients: A semi-parametric approach applied to real estate markets. *Real Estate Econ.* **2000**, *28*, 249–283. [CrossRef]
13. Morano, P.; Tajani, F.; Guarini, M.R.; Di Liddo, F.; Anelli, D. A multivariate econometric analysis for the forecasting of the interdependences between the housing prices and the socio-economic factors in the city of Barcelona (Spain). In *Proceedings of the International Conference on Computational Science and Its Applications*; Springer: Cham, Switzerland, 2019; pp. 13–22.
14. Su, S.; Zhang, J.; He, S.; Zhang, H.; Hu, L.; Kang, M. Unraveling the impact of TOD on housing rental prices and implications on spatial planning: A comparative analysis of five Chinese megacities. *Habitat Int.* **2021**, *107*, 102309. [CrossRef]
15. Ju, H.; Zhang, Z.; Zuo, L.; Wang, J.; Zhang, S.; Wang, X.; Zhao, X. Driving forces and their interactions of built-up land expansion based on the geographical detector–A case study of Beijing, China. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 2188–2207. [CrossRef]
16. Zhan, Y.; Luo, Y.; Deng, X.; Zhang, K.; Zhang, M.; Grieneisen, M.L.; Di, B. Satellite-based estimates of daily NO2 exposure in China using hybrid random forest and spatiotemporal kriging model. *Environ. Sci. Technol.* **2018**, *52*, 4180–4189. [CrossRef]
17. Hu, L.; He, S.; Han, Z.; Xiao, H.; Su, S.; Weng, M.; Cai, Z. 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* **2019**, *82*, 657–673. [CrossRef]
18. Li, G.; Cai, Z.; Liu, X.; Liu, J.; Su, S. A comparison of machine learning approaches for identifying high-poverty counties: Robust features of DMSP/OLS night-time light imagery. *Int. J. Remote Sens.* **2019**, *40*, 5716–5736. [CrossRef]
19. Wheeler, D.; Tiefelsdorf, M. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *J. Geogr. Syst.* **2005**, *7*, 161–187. [CrossRef]
20. Kang, Y.; Zhang, F.; Peng, W.; Gao, S.; Rao, J.; Duarte, F.; Ratti, C. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy* **2020**, 104919. [CrossRef]
21. Long, Y.; Liu, L. How green are the streets? An analysis for central areas of Chinese cities using Tencent Street View. *PLoS ONE* **2017**, *12*, e0171110. [CrossRef]
22. Liang, X.; Liu, Y.; Qiu, T.; Jing, Y.; Fang, F. The effects of locational factors on the housing prices of residential communities: The case of Ningbo, China. *Habitat Int.* **2018**, *81*, 1–11. [CrossRef]
23. Yuan, F.; Wei, Y.D.; Wu, J. Amenity effects of urban facilities on housing prices in China: Accessibility, scarcity, and urban spaces. *Cities* **2020**, *96*, 102433. [CrossRef]
24. Cui, N.; Gu, H.; Shen, T.; Feng, C. The impact of micro-level influencing factors on home value: A housing price-rent comparison. *Sustainability* **2018**, *10*, 4343. [CrossRef]
25. Yuan, F.; Wu, J.; Wei, Y.D.; Wang, L. Policy change, amenity, and spatiotemporal dynamics of housing prices in Nanjing, China. *Land Use Policy* **2018**, *75*, 225–236. [CrossRef]
26. Chin, H.C.; Foong, K.W. Influence of school accessibility on housing values. *J. Urban Plan. Dev.* **2006**, *132*, 120–129. [CrossRef]
27. Jim, C.Y.; Chen, W.Y. Impacts of urban environmental elements on residential housing prices in Guangzhou (China). *Landsc. Urban Plan.* **2006**, *78*, 422–434. [CrossRef]
28. Wen, H.; Tao, Y. Polycentric urban structure and housing price in the transitional China: Evidence from Hangzhou. *Habitat Int.* **2015**, *46*, 138–146. [CrossRef]
29. Debrezion, G.; Pels, E.; Rietveld, P. The impact of railway stations on residential and commercial property value: A meta-analysis. *J. Real Estate Financ. Econ.* **2007**, *35*, 161–180. [CrossRef]
30. Zhang, Y.; Dong, R. Impacts of street-visible greenery on housing prices: Evidence from a hedonic price model and a massive street view image dataset in Beijing. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 104. [CrossRef]
31. Su, S.; Zhou, H.; Xu, M.; Ru, H.; Wang, W.; Weng, M. Auditing street walkability and associated social inequalities for planning implications. *J. Transp. Geogr.* **2019**, *74*, 62–76. [CrossRef]
32. Yin, L.; Wang, Z. Measuring visual enclosure for street walkability: Using machine learning algorithms and Google Street View imagery. *Appl. Geogr.* **2016**, *76*, 147–153. [CrossRef]
33. Helbich, M.; Yao, Y.; Liu, Y.; Zhang, J.; Liu, P.; Wang, R. Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environ. Int.* **2019**, *126*, 107–117. [CrossRef]
34. Zhang, F.; Zhou, B.; Liu, L.; Liu, Y.; Fung, H.H.; Lin, H.; Ratti, C. Measuring human perceptions of a large-scale urban region using machine learning. *Landsc. Urban Plan.* **2018**, *180*, 148–160. [CrossRef]
35. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining; ACM: New York, NY, USA, 2016; pp. 785–794.
36. Fu, X.; Jia, T.; Zhang, X.; Li, S.; Zhang, Y. Do street-level scene perceptions affect housing prices in Chinese megacities? An analysis using open access datasets and deep learning. *PLoS ONE* **2019**, *14*, e0217505. [CrossRef]

37. Rafiei, M.H.; Adeli, H. A novel machine learning model for estimation of sale prices of real estate units. *J. Constr. Eng. Manag.* **2016**, *142*, 04015066. [CrossRef]
38. Ma, J.; Ding, Y.; Cheng, J.C.; Tan, Y.; Gan, V.J.; Zhang, J. Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: A city management perspective. *IEEE Access* **2019**, *7*, 148059–148072. [CrossRef]
39. Li, R.; Cui, L.; Fu, H.; Meng, Y.; Li, J.; Guo, J. Estimating high-resolution PM1 concentration from Himawari-8 combining extreme gradient boosting-geographically and temporally weighted regression (XGBoost-GTWR). *Atmos. Environ.* **2020**, *229*, 117434. [CrossRef]
40. Jin, Q.; Fan, X.; Liu, J.; Xue, Z.; Jian, H. Estimating Tropical Cyclone Intensity in the South China Sea Using the XGBoost Model and FengYun Satellite Images. *Atmosphere* **2020**, *11*, 423. [CrossRef]
41. Du, Q.; Wu, C.; Ye, X.; Ren, F.; Lin, Y. Evaluating the effects of landscape on housing prices in urban China. *Tijdschr. Voor Econ. En Soc. Geogr.* **2018**, *109*, 525–541. [CrossRef]
42. Wu, C.; Peng, N.; Ma, X.; Li, S.; Rao, J. Assessing multiscale visual appearance characteristics of neighbourhoods using geographically weighted principal component analysis in Shenzhen, China. *Comput. Environ. Urban Syst.* **2020**, *84*, 101547. [CrossRef]
43. Yang, W.; Zhen, X.; Gao, W.; Ouyang, S. An examination of the impact of neighbourhood walking environments on the likelihood of residents of dense urban areas becoming overweight or obese. *Can. Geogr./ Géogr. Can.* **2020**, *64*, 619–633. [CrossRef]
44. Li, X.; Zhu, J.; Wang, Y. Spatial differences of residential quarter floor area ratio: A case study of Dalian. *Prog. Geogr.* **2015**, *34*, 687–695.
45. Wen, H.; Xiao, Y.; Zhang, L. School district, education quality, and housing price: Evidence from a natural experiment in Hangzhou, China. *Cities* **2017**, *66*, 72–80. [CrossRef]
46. Bennett, M.M.; Smith, L.C. Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sens. Environ.* **2017**, *192*, 176–197. [CrossRef]
47. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
48. Kong, F.; Yin, H.; Nakagoshi, N. Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. *Landsc. Urban Plan.* **2007**, *79*, 240–252. [CrossRef]
49. Ren, X.; Guo, H.; Li, S.; Wang, S.; Li, J. A novel image classification method with CNN-XGBoost model. In *International Workshop on Digital Watermarking*; Springer: Cham, Switzerland, 2017; pp. 378–390.
50. Gumus, M.; Kiran, M.S. Crude oil price forecasting using XGBoost. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 1100–1103.
51. Rosen, S. Hedonic prices and implicit markets: Product differentiation in pure competition. *J. Political Econ.* **1974**, *82*, 34–55. [CrossRef]
52. Wang, R.; Lu, Y.; Wu, X.; Liu, Y.; Yao, Y. Relationship between eye-level greenness and cycling frequency around metro stations in Shenzhen, China: A big data approach. *Sustain. Cities Soc.* **2020**, *59*, 102201. [CrossRef]
53. Wang, R.; Liu, Y.; Lu, Y.; Zhang, J.; Liu, P.; Yao, Y.; Grekousis, G. Perceptions of built environment and health outcomes for older Chinese in Beijing: A big data approach with street view images and deep learning technique. *Comput. Environ. Urban Syst.* **2019**, *78*, 101386. [CrossRef]
54. Wei, Y.D. Zone fever, project fever: Development policy, economic transition, and urban expansion in China. *Geogr. Rev.* **2015**, *105*, 156–177. [CrossRef]