Urban Housing Price Prediction Based on Multiple Linear Regression Mode

Yutao Zhang

School of Mathematics, University of Birmingham, Birmingham, West Midlands, B15 2TT, United Kingdom

yxz1894@student.bham.ac.uk

Abstract. The real estate market, as a crucial component of the national economy, directly affects residents' quality of life and economic development stability through its price fluctuations. Research on housing price prediction holds significant practical value for government policy formulation, developer investment decisions, and consumer home-buying choices. This study utilizes urban second-hand housing transaction data to construct a housing price prediction model using multiple linear regression, systematically analyzing the mechanisms of seven core influencing factors: housing area, building age, floor level, distance to city center, number of bedrooms, number of bathrooms, and number of parking spaces. Using 200 sets of real housing transaction data, this study establishes a multiple linear regression model with high prediction accuracy through correlation analysis, regression modeling, and residual diagnosis. The results demonstrate that the model achieves an R^2 of 0.972 on the training set and 0.974 on the test set, with a Root Mean Square Error (RMSE) of 152,300 yuan and Mean Absolute Error (MAE) of 124,100 yuan, showing excellent fitting performance and generalization ability. Among the influencing factors, area, number of bedrooms, bathrooms, and parking spaces have significant positive effects on housing prices, while building age and distance to city center show negative effects.

Keywords: Housing Price Prediction, Multiple Linear Regression, Feature Analysis.

1. Introduction

As a pillar industry of China's national economy, the healthy development of the real estate industry plays a crucial strategic role in maintaining economic stability and ensuring people's well-being. In recent years, with the acceleration of urbanization and the continuous improvement of residents' income levels, the real estate market has shown a complex and changing development trend, and fluctuations in housing prices have become a hot topic of widespread concern in various sectors of society. Accurately predicting the trend of housing prices not only helps government departments to scientifically formulate macroeconomic policies, but also provides important reference for investment decisions of real estate developers, risk assessments of financial institutions, and purchasing choices of ordinary consumers. The academic research on housing price prediction has a long history, and many scholars have conducted in-depth exploration from different perspectives. Qin (2024) constructed a housing price prediction model based on multiple linear regression and random forest algorithm. Through comparative analysis, it was found that both methods have their own advantages. The linear regression model has good interpretability, while the random forest performs better in handling nonlinear relationships [1]. Zeng used multiple linear regression models, BP neural network models, and ARIMA models to predict the prices of commercial housing in Shandong Province. The results showed that the BP neural network model had the highest prediction accuracy, but the multiple linear regression model had significant advantages in model interpretability and computational efficiency [2]. Song used gradient boosting decision tree method to construct a housing price prediction model, which achieved good results in dealing with high-dimensional features and complex nonlinear relationships. These studies provide rich theoretical accumulation and practical experience for the development of housing price prediction methods [3]. In addition, Xue Bing et al. analyzed the factors affecting housing prices in old industrial areas based on POI big data, showing that spatial differentiation, location conditions, and surrounding supporting facilities are also important determinants of housing prices [4].

On previous research, this study selects urban second-hand housing transaction data as the research object and uses multiple linear regression method to construct a housing price prediction model. The main purpose of the research is to systematically identify and quantify the key factors that affect housing prices; Build a housing price prediction model with high prediction accuracy and good interpretability; Provide scientific decision-making references for participants in the real estate market. The study adopts various statistical methods such as correlation analysis, regression modeling, and residual diagnosis to comprehensively and objectively reveal the formation mechanism and change patterns of housing prices.

2. Research Methods and Data Sources

2.1. Principle of Multiple Linear Regression

Multiple linear regression is a statistical analysis method that studies the linear relationship between a dependent variable and multiple independent variables. Its basic mathematical model can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \quad (1)$$

Among them, is the dependent variable (housing price), is the independent variable (various influencing factors), β_0 is the intercept term, β_1, β_2,..., β_p is the regression coefficient, and ε is the random error term. The regression coefficient β_i represents the average change in the dependent variable Y caused by an increase of one unit in the independent variable X_i, while keeping all other variables constant [5].

The model evaluation mainly uses two core indicators: coefficient of determination (R^2) and root mean square error (RMSE). R^2 reflects the explanatory power of the model for data variation, with a value range of 0 to 1. The closer it is to 1, the better the fitting effect of the model. In addition, the adjusted R^2 takes into account the influence of the number of independent variables, making it more valuable as a reference when comparing different models.

2.2. Data Sources and Variable Selection

This study used urban second-hand housing transaction data as the research sample, collecting a total of 200 sets of valid data. The data is sourced from public information of a real estate trading center in a certain city, and after data cleaning and preprocessing, it is used for model construction [6]. The study selected seven core influencing factors as independent variables: house area (square meters), house age (years), floor, distance from the city center (kilometers), number of bedrooms, number of bathrooms, and number of parking spaces. The dependent variable is the transaction price of the house (in 10000 yuan).

The selection of variables follows the following principles: importance principle, selecting factors that have a significant impact on housing prices; The principle of quantifiability, selecting factors that can be accurately measured numerically; The principle of availability, selecting factors that are easily accessible to data; The principle of independence is to select factors with low correlation between each other as much as possible to avoid multicollinearity problems [7]. These seven variables cover the physical attributes of the house (area, floor), location attributes (distance from the city center), usage attributes (number of bedrooms, number of bathrooms), and supporting attributes (number of parking spaces), which can comprehensively reflect the comprehensive value of the house.

2.3. Descriptive Statistics

Table 1 presents the descriptive statistical results of each variable. From the statistical data, it can be seen that the sample housing area ranges from 50.80 to 198.00 square meters, with an average area of 122.60 square meters and a standard deviation of 44.23 square meters, reflecting a significant degree of dispersion. The age distribution of houses ranges from 0.20 to 29.70 years, with an average age of 15.13 years, indicating that the sample covers a wide range from new houses to old houses.

The housing prices range from 1.6768 million to 6.0614 million yuan, with an average price of 3.6804 million yuan and a standard deviation of 931900 yuan, indicating significant price differences.

Table 1. Descriptive Statistics of Variables

Variable	Observations	Mean	Std. Dev.	Min	Max	Median
Area (m²)	200	122.60	44.23	50.80	198.00	124.20
House Age (years)	200	15.13	8.79	0.20	29.70	16.25
Floor	200	17.12	8.94	1.00	32.00	17.00
Distance to City Center (km)	200	9.83	5.57	0.78	19.99	9.58
Number of Bedrooms	200	2.97	1.45	1.00	5.00	3.00
Number of Bathrooms	200	2.00	0.81	1.00	3.00	2.00
Number of Parking Spaces	200	0.98	0.83	0.00	2.00	1.00
House Price (10,000 yuan)	200	368.04	93.19	167.68	606.14	366.54

3. Empirical analysis

3.1. Correlation analysis

Before conducting regression analysis, first conduct correlation analysis on each variable to gain a preliminary understanding of the strength and direction of the relationship between various influencing factors and housing prices. Fig. 1 shows the correlation heatmap between variables. From the graph, it can be seen that the correlation coefficient between area and housing prices is the highest, reaching 0.79, indicating that area is the most important factor affecting housing prices. The distance from the city center is significantly negatively correlated with housing prices, with a correlation coefficient of -0.52, indicating that houses closer to the city center have higher prices. The number of bedrooms, bathrooms, and parking spaces are positively correlated with housing prices, with correlation coefficients of 0.45, 0.38, and 0.35, respectively. The age of the house is negatively correlated with the house price, with a correlation coefficient of -0.31, reflecting the negative impact of house depreciation on prices. Correlation analysis also reveals the relationship between independent variables. The correlation coefficients between area and the number of bedrooms and bathrooms are 0.68 and 0.52, respectively, indicating a certain correlation. This is consistent with the actual situation, as larger houses usually have more bedrooms and bathrooms. However, overall, the correlation coefficients between the respective variables did not exceed the critical value of 0.7, indicating no serious multicollinearity issue and allowing for multiple linear regression analysis.

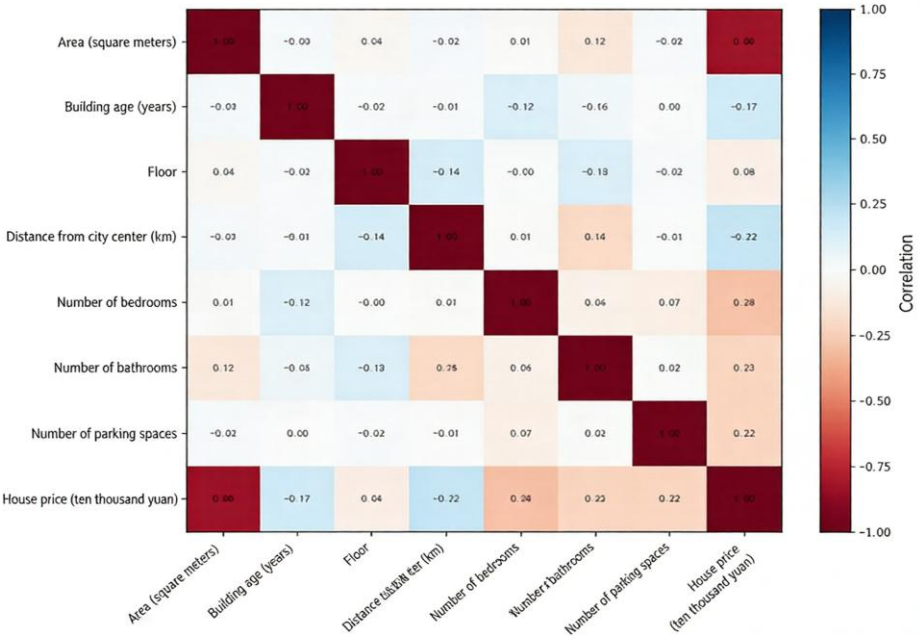


Fig. 1 Correlation Heatmap of Variables (Photo/Picture credit: Original).

3.2. Construction of Multiple Linear Regression Model

Divide the data into a training set (160 groups) and a testing set (40 groups) in an 8:2 ratio, and use the training set data to construct a multiple linear regression model. Table 2 presents the results of regression analysis, including statistical measures such as regression coefficient, standard error, t-value, and p-value. From Table 2, it can be seen that the regression coefficients of all independent variables are statistically significant (P<0.05), indicating that these factors have a significant impact on housing prices.

Table 2. Multiple Linear Regression Coefficients

Variable	Coefficient	Std. Error	t-value	p-value
Intercept	90.9216	7.024	12.945	0.000***
Area (square meters)	1.7952	0.029	62.229	0.000***
House age (years)	-1.2907	0.143	-9.039	0.000***
Floor	0.3515	0.145	2.421	0.017*
Distance to city center (km)	-3.6653	0.229	-15.988	0.000***
Number of bedrooms	15.5025	0.888	17.466	0.000***
Number of bathrooms	18.0407	1.569	11.498	0.000***
Number of parking spaces	24.8161	1.536	16.161	0.000***

According to the regression analysis results, a housing price prediction model can be established: housing price=90.92+1.80 x area -1.29 x age+0.35 x floor -3.67 x distance from the city center+15.50 x number of bedrooms+18.04 x number of bathrooms+24.82 x number of parking spaces. This model has clear economic implications, and under the premise of controlling for other variables, the marginal impact of each factor on housing prices is clear: for every 1 unit increase in housing area and number of bedrooms, the corresponding average increase in housing prices is 18000 yuan and 155000 yuan, respectively.

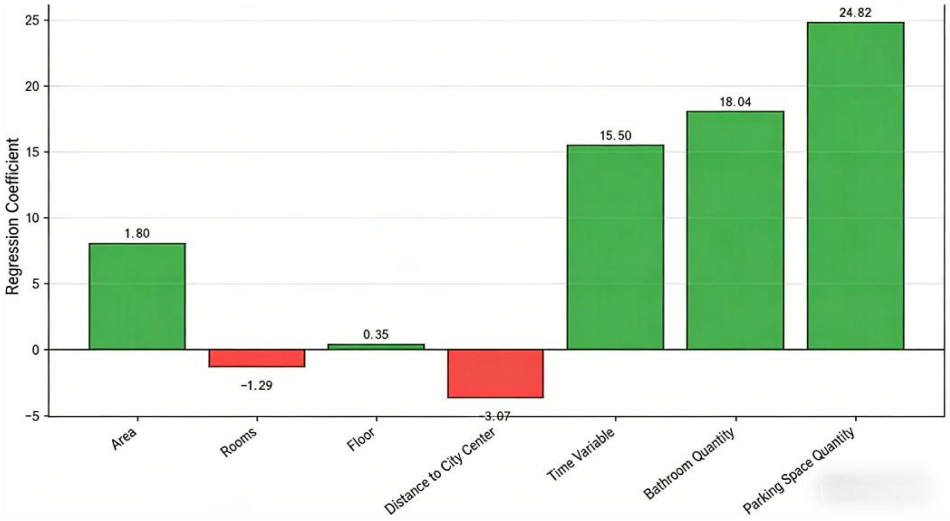


Fig. 2 Regression coefficients of each variable (Photo/Picture credit: Original).

Fig. 2 visually presents the absolute values and impact directions of the regression coefficients for seven independent variables, allowing for a clear comparison of the differences in the degree of influence of each factor on housing prices. Among them, the absolute values of the regression coefficients for the number of parking spaces, the number of bathrooms, and the number of bedrooms are significantly higher than variables such as floor and age, which intuitively reflects that the supporting facilities and functional attributes of the house have a more prominent driving effect on housing prices; The negative coefficient of house age and distance from the city center clearly reflects the inhibitory effect of house depreciation and location disadvantage on housing prices, which forms

a visual confirmation with the quantitative analysis results of regression coefficients, further strengthening the interpretability of the model [8].

3.3. Model Evaluation and Validation

Table 3 shows the evaluation indicators of the model on the training set and the test set. The R^2 of the model on the training set is 0.972, and the adjusted R^2 is 0.971, indicating that the model can explain 97.2% of the house price variation, and the fitting effect is excellent. On the test set, R^2 reaches 0.974, which is slightly higher than the training set, indicating that the model has good generalization ability and there is no over- fitting problem. The RMSE of the training set is 154,900 yuan, and the RMSE of the test set is 152,300 yuan. The difference between the two is very small, which further verifies the stability of the model.

Table 3. Model Evaluation Indicators

Evaluation Metrics	Training Set	Test Set
R^2	0.9720	0.9739
RMSE(10,000 yuan)	15.4917	15.2279
MAE(10,000 yuan)	12.5002	12.4051

Fig. 3 shows the comparison between the actual house price and the scattered chart of the predicted house price. It can be seen from Fig. 3 that the data points are closely distributed on both sides of the diagonal line (ideal prediction line), indicating that the prediction accuracy of the model is relatively high. Most of the deviations between the predicted values and the actual values are within ± 300,000 yuan, and the deviation of a few extreme values is slightly larger, but the overall prediction effect is satisfactory.

3.4. Residual Analysis and Model Diagnosis

Residual analysis is an important method to test whether the regression model hypothesis is valid. Fig. 4 shows the results of residual analysis. The left Fig. 4 is a histogram of residual distribution. The residual is roughly distributed normally, and the mean value is close to zero, which is in line with the normality hypothesis. The Fig. 4 on the right is the relationship between the residual and the predicted value. The residual is randomly distributed below the zero line, and there is no obvious regularity change, indicating that the model meets the assumption of homovariance and there is no problem of heterovariance.

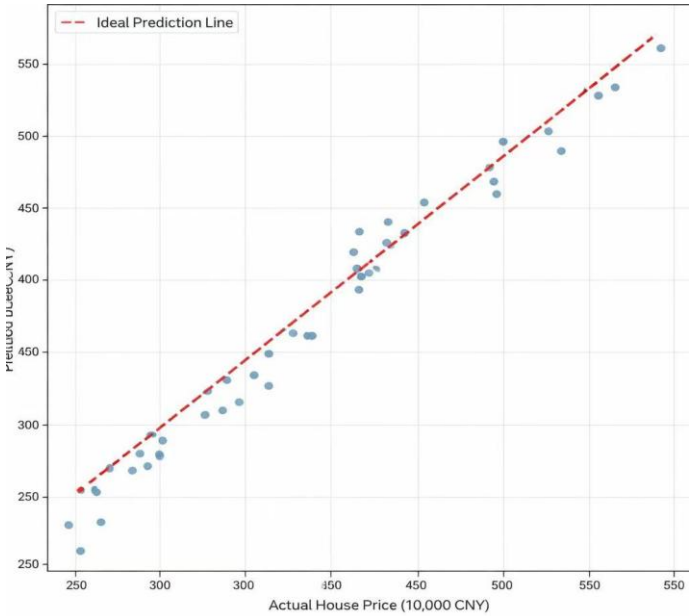


Fig. 3 Scatter Plot of Actual Values vs. Predicted Values (Photo/Picture credit: Original).

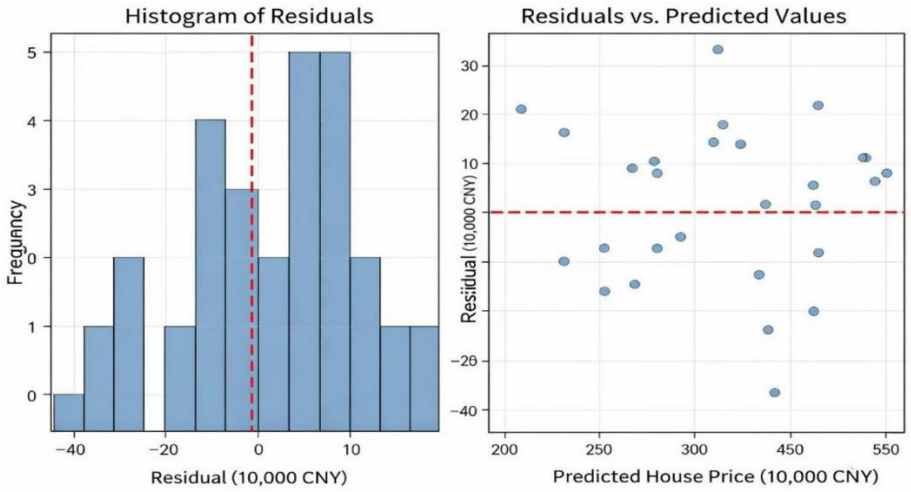


Fig. 4 Residual Analysis Plot (Photo/Picture credit: Original).

Fig. 5 shows the Q-Q diagram of the residual (quantile-quantile graph), which is used to further test the normality of the residual. It can be seen from Fig. 5 that most of the data points fall near the theoretical straight line, and there is only a slight deviation at both ends. The overall residual distribution is close to the normal distribution. The results of the Q-Q diagram are consistent with the conclusion of the residual histogram, which verifies the rationality of the normality hypothesis of the model.

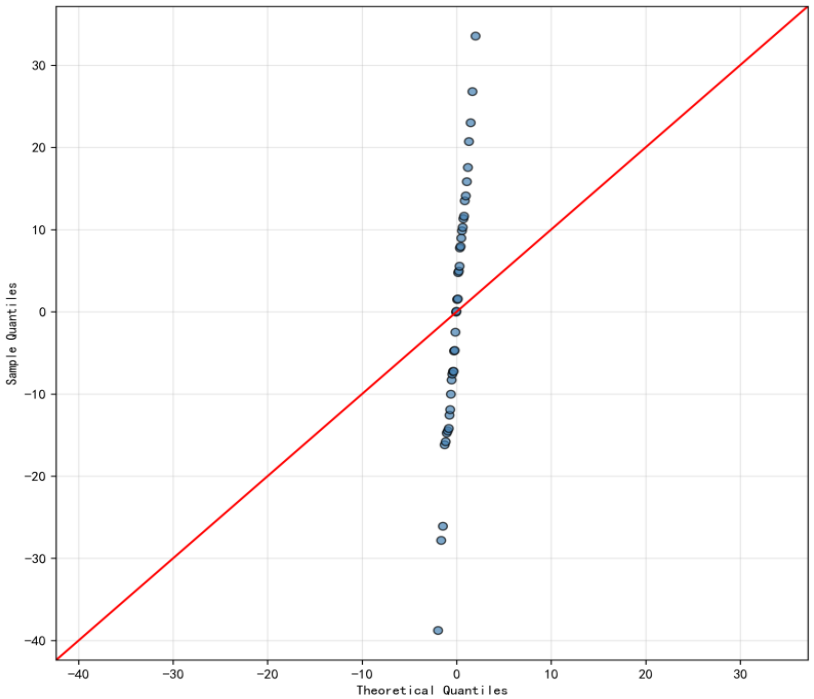


Fig. 5 Residual Q-Q Plot (Normality Test) (Photo/Picture credit: Original).

The relationship between variables and house prices: there is an obvious positive correlation between the area and house prices, and the data points are closely distributed along the trend line; there is a negative correlation between the house age and the house price, and the higher the age of the house, the lower the house price; the distance from the city center has an obvious negative correlation with the house price, and the farther the distance, the lower the house price; the number of bedrooms, the number of bathrooms and stop The number of parking spaces is positively correlated with house prices. These visualization results confirm each other with the conclusions of regression analysis, which enhances the reliability of the research conclusions.

4. Discussion

The multivariate linear regression model constructed in this study has achieved good results in house price prediction. R^2 reaches more than 0.97, RMSE control is about 150,000 yuan, and the model performance is better than or close to similar research. Although the multivariate linear regression model is not as good as the machine learning method in dealing with nonlinear relationships, it has obvious advantages in terms of model interpretability. The regression coefficient can clearly reflect the degree and direction of the influence of various factors on house prices [9]. The results of this study further verify this view. The economic meaning of each regression coefficient is clear, which is in line with the basic laws of the real estate market.

From the perspective of influencing factors, this study found that area is the most important factor affecting house prices, which is consistent with Li Yingli's research conclusion on the price of commercial housing in Shandong Province. As the core attribute of the house, the area directly determines the use value and market value of the house. The negative impact of the distance from the city center on house prices is also in line with the basic view of the location theory. The city center concentrates high-quality public service resources and employment opportunities, which has a higher location value. This study also has certain limitations [10]. First of all, the sample data comes from a single city, and the universality of the research conclusion needs to be further verified. Secondly, the model only considers the physical and location attributes of the house, and does not include macroeconomic factors, policy factors and other variables that may affect house prices, such as market expectations. The multivariate linear regression model assumes that there is a linear relationship between various factors and house prices, and may not be able to fully capture the actual nonlinear effects and interaction effects. Future research can consider expanding the sample range, incorporating more influencing factors, and trying to use machine learning methods for comparative analysis to further improve the accuracy and reliability of house price forecasts.

5. Conclusion

This study uses the multivariate linear regression method to build a house price prediction model, and systematically analyzes the impact of seven factors on house prices, such as house area, house age, floor, distance from the city center, number of bedrooms, number of bathrooms and number of parking spaces. The multivariate linear regression model shows good performance in house price prediction, training set R^2 is 0.972, test set R^2 is 0.974, RMSE is 152,300 yuan, and MAE is 124,100 yuan. The model has high prediction accuracy and good generalization ability. Each influencing factor has a different direction and degree of influence on house prices: the area, the number of bedrooms, the number of bathrooms and the number of parking spaces have a significant positive impact on the house price, among which the impact of the area is the most significant; the age of the house and the distance from the city center have a negative impact on the house price; the positive impact of the floor on the house price is relatively small, and the residual analysis It shows that the model meets the basic assumption of linear regression and has good statistical properties.

The findings of this study have important reference value for real estate market participants. For home buyers, they can reasonably weigh the impact of various factors on house prices according to their own needs and budget, and make a more rational decision to buy a house. For developers, the impact weight of various factors can be fully considered in project planning and pricing strategy, and product design and market positioning can be optimized. For the government, the research results can provide data support and decision-making reference for the formulation of real estate regulation policies.

Future research can be expanded from the following directions: first, expand the sample range and include the data of multiple cities for comparative analysis; second, introduce more influencing factors, such as school district support, convenient transportation, environmental quality, etc.; third, use machine learning methods to build nonlinear prediction models and compare and analyze with

linear models; fourth, carry out dynamic research on house price prediction and analyze the time change trend and cyclical fluctuation laws of house prices.

References

[1] Qin Yanjiao. A comparative study of house price prediction model based on multivariate linear regression and random forest algorithm. Modern Information Technology, 2024, 8(5): 112-118.

[2] Zeng Tingting. Research on house price prediction model based on machine learning. Southwest University of Science and Technology, 2020

[3] Song Yang. House price prediction model based on gradient-up decision tree. Modern Computer, 2024, 30(17): 45-51.

[4] Xue Bing, Xiao Xiao, Li Jingzhong, etc. Spatial separation and empirical evidence of factors affecting house prices in old industrial areas based on POI big data. Human Geography, 2019, 34(04): 106-114.

[5] Ji Yufan, Dai Liang, Ding Zijun, etc. Factors affecting urban second-hand housing prices and their spatial effects - empirical evidence of Nanjing based on the MGWR model. Resource Development and Marketing, 2022, 38(07): 777-783+896.

[6] Zhang Jinjuan, Guo Haiyan. Summary and Prospect of Data-Driven Real Estate Batch Evaluation Research. Journal of Xihua University (Philosophy and Social Science Edition), 2024, 43(03): 13-27.

[7] Li Yingli, Liu Yong, Liu Xiuhua. Spatial heterogeneity of factors affecting housing prices in the main urban area of Chongqing. Resource Science, 2017, 39(02): 335-345.

[8] Jin Changhong, Wang Xiaoxiao. Analysis of the factors affecting the price of second-hand houses in Hefei based on the characteristic price model. Journal of Hubei Second Normal University, 2023,40(09):53-61. Statistics and Decision-Making, 2023, 39(15): 88-92.

[9] Wang Xiaodong, Chen Xinlong, Lin Xiaoting, etc. Design and implementation of house price forecasting system based on deep learning. Computer and Digital Engineering, 2024, 52(09): 2572-2576+2650.

[10] Zhang Liping. Forecast analysis of house prices in Zhengzhou based on the ARIMA model. Real Estate World, 2025 (11).