sustainability

MDPI

Article

Research on the Sustainable Development of Urban Housing Price Based on Transport Accessibility: A Case Study of Xi'an, China

Chao Xue, Yongfeng Ju *, Shuguang Li * and Qilong Zhou

School of Electronic and Control Engineering, Chang'an University, Xi'an 710064, China; 2015032001@chd.edu.cn (C.X.); 2017132046@chd.edu.cn (Q.Z.)

* Correspondence: yfju@chd.edu.cn (Y.J.); shgli@chd.edu.cn (S.L.); Tel.: +86-29-8233-4551 (S.L.)

Received: 10 January 2020; Accepted: 12 February 2020; Published: 17 February 2020

Abstract: The development of a real estate economy is beneficial to urban stability. A method of real estate price prediction based on transport accessibility is proposed. The method adds bus accessibility and metro accessibility into the model, which has higher prediction accuracy than the traditional model. Firstly, bus accessibility and metro accessibility are calculated according to the space syntax theory. Then, four models, the traditional hedonic price model (HPM) with transport accessibility, the traditional hedonic price model without transport accessibility, the random forest (RF) model with transport accessibility, and the random forest model without transport accessibility, are introduced. Finally, the four models are compared and analyzed in terms of precision and importance of index contributions. Taking Xi 'an, China, as an example, the experimental results show that the transport accessibility calculated based on space syntax can accurately represent the transport convenience in an urban space structure. Furthermore, it has a great influence on the contribution of indexes in the model. With the introduction of bus accessibility and metro accessibility, the accuracy of the real estate price prediction model is greatly improved.

Keywords: sustainable development; bus accessibility; metro accessibility; real estate price estimation; model comparison

1. Introduction

The real estate economy to some extent represents the economic development level of a city. In the process of sustainable urban development, urban spatial planning and real estate price planning have gradually become the focus of governments, consumers, investors and academic researchers. The causes of real estate prices are generally based on local conditions, which are complicated and changeable. Choosing the influencing factors and different models of real estate price is an effective way to estimate the real estate price.

The economic significance, collectability and data quantification of the feature need to be focused when considering the feature selection of house price prediction [1]. At present, there are two main factors in feature selection: the internal property factor and the location factor of the house. The internal property factor mainly includes the housing area, floor, decoration level, elevator and so on. The location factor mainly includes the community greening rate, plot ratio, parking space, property fees and so on. When analyzing the real estate price, it is also necessary to consider the influence of transport factors from the perspective of urban spatial planning. The transport factor can explain the transport convenience of the house in the urban space. According to a random street survey, residents always want houses that are closer to a bus station, metro station or urban main road, and they prefer to buy houses near bus stations and metro stations. The convenience of transportation

is an important feature of real estate. Therefore, bus accessibility and metro accessibility are calculated to describe the convenience of transportation.

The quantification of the transport factor is generally simply measured by distance in the current research. Martínez et al. studied the distance distribution between the house location in Lisbon, Portugal, and the public transport facilities, such as metro and light rail, and analyzed the impact of public transport on land price promotion [2]. Tao et al. analyzed the change rule of commercial value of real estate along a metro line and the influence of the metro radiation radius on commercial value by taking Wuhan Metro Line 2, China, as an example [3]. Li et al. analyzed the impact of the location of the real estate and the distance from the metro station on the price by taking Beijing, China, as an example [4]. However, in urban spatial planning, due to the complexity of the transportation network, it is not practical to simply take the distance between two points as the research scale. Attribute weights at different points should also be taken into account.

In this paper, a calculation method describing the influence of bus and metro factors on house prices is proposed. This method can describe the real situation of urban transportation better than a single distance.

In the model selection, the traditional method is to use the hedonic price model (HPM) or its improved model to explore the influencing factors of the real estate price. In 1928, Waugh proposed the functional relationship between commodity characteristics and price when applying the regression equation to study the relationship between vegetable quality and price change in Boston [5]. In 1974, Rosen introduced HPM into the real estate industry and made it develop vigorously [6]. Schollenberg et al. studied the house price change in Sweden based on HPM [7]. Andrewson et al. found that there are spatial autocorrelations and spatial error variables in the sample data when analyzing the relationship between house price and spatial influencing factors. Therefore, they proposed an improved model based on the HPM, spatial auto-regression model (SAR) and spatial auto-regression model with error term (SAE) [8]. Shin et al. used the SAE model and introduced the spatial weight matrix to describe the spatial dependence between variables when analyzing the real estate price in Seoul, South Korea [9]. Axhausen et al. compared and analyzed the HPM, SAR and SAE models, and improved the traditional regression model to better estimate the interaction between house price and characteristic factors [10]. Tao et al. built multi HPM, SAR and SAE, and analyzed the change rule of commercial value of real estate along a metro line and the influence of metro radiation radius on commercial value by taking Wuhan Metro Line 2 as an example [3]. Huang et al. analyzed the value-added changes of real estate in Zhengzhou based on the multiple linear regression (MLR) semi-logarithm model by taking Zhengzhou as an example [11]. Zhang et al. used HPM, SAR and SAE multiple models to make a comparative analysis of house prices, and explored the spatial effect of urban rail transit on house prices by taking Hangzhou Metro Line 1 as an example [12]. In the above research, although the common HPM model has high interpretability, with the complexity of urban spatial structure and the diversification of urban characteristics, it cannot be described clearly. Although the SAR and SAE models can explain the influence of spatial factors on house prices, the spatial weight matrix was assumed according to personal experience, so the experimental results are not objective.

In recent years, the rise of machine learning provides a new way for real estate price research. The genetic algorithm (GA), the support vector machine (SVM) algorithm, the random forest (RF) algorithm and the gradient boosting decision tree (GBDT) algorithm, which are commonly used in machine learning, have many advantages, such as fast iteration speed and high precision, and are able to characterize the nonlinear relationship between variables well.

Manganelli et al. used the GA to analyze the relationship between real estate price and geographical location in Potenza city, demonstrating the superiority of the genetic algorithm in this respect [13]. Giudice et al. used the multiple linear regression (MLR) algorithm and GA to analyze the real estate prices in Naples. The results showed that the GA was better than the MLR algorithm [14]. Vineeth et al. analyzed house price and its influencing factors by applying the machine learning algorithm [15]. Phan et al. used the machine learning algorithm to predict the trend of house prices based on the

historical transaction prices in Melbourne, Australia [16]. Xianan used the random forest algorithm to evaluate the risk of real estate development projects in different regions [17].

In this paper, a real estate price estimation model based on the RF method is proposed, and on this basis, bus accessibility and metro accessibility are introduced as transport factors of the model. Through the comparison with the traditional HPM method, the optimal model of real estate price estimation is determined. In the following sections, data sources, characterization, and modeling methods are described in detail.

2. Materials and Methods

2.1. Data Source

Data were from the house property data and urban transportation data in the main urban area of Xi'an, China. The spatial projection distribution of data in the ArcGIS software is shown in Figure 1. The main collection method for the house attribute data was to capture the information of ANJUKE (a house information publishing website in China), and the collection time was June 2019. The urban basic road network, urban bus and urban metro transportation data were obtained mainly through the API interface of the GAODE app, and the data collection time was May 2019.

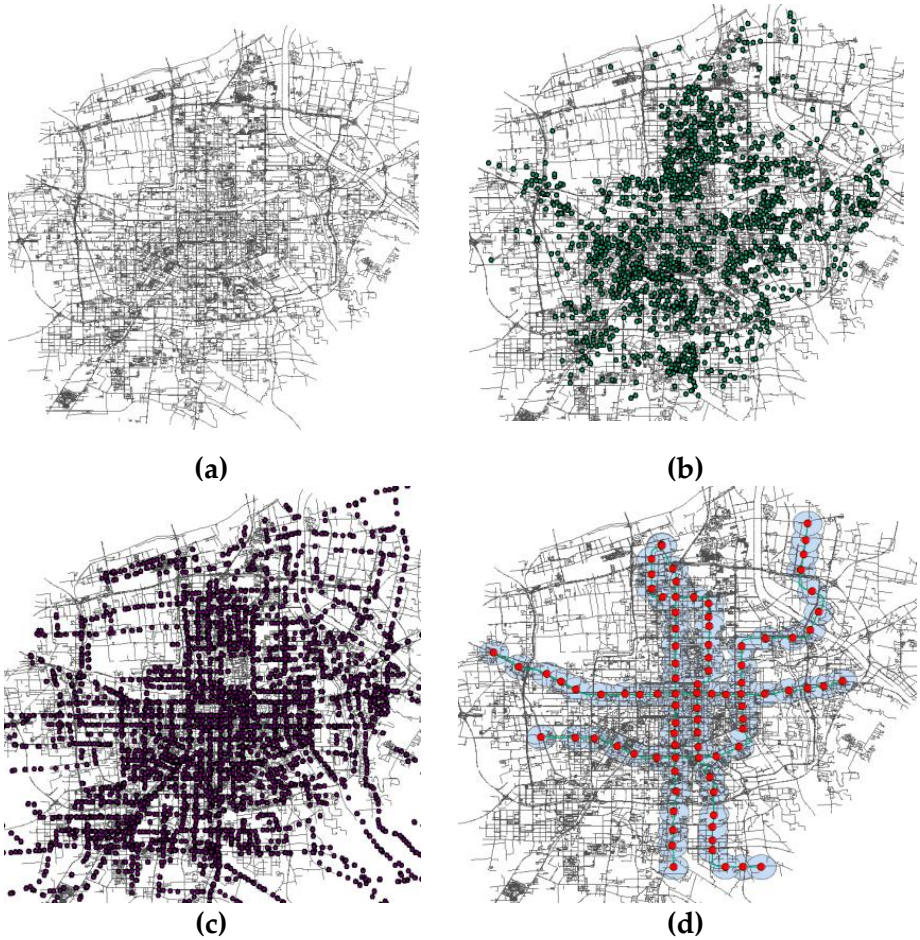

Figure 1. Data spatial distribution in Xi'an, China. (**a**) Urban road network spatial distribution. (**b**) Housing spatial distribution. (**c**) Bus spatial distribution. (**d**) Metro spatial distribution.

Figure 1a shows the road network data map of Xi'an City, including urban main roads, auxiliary roads and branch roads. Figure 1b shows the property data of houses in the urban area with the road network as the base map. Figure 1c,d respectively projects and marks the bus and metro

data of Xi'an city. The acquisition of these data provides data support for the later construction of characteristic indicators.

2.2. Analysis Framework

Figure 2 shows the analysis framework on real estate price estimation in this study. First, data need to be collected, mainly including real estate property data, urban geospatial data, bus and metro data. Second, the indexes of real estate price estimation are constructed based on the obtained data. In our study, two important indexes, bus accessibility and metro accessibility, are introduced in detail. Next, four models, the traditional hedonic price model (HPM) with transport accessibility, the traditional hedonic price model without transport accessibility, the random forest (RF) model with transport accessibility, and the random forest model without transport accessibility, are constructed. Finally, the optimal real estate price estimation model is determined through model comparison. Furthermore, the sustainable development and application of the model in the city is also discussed and analyzed in our study.

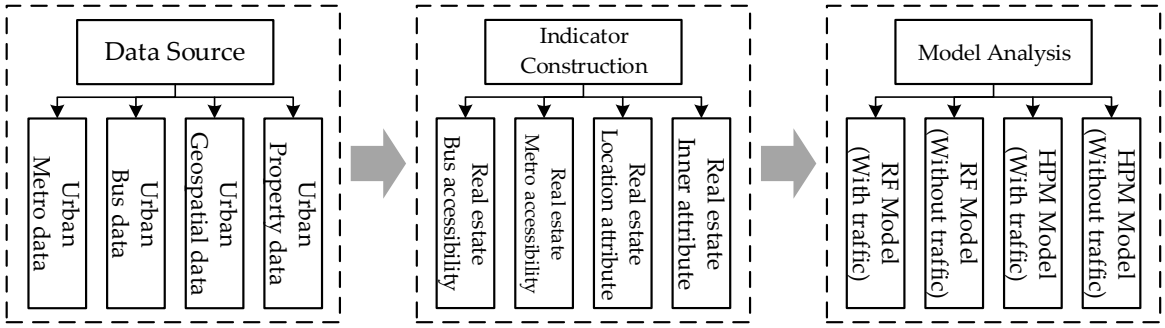

Figure 2. Analysis framework.

2.3. Data Processing

2.3.1. Bus Accessibility

Bus accessibility reflects the convenience of residents when they travel through the city. The calculation method of bus accessibility is based on the analysis method of the OD (origin to destination) cost matrix of minimum impedance [18]. From a bus station to another one, the smaller the impedance between two points, the better the accessibility of the point. The minimum impedance here refers to the shortest path distance between two bus stops. The calculation method is shown in Formula (1):

$$B_i = \frac{1}{n-1} \sum_{j=1}^{n} \frac{d_{ij}}{k_i} \quad (1)$$

where B_i represents the accessibility of stop i, k_i is the number of bus lines passing through stop i, as the station weight. The more bus lines passing through stop i, the better the accessibility, the less the impedance, and the less the weight of the station. n is the number of all bus stops, and d_{ij} is the shortest path distance from stop i to stop j.

When calculating the bus accessibility, first calculate the shortest path length from a bus stop to all bus stops through the ArcGIS software, where the path is obtained based on the real urban road network map. Secondly, calculate the accessibility of each bus stop through Formula (1). Then generate the thermodynamic diagrams of bus accessibility by using the kernel density estimate, as shown in Figure 3. Finally, the bus accessibility is associated with the house price attribute table through nearest neighboring analysis, regarding it as the characteristic index of the house's bus accessibility.

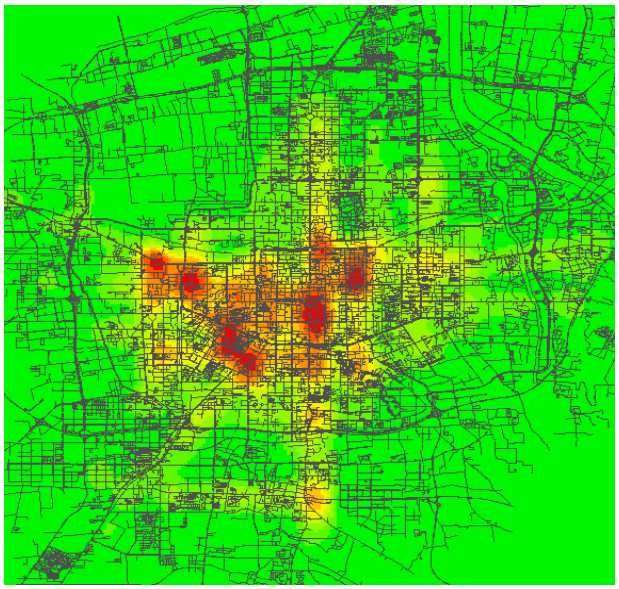

Figure 3. Thermal map of bus accessibility.

2.3.2. Metro Accessibility

The metro accessibility reflects the convenience of the residents when they travel through the metro in the city. The calculation method of metro accessibility is based on the space syntax theory [19]. Through describing the urban spatial form, space syntax analyzes how the spatial behavior is affected by the spatial form and its relationship. At present, space syntax is widely used in crime analysis, traffic pollution control, urban traffic analysis and prediction and so on. Accessibility calculation is also one of its important applications. In space syntax theory, the degree of integration reflects the degree of agglomeration or dispersion between a single node and all other nodes in urban space [20]. The smaller the integration value is, the lower the convenience of the node in urban space, and the lower the node is in an inconvenient position. In contrast, the larger the integration value is, the higher the convenience of the node in urban space, and the higher the node is in a convenient position. The calculation formula is shown in Formula (2):

$$\begin{cases} I_i = \dfrac{D_n(n-2)}{2(MD_i - 1)} \\ MD_i = \dfrac{\sum_{j=1}^{n} d_{ij}}{n-1} \\ D_n = \dfrac{2n\left[\log_2\left(\frac{n+2}{3} - 1\right) + 1\right]}{(n-1)(n-2)} \end{cases} \tag{2}$$

where I_i represents the integration degree in spatial syntax and the accessibility of metro i stations defined in this study; n represents the total number of metro stations; MD_i represents the average depth of station i; d_{ij} represents the distance between station i and station j; and D_n is the standardized parameter defined in spatial syntax.

The specific steps to calculate the metro accessibility are as follows. Firstly, the map of the metro lines of Xi'an are interrupted according to the metro station, and then it is converted into an axis map by Depthmap software, as shown in Figure 4. Secondly, Formula 2 is used to calculate the degree of integration at each metro station. Next, the map is inversely transformed and the core density analysis function in ArcGIS software is used to generate the metro accessibility trend surface, as shown in Figure 5. Finally, the metro accessibility and the property properties are analyzed in nearest neighbor to obtain the metro accessibility of each real estate location.

In Figure 4, the lines in different colors represent the lines connected by the calculated integration of each metro station. In order to distinguish the common station from the transfer station, the transfer number is taken as the weight of the station, and the metro accessibility is finally obtained.

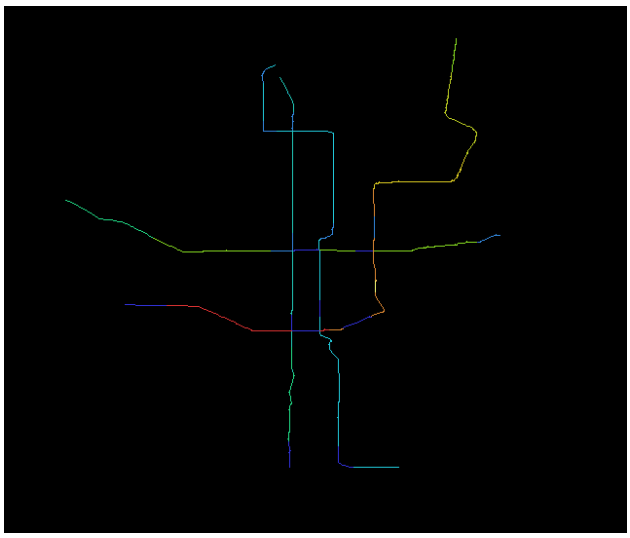

Figure 4. Axis map of metro lines.

In Figure 5, the darker the color, the higher the metro accessibility value. Through the function of neighbor analysis in ArcGIS software, the metro accessibility is related to each real estate in Xi'an. The spatial syntax is used to calculate the metro accessibility. The advantage is that metro lines are regarded as a network and its spatial characteristics are analyzed. Therefore, the calculation of the metro accessibility is more realistic.

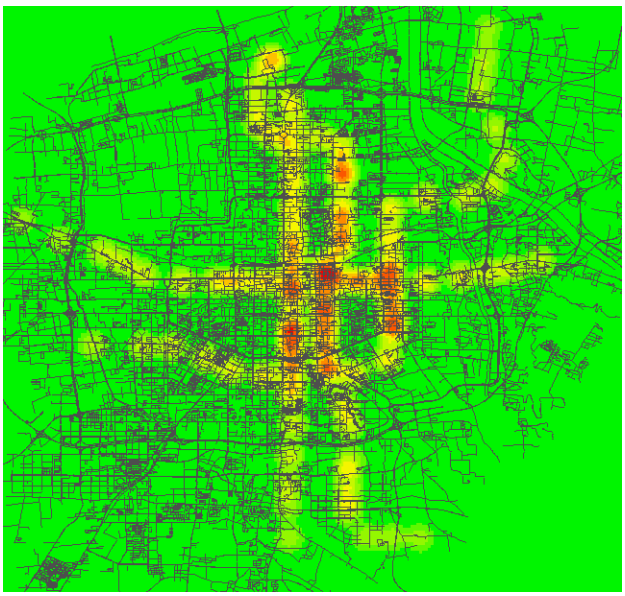

Figure 5. Thermal map of metro accessibility.

Transport accessibility refers to the degree of transport convenience when residents choose different modes of transportation to arrive at the destination from the starting point. Transport accessibility plays an important role in road network optimization, land use planning, land use evaluation and location analysis [21]. Compared with the value calculated by a single distance factor, the two transport

characteristics of bus accessibility based on the OD cost matrix of minimum impedance, and the metro accessibility based on the space syntax theory, are more complex. However, they can more reasonably represent the public transport operation in the urban space, make the data analysis results more authentic, and also better reflect the impact of urban public transport factors on the house price.

2.3.3. Indicator Construction

Before the establishment of the real estate price evaluation model, it is necessary to construct the indexes of the model. Table 1 gives a detailed explanation of the model indexes.

Table 1. The description of model indexes.

Categories	Indexes	Description	Mean	SD
Inner attribute	Price	The price of individual real estate, unit is yuan/m^2	14,270.68	5319.307
	Type	Types of real estate (Residence = 1, apartment = 2, villa = 3)	NA	NA
	Period	The years people can live after they buy the house (40 years, 50 years, 70 years)	NA	NA
	S	The size of building area, unit is m^2	107.72	52.28
	Direction	Building head(east = 1, west = 2, south = 3, north = 4)	NA	NA
	Floor	The floor of the real estate(lower = 1, middle = 2, upper = 3)	NA	NA
	Elevator	Elevator = 1, no elevator = 0	NA	NA
	Decoration	The degree of decoration of the real estate(simple decoration = 1, medium decoration = 2, high decoration = 3, luxury decoration = 4)	NA	NA
Location attribute	Plot	Plot ratio	3.52	1.37
	Car	The number of parking spaces in the real estate	1032	1380
	Green	Afforestation rate	0.37	0.07
	Fee	Property management fee, unit is yuan/m^2per month	1.29	1.05
	Year	Age of construction of the real estate	2011	4.2
Transport attribute	Bus	Bus accessibility	2.73	2.1
	Metro	Metro accessibility	1507.91	1353.65

SD: standard deviation; yuan: 1 yuan = US$ 0.1413 in 2019; NA: categorical variables not applicable.

According to Table 1, there are 15 kinds of indexes divided into three categories, and each index has been digitized. It should be noted that the numerical classification of features is done by manually marking. For example, during the decoration data collection, the interior pictures of each real estate are collected. Through manual identification, the decoration degree of each house is digitally labeled. This ensures the authenticity of the feature. Among these indexes, price is the output of the model, and the other 14 indexes are the input of the model. In order to ensure the authenticity and validity of the data, we cleaned the collected data and finally obtained 29,180 pieces of data to evaluate the real estate price model.

2.4. Analysis Method

This section focuses on two methods, the traditional HPM method and the RF method, for real estate price model construction.

2.4.1. Traditional HPM Method

The traditional HPM method uses price as the dependent variable and other indicators as independent variables to construct multiple linear regression (MLR) function. Formula 3 shows the general expression of the traditional HPM method:

$$p = f(I, L, T) \tag{3}$$

where p represents the price of real estate, I represents the inner indexes of real estate, L represents the location indexes of real estate, and T represents the transport indexes of real estate.

When using the traditional HPM method, the real estate price estimation models with transport indexes and without transport indexes are constructed simultaneously. Therefore, the influence of transport factors on the real estate price can be better judged.

2.4.2. Random Forest Method

The RF method is an algorithm based on the decision tree proposed by Breiman and Cutler, which uses multiple decision trees to train and predict samples [22]. It combines multiple weak classifiers (decision trees) to learn the data, and obtains the final results by voting or taking the mean value of all the final results of decision tree learning, so that the results of the overall model have a higher accuracy and maintain a certain generalization performance. The RF algorithm has the advantages of being applicable to high-dimensional large sample data, strong anti-interference ability and high model accuracy [23].

The process of building the real estate price model based on the RF method is firstly to adjust the model parameters. The grid-search method is also used to determine parameters. Secondly, cross validation is used to prevent model overfitting [24]. The K-fold cross validation method is used in the study [25]. Finally, the training data are used to build the model, and the testing data are used to predict the real estate price. Similar to the traditional HPM method, the real estate price model based on RF method can also be divided into two types: the transport index model and the non-transport index model.

2.4.3. Data division and Model Evaluation Method

Usually, before model construction, it is necessary to determine the division of the data set and the model evaluation method.

The common method for data division is to divide the data into a training set and a test set in the process of modeling. The test set is independent of the training set data, which are not involved in the training at all. In this study, the ratio of the training set to the testing set is 7:3.

R^2 and *RMSE* are commonly used model evaluation indexes. The calculation of the two indexes are shown in Formula (4) and Formula (5).

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y_i})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \overline{y_i})^2} \tag{5}$$

where m represents the number of samples; y_i represents the real value of the sample i; $\hat{y}_i$ represents the predicted value of the sample i; and $\overline{y_i}$ represents the mean value of the samples. R^2 is used to measure the accuracy of the model. The closer the value is to 0, the more inaccurate the representation is, and the closer the value is to 1, the more accurate the representation is. *RMSE* is the root mean square error, and the smaller the values are, the better the model is.

3. Experimental Results

In this section we construct four real estate price estimation models. Moreover, the optimal real estate price estimation model is determined by the model prediction accuracy, model error and model real-time performance.

3.1. Real Estate Price Estimation for Traditional HPM

The traditional HPM model is used to construct the real estate price estimation model. First, build the traditional HPM method without transport indexes. The traditional HPM function is shown in Formula (6):

$$p = c_0 + \sum c_i x_i + \sum c_j x_j + \sum c_k x_k \tag{6}$$

where p represents the price of real estate, c_0 is the constant term of the model, and x_i represents the inner property index of real estate, including the type, period, s, direction, floor, elevator and decoration. c_i is the coefficient on x_i. x_j represents the location index of real estate, including the plot, car, green, fee and year. c_j is the coefficient on x_j.

Next, the data are randomly divided into a training set and a testing set according to the ratio of 7:3, and then 10% of the data are randomly sampled in the training set to get the verification set. Finally, the training set is used to fit the function of the traditional HPM, and the real estate price evaluation function based on the traditional HPM method without transport indexes is shown in Formula (7):

$$\begin{aligned} p = \quad & 0.309x_1 - 0.341x_2 + 3.392x_3 + 0.739x_4 - 0.155x_5 + 1.746x_6 + 2.105x_7 - \\ & 0.277x_8 - 0.144x_9 + 0.811x_{10} + 0.287x_{11} - 0.119x_{12} + 1.224e - 14 \end{aligned} \tag{7}$$

where x_1 is type, x_2 is period, x_3 is S, x_4 is direction, x_5 is floor, x_6 is elevator, x_7 is decoration, x_8 is plot, x_9 is car, x_{10} is green, x_{11} is fee, and x_{12} is year. The prediction of the model under the verification set is $R^2 = 0.258$ and the $RMSE = 4769$. The prediction of the model under the testing set is $R^2 = 0.219$ and the $RMSE = 4801$. According to the results, the prediction accuracy of the model is very low and the error is very large, so the model cannot accurately predict the real estate price. This model is labeled as M1.

The same experimental process is used to construct the real estate price estimation model with transport indexes based on the traditional HPM method. The function is shown in Formula (8):

$$\begin{aligned} p = \quad & 0.306x_1 - 0.347x_2 + 3.39x_3 + 0.76x_4 - 0.15x_5 + 1.73x_6 + 2.09x_7 - 0.359x_8 - \\ & 0.09x_9 + 0.84x_{10} + 0.27x_{11} - 0.002x_{12} + 0.37x_{13} - 0.007x_{14} + 1.15e - 14 \end{aligned} \tag{8}$$

where x_{13} is bus, and x_{14} is metro. The prediction of the model under the verification set is $R^2 = 0.326$ and the $RMSE = 4369$. The prediction of the model under the testing set is $R^2 = 0.227$ and the $RMSE = 4735$. According to the results, with the addition of bus accessibility and metro accessibility, the prediction accuracy of the model is improved compared with M1, and the error is also reduced, but it still fails to meet the requirements and cannot accurately predict the real estate price. This model is labeled as M2.

3.2. Real Estate Price Estimation for RF

Before starting, we mark the real estate price estimation model without transport indexes based on the RF method as M3 and mark the real estate price estimation model with transport indexes based on the RF method as M4 for the convenience of expression. When building M3 and M4, there are two model parameters that need to be adjusted, which are max_features and n_estimators. max_features represents the maximum number of features that nodes of the decision tree participate in judgment during classification. n_estimators represents the number of decision trees when constructing a random forest. Figures 6 and 7 respectively show the parameter adjustment curve about M3 and M4.

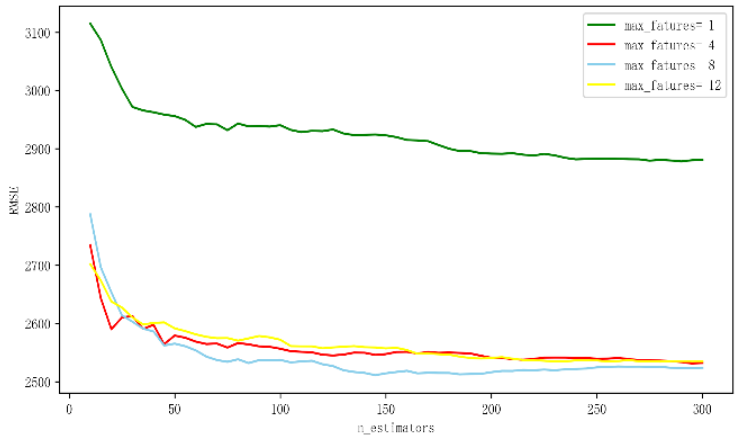

Figure 6. The parameters adjustment curve of M3.

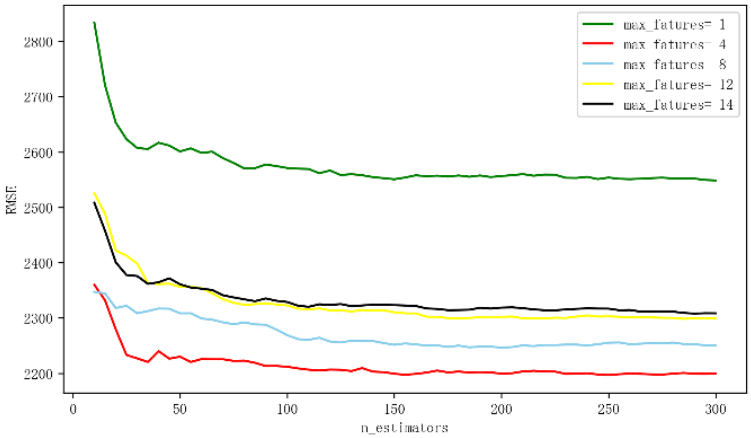

Figure 7. The parameters adjustment curve of M4.

In Figure 7 above, max_features = 8 and n_estimators = 150 in M3 are determined, and max_features = 4 and n_estimators = 150 in M4 are determined.

Next, the k-fold cross-validation (K = 10) is performed. The training set is randomly divided into K parts (K = 10). Each time the model is trained, 9 of them are taken as the training data, and the remaining 1 is taken as the validation data. The model is trained in turn, and the R^2 and RMSE of the validation set are obtained after 10 trainings. The results for M3 and M4 are shown in Table 2 below.

Table 2. The results of the K-Fold cross validation (K = 10).

K	Validation Set in M3		Validation Set in M4	
	R^2	RMSE	R^2	RMSE
1	0.792	2347.898	0.832	2126.106
2	0.755	2526.538	0.794	2341.029
3	0.728	2777.711	0.801	2379.009
4	0.808	2320.038	0.856	1970.579
5	0.767	2526.512	0.829	2193.542
6	0.766	2500.781	0.816	2173.348
7	0.774	2516.819	0.834	2158.122
8	0.770	2606.325	0.823	2285.631
9	0.794	2373.399	0.838	2122.621
10	0.771	2675.686	0.819	2368.503
mean	0.7729	2520.934	0.8243	2215.302

In Table 2, both M3 and M4 have stable accuracy and error under the verification set, indicating that the model is not overfitted. Furthermore, when M3 and M4 are used to predict the testing set, R^2 = 0.797 and *RMSE* = 2415 in M3, and R^2 = 0.840 and *RMSE* = 2142 in M4. The results show that M4 has better prediction accuracy and smaller error than M3.

3.3. Model Comparison

In Section 3.2, four real estate price estimation models are constructed. This section compares and analyze the prediction accuracy, error and running time of the models to determine the optimal model.

Figure 8 shows the results predicted by the four models on the testing set. The x-coordinate is the true value of real estate prices. The y-coordinate is the predictive value of real estate prices.

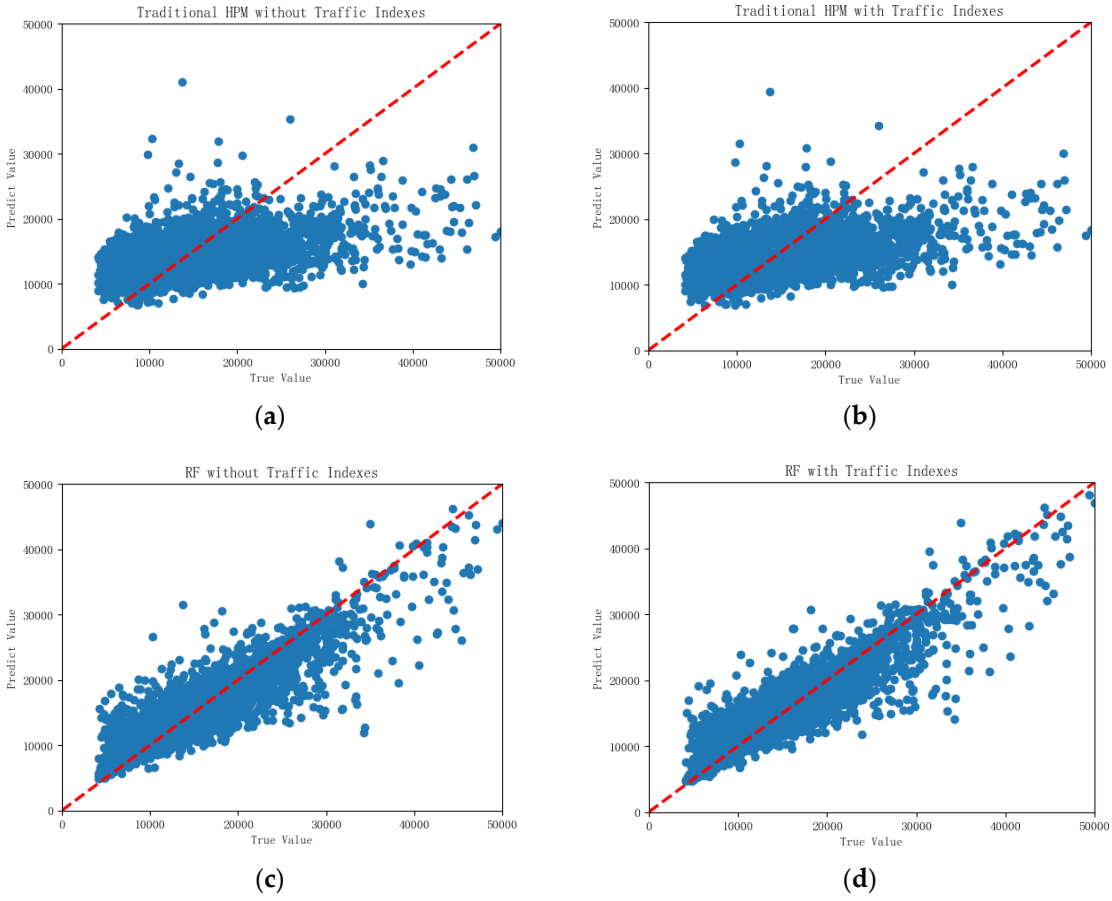

Figure 8. The accuracy of real estate price prediction. (**a**) Prediction accuracy of M1. (**b**) Prediction accuracy of M2. (**c**) Prediction accuracy of M3. (**d**) Prediction accuracy of M4.

According to Figure 8, compared with the RF method, the traditional HPM method has a lower accuracy in real estate price estimation. In addition, transport indexes (bus accessibility and metro accessibility) play an important role in improving the model accuracy. Table 3 shows the comparison of prediction accuracy, prediction error and running time of the four models in the testing set.

Table 3. Comparison of the models in the testing set.

Model	R^2	RMSE	Runtime(s)
M1	0.219	4801	0.006
M2	0.227	4735	0.006
M3	0.797	2415	0.417
M4	0.840	2142	0.421

In Table 3, Runtime refers to the total time taken by the models to predict the testing set data. Through a comparison, we can see that among the four models, M4 has the best estimation effect of real estate price. In terms of real-time performance, since the RF method is more complex than the HPM method, the model size is larger and the running time is longer. However, the running time of the four models is less than one second, which can guarantee real-time performance. Therefore, in this paper, the real estate price estimation model based on the RF method with transport indexes is selected as the final model.

4. Discussion

4.1. Application of the RF Method in Real Estate Price Prediction

At present, most researchers in China only think that the RF algorithm has advantages in classification prediction, but seldom apply it to regression. This study shows that the RF algorithm has a good effect in regression prediction, especially in real estate price prediction. In addition, with the wide application of big data, the RF algorithm can describe complex nonlinear relationships in the process of modeling different types of variables. Therefore, in the era of big data, the application of the RF algorithm to predict the real estate price is an effective method.

4.2. Application of Transport Accessibility in Real Estate Price Prediction

Buses and metros are the main ways of travel for city residents. Figure 9 shows the contribution of 14 features in the real estate model, which represent the importance of each feature in the model construction.

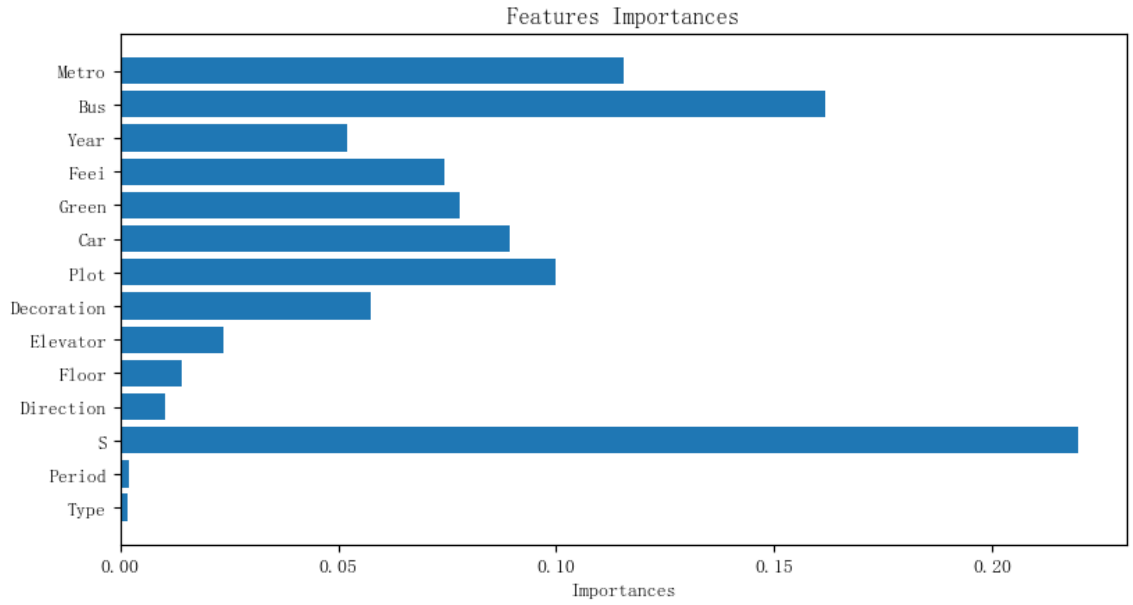

Figure 9. The feature importance of M4.

In Figure 9, the x-coordinate represents the percentage of importance and the y-coordinate represents 14 characteristic indexes. From the distribution results, the first three most important characteristics of house price forecast are housing area, bus accessibility and metro accessibility. This shows that when residents consider the purchase of houses, in addition to considering the internal attributes of the house, the transport conditions where the houses are located are also very important. Based on the Xi'an road network calculation, the two characteristic indexes of bus accessibility and metro accessibility have a high influence on house price forecast, which can be used as an important index of real estate price evaluation.

Nowadays, more and more people advocate green travel, and urban public transportation infrastructure is becoming perfected. In urban spaces, newly planned bus and metro lines can drive the growth of the regional real estate economy. The indexes of bus accessibility and metro accessibility proposed in this study can accurately estimate the real estate price through the real estate price model, which provides data support for urban spatial planning and is conducive to promoting the balanced and stable development of an urban regional economy.

5. Conclusions

In this study, two characteristic indexes of bus accessibility and metro accessibility were calculated and introduced into the influencing factors of real estate price. By comparing the traditional HPM method and the RF method, the real estate price estimation model based on the RF method with transport indexes was finally determined. The prediction accuracy of the model reached 0.84, and the root-mean-square error was 2142. The research shows that transport factors to some extent affect the real estate prices, and are an indispensable factor.

However, the study also has certain limitations. Firstly, some external amenities and spatial variables, such as hospitals, shopping malls, schools, etc., were not taken into account. The omission of these variables may lead to inaccurate results. Secondly, our study only compared the traditional HPM and RF model. More models should be selected for analysis. Finally, the study only focused on the real estate price in Xi 'an, China, and more cities with different spatial characteristics should be studied to reach a more universal conclusion. Further research will be carried out from three perspectives. The first is to describe the bus accessibility and metro accessibility more carefully. In this paper, the calculation of the bus accessibility index does not fully reflect the real situation, because the actual operation route data of the bus were not obtained, so it can only be replaced by the shortest path. Second, the factors that affect the real estate price are far more than the 14 features mentioned in the paper and we hope to take more factors into account to further improve the accuracy of the model, such as noise levels, crime rates, security levels, green areas and so on. The last and the most important one is that we hope to explore more possibilities for applying the real estate price estimation model to the sustainability of urban development. The real estate economy can reflect the development level of a city or region to a certain extent. In the inner space of the city, the areas with high real estate prices represent the locations that may have better education, better medical care and more convenient transportation. Therefore, we hope to propose a more accurate real estate price estimation model and conduct a deeper study on the regional development imbalance in the process of urban sustainable development.

Data Availability: The data used to support the findings of this paper are available from the corresponding author upon request. All data are included within the manuscript.

Author Contributions: C.X. and Y.J. conceived and designed the experiments; C.X., S.L. and Q.Z. presented tools and carried out the data analysis; C.X. wrote the paper; Y.J. and S.L. guided and revised the paper; C.X. rewrote and improved the theoretical section; S.L. and Q.Z. collected the materials and did a lot of format editing work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the General Program of National Natural Science Foundation of China (NSFC) under Grant No.61603057. This research was also partially supported by the Shaanxi Provincial Natural Science Foundation of China under Grant No. 2016JM5052 and the Special Fund for Basic Scientific Research of Central Colleges by Chang'an University under Grant No. 300102328402.

Acknowledgments: The authors are grateful for the comments and reviews from the reviewers and editors.

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

References

1. Ahmad, S.R.; Bakar, A.A.; Yaakub, M.R. A review of feature selection techniques in sentiment analysis. *Intell. Data Anal.* **2019**, *23*, 159–189. [CrossRef]

2. Martínez, L.M.; Viegas, J.M. Effects of transportation accessibility on residential property values: Hedonic price model in the Lisbon, Portugal, metropolitan area. *Transp. Res. Rec.* **2009**, *2115*, 127–137. [CrossRef]
3. Xu, T.; Zhang, M.; Aditjandra, P.T. The impact of urban rail transit on commercial property value: New evidence from Wuhan, China. *Transp. Res. Part A Policy Pract.* **2016**, *91*, 223–235. [CrossRef]
4. Li, S.; Chen, L.; Zhao, P. The impact of metro services on housing prices: A case study from Beijing. *Transportation* **2019**, *46*, 1291–1317. [CrossRef]
5. Waugh, F.V. Quality Factors Influencing Vegetable Prices. *J. Farm Econ.* **1928**, *10*, 185. [CrossRef]
6. Rosen, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *J. Polit. Econ.* **1974**, *82*, 34–55. [CrossRef]
7. Schollenberg, L. Estimating the hedonic price for Fair Trade coffee in Sweden. *Br. Food J.* **2012**, *114*, 428–446. [CrossRef]
8. Anderson, S.T.; West, S.E. Open space, residential property values, and spatial context. *Reg. Sci. Urban Econ.* **2006**, *36*, 773–789. [CrossRef]
9. Shin, K.; Washington, S.; Choi, K. Effects of Transportation Accessibility on Residential Property Values: Application of Spatial Hedonic Price Model in Seoul, South Korea, Metropolitan Area. *Transp. Res. Rec.* **2007**, *1994*, 66–73. [CrossRef]
10. Felsenstein, D.; Axhausen, K.; Waddell, P. Land use-transportation modeling with UrbanSim: Experiences and progress. *J. Transp. Land Use* **2010**, *3*, 1–3. [CrossRef]
11. Weijie, H. The Impact of Urban Rail Transit on Surrounding Real Estate Price—Taking Zhengzhou Metro Line 1 as an Example. *J. Luoyang Norm. Univ.* **2017**. Available online: https://www.cnki.net/kcms/doi/10.16594/j.cnki.41-1302/g4.2017.02.019.html (accessed on 10 January 2020).
12. Shu, Z. *Study on the Spatial Effect of Urban Rail Transit on Real Estate Price—Taking Hangzhou Metro Line 1 as an Example*; Zhejiang University: Hangzhou, China, 2014.
13. Manganelli, B.; de Mare, G.; Nesticò, A. Using genetic algorithms in the housing market analysis. In *Lecture Notes in Computer Science*; (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics); In the Proceedings of the International Conference on Computational Science and Its Applications; Springer: Cham, Germany, 2015; pp. 36–45. [CrossRef]
14. Del Giudice, V.; De Paola, P.; Forte, F. Using genetic algorithms for real estate appraisals. *Buildings* **2017**, *7*, 31. [CrossRef]
15. Vineeth, N.; Ayyappa, M.; Bharathi, B. House Price Prediction Using Machine Learning Algorithms. In Proceedings of the International Conference on Soft Computing Systems, Kollam, India, 19–20 April 2018; Springer: Singapore, 2018; pp. 425–433. [CrossRef]
16. Phan, T.D. Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In Proceedings of the 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Parramatta, Australia, 3–7 December 2019; pp. 8–13. [CrossRef]
17. Xianan, L. *Research on Risk Evaluation of Real Estate Project Based on Random Forest Algorithm*; Tianjin University: Tianjin, China, 2017.
18. Volchenkov, D.; Blanchard, P. Scaling and universality in city space syntax: Between Zipf and Matthew. *Phys. A Stat. Mech. Appl.* **2008**, *387*, 2353–2364. [CrossRef]
19. Telega, A. Urban Street Network Analysis Using Space Syntax in GIS—Cracow Case Study. In Proceedings of the 2016 Baltic Geodetic Congress (BGC Geomatics), Gdansk, Poland, 2–4 June 2016; pp. 282–287. [CrossRef]
20. Hou, W. Correlation Analysis of Human Traffic and Public Space Integration. *Anhui Archit.* **2017**, *24*. [CrossRef]
21. Luo, J.; Jin, X.; Dang, A. Study of Spatial Accessibility Based on GIS and Value of Time. In Proceedings of the 2010 International Conference on Optoelectronics and Image Processing, Haiko, China, 11–12 November 2010; pp. 416–419. [CrossRef]
22. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
23. Kulkarni, V.Y.; Sinha, P.K. Pruning of random forest classifiers: A survey and future directions. In Proceedings of the 2012 International Conference on Data Science & Engineering (ICDSE), Cochin, India, 18–20 July 2012; pp. 64–68. [CrossRef]

24. Pourtaheri, Z.K.; Zahiri, S.H. Ensemble classifiers with improved overfitting. In Proceedings of the 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), Bam, Iran, 9–11 March 2016; pp. 93–97. [CrossRef]
25. Bengio, Y.; Grandvalet, Y. Bias in estimating the variance of K-fold cross-validation. *Stat. Model. Anal. Complex Data Probl.* **2005**, 75–95. [CrossRef]