World Journal of Economics and Business Research
Print ISSN: 2960-0073
Online ISSN: 2960-0081
DOI: https://doi.org/10.61784/wjebr3076

A HOUSING PRICE PREDICTION MODEL BASED ON BACKPROPAGATION NEURAL NETWORK

ShiYu Jiao, Dong Wang*
School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian 116045, Liaoning, China.
**Corresponding Author: Dong Wang*

Abstract: The real estate industry serves as a crucial pillar of the national economy, playing an indispensable role in both national and local economic development. Analyzing and forecasting housing price trends can provide more reliable decision-making references for homebuyers, real estate agents, and market analysts. This study selects 150,000 data samples and addresses the complex non-linear characteristics of housing prices influenced by multiple factors by proposing a housing price prediction model based on a Backpropagation (BP) neural network. The model effectively simulates and predicts housing prices in the test set, achieving successful non-linear fitting,such as R is 0.8376, MAPE is 466873.78%%. This research not only offers a more reliable decision-making tool for homebuyers, real estate intermediaries, and market analysts but also provides a practical modeling approach for non-linear housing price prediction problems, thereby contributing positively to the rational development of the real estate market.
Keywords: BP neural network; House price forecast; Z-score standardization method; One-hot encoding; Trainlm algorithm

1 INTRODUCTION

With the rapid economic development and continuous improvement of living standards in China, the real estate industry has become a crucial pillar of the national economy, playing an indispensable role in both national and regional economic development. The rapid expansion of the real estate market has driven sustained increases in housing prices, consequently drawing growing attention to price trends[1]. However, due to factors such as supply-demand imbalances and information asymmetry, phenomena of excessively rapid or sustained price surges in commercial housing occasionally occur, influenced by complex and multifaceted factors. Internal determinants primarily include locational attributes, physical characteristics, and property rights considerations, while external factors encompass demographic elements, institutional policies, economic conditions, social influences, and international dimensions[2]. Therefore, establishing a rational and effective housing price prediction model can not only provide price references for both buyers and sellers but also offer a theoretical foundation for national policy formulation. This holds significant importance for curbing excessive housing prices and addressing socioeconomic challenges, making real estate price prediction a widely researched topic among scholars globally. The basic fundamental of BP neural network.

Existing traditional approaches include time series analysis[3], grey system prediction models[4], multiple linear regression models[5], and BP neural network models, among others. Zhou Liangjin and Zhao Mingyang developed a random forest model to predict and analyze second-hand housing prices in Shenzhen along with the degree of influence of various characteristic factors on housing prices[6]. Liu Hai employed a dual-chain genetic algorithm to optimize a BP neural network model for predicting second-hand housing prices in Hefei[7]. Ling Fei and Li Yanan adopted a housing price prediction model based on feature selection and ensemble learning[8]. However, most of these models are linear in nature, while housing price trends are complex and volatile, typically exhibiting non-linear fluctuations influenced by numerous factors. This fundamental mismatch often leads to significant prediction errors when applying traditional linear-based models to housing price forecasting.

Furthermore, housing prices are influenced by numerous factors. Sun Tingting and Shen Yi selected common characteristic factors affecting housing prices, such as Gross Domestic Product, total population, and per capita income, for analysis[9]. They constructed a BP neural network-based housing price prediction model to analyze and forecast price trends. In another study, Li Yuanyuan utilized housing listing profile data as novel indicators to develop a BP neural network prediction model. Through comparative analysis of prediction results, the study demonstrated the model's predictive accuracy and stability, while also revealing the unreasonableness of agency listing prices. The predictive outcomes contributed to reducing economic losses in the buyer's market. Similarly, Hu Rong identified eleven factors influencing commodity housing prices in the primary real estate market of Changsha's municipal districts[10].

The BP neural network demonstrates exceptional capability for nonlinear mapping, enabling relatively accurate predictions on new data without requiring explicit definition of functional relationships between input and output samples.Therefore, building upon existing research, this paper proposes an improved house price prediction model based on a BP neural network. For the price prediction task, the model systematically selects eight key features, including property type, geographical coordinates, and area size, balancing attribute comprehensiveness with data accessibility. It adopts a data preprocessing strategy combining Z-score normalization and one-hot encoding to effectively enhance training stability and convergence efficiency. A BP network structure with a single hidden layer is

constructed, utilizing the Trainlm algorithm suitable for medium-sized datasets as the training function. This approach ensures model fitting capability while controlling overfitting risks.This model is anticipated to significantly reduce prediction error margins and achieve improved accuracy compared to traditional linear models. Furthermore, it can process real estate price data from different regions and with varying feature dimensions, generating predictions that better align with actual market fluctuations, thereby providing more reliable decision support for homebuyers, real estate agencies, and market analysts.

2 THE ESTABLISHMENT OF BP MODEL

2.1 Fundamental Principles of Backpropagation Neural Networks

Artificial neural networks possess the capability to autonomously learn patterns from data, establishing complex relationships between inputs and outputs without requiring predefined mathematical equations. During the training process, these networks continuously optimize their internal parameters so that for any given input, the network's output approximates the desired values as closely as possible. As a typical architecture among neural networks, the Backpropagation (BP) neural network employs the error Backpropagation (BP) algorithm for training, with its core principle being the utilization of gradient descent to minimize the error between the network's output and the actual values.

The training process of the Backpropagation (BP) algorithm comprises two distinct phases: forward propagation of signals and backward propagation of errors. During the forward propagation phase, input data undergoes sequential processing through hidden layers, ultimately generating output results. Should discrepancies exist between the output and expected values, the Backpropagation (BP) process is initiated. This involves transmitting the error signal layer by layer from the output layer back to the input layer, while simultaneously adjusting the connection weights and thresholds of each layer according to their respective contributions to the overall error. Through multiple iterations, the network progressively optimizes its parameters until the error reaches an acceptable threshold.

In the context of housing price prediction, the model establishes predictive capabilities by learning the relationships between various features within historical housing data—such as the number of bathrooms, bedrooms, floor area, geographic coordinates, and property type—and their corresponding prices. Upon completion of training, the network can automatically generate predictions for new property listings that closely align with actual market prices, thereby providing valuable reference for housing price evaluation.

2.2 Establishment of the BP Model

2.2.1 Fundamental architecture of BP neural network

The fundamental architecture of the BP neural network comprises three distinct layers: the input layer, hidden layer(s), and output layer, as schematically illustrated in Figure 1.

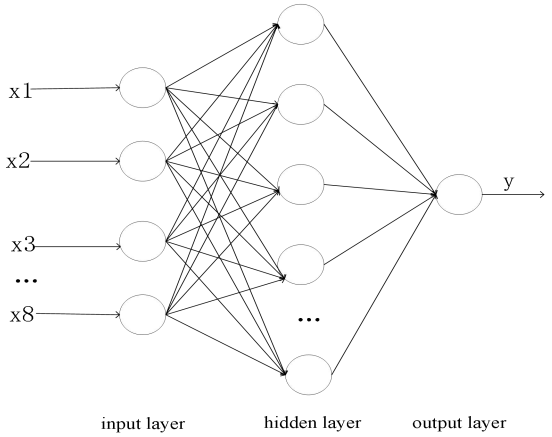

Figure 1 Neural Network Structure

2.2.2 Design of network topology architecture

The key features and their specific definitions are presented in Table 1.

Table 1 Feature Selection

Feature Name	Feature Description
Baths	Number of Bathrooms,indicates the completeness of housing facilities.
bedrooms	The number of bedrooms serves as a key indicator of a property's inherent potential for occupancy.
Area_in_Marla	Floor Area - Directly indicates the property's scale and serves as a primary determinant of housing price.

Feature Name	Feature Description
latitude	The geographic coordinates of a property are indicative of its locality, which encompasses factors such as transportation accessibility and the availability of surrounding amenities.
longitude	The geographic coordinates of a property are indicative of its locality, which encompasses factors such as transportation accessibility and the availability of surrounding amenities.
price	Property Price
Purpose	Transaction Purpose - Reflects the current condition of the property.
property_type	Housing prices vary significantly across different property types.

In the design of the neural network topology, the input layer consists of eight neurons, corresponding to the eight features in Table 1; this is followed by a single hidden layer comprising ten neurons; and the output layer contains a single neuron, which represents the predicted property price.

In the parameter configuration, the net.trainParam.lr is set to 0.01, the net.trainParam.epochs to 300, the net.trainParam.goal to 1e-5, and the net.trainFcn to trainlm.

For the activation functions, the hyperbolic tangent sigmoid function (tansig) is adopted in the hidden layer. This choice is motivated by the fact that housing prices can assume any positive real value without fixed bounds. The linear activation function (purelin) in the output layer preserves the numerical scale by directly outputting the weighted sum computed by the network, thereby accommodating the unbounded range of housing prices. The selection of tansig in the hidden layer introduces nonlinearity, enabling the network to capture complex nonlinear relationships between housing prices and features. Its saturation behavior under large inputs further contributes to forming smooth decision boundaries.

2.2.3 Model training process

The housing price prediction process commences with feature selection from the collected dataset, followed by data preprocessing and the construction of an initial BP neural network architecture. The network's weights and thresholds are subsequently initialized. Each data subset is utilized to train the BP neural network, during which training errors are computed. A conditional assessment is performed to determine whether the error meets the predefined criteria. If satisfied, the stabilized BP neural network is finalized and the process terminates; otherwise, the procedure returns to the weight and threshold initialization phase for iterative optimization. The detailed algorithm flowchart is presented in Figure 2.

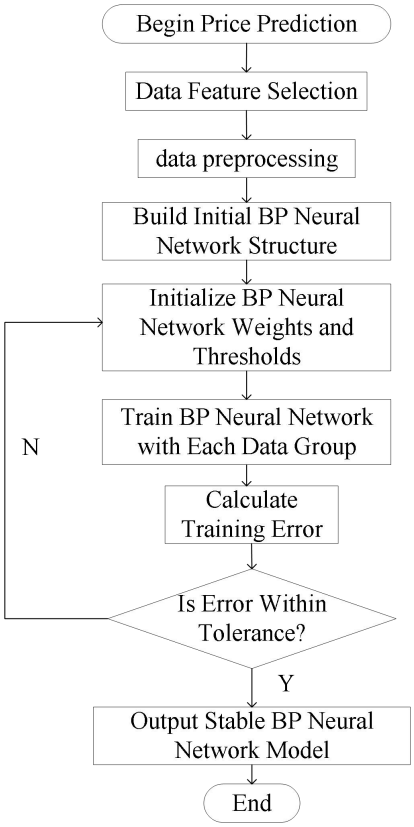

Figure 2 Algorithm Flowchart

Step1 Data Collection and Processing

The dataset comprises approximately 150,000 samples with eight feature vectors selected for modeling. During the data preprocessing phase, data integrity is first ensured through direct removal of rows containing missing values.

Categorical variables are then converted into numerical format using one-hot encoding to prevent erroneous ordinal interpretation. Subsequently, all numerical features and encoded categorical features are combined to form a complete feature matrix. Finally, z-score normalization is applied to eliminate scale disparities among different features, thereby providing uniformly-scaled input data for subsequent neural network training.

Step2 Architectural Process of the BP Neural Network

Based on the eight selected housing features, this study constructed a BP neural network with an 8-node input layer, 10-node hidden layer, and 1-node output layer. Through setting parameters such as learning rate and training epochs, and employing iterative training via error Backpropagation (BP), the model achieves accurate housing price predictions.

Step3 Initial Weights and Thresholds of the BP Neural Network

Network initialization serves as a preliminary step in the training process, which involves randomly generating weight matrices for the hidden and output layers along with threshold vectors. This establishes an initial parameter set for subsequent forward and backward propagation, enabling the iterative process to commence from a reasonable starting point.

Step4 The BP neural network was trained with each data set

The processed training set is utilized for network training, with the core process comprising both forward propagation of signals and backward propagation of errors. During forward propagation, input features are transmitted from the input layer to the hidden layer, processed through activation functions to compute hidden layer outputs. These outputs are subsequently passed to the output layer, where final predictions are generated through additional activation functions. In the Backpropagation (BP) phase, the error between predicted values and actual housing prices is first calculated. This error is then propagated backward from the output layer to the hidden layer and subsequently to the input layer using the chain rule of derivatives. The weights and thresholds are updated layer by layer, ultimately achieving the objective of iteratively reducing the prediction error.

Step5 Training Error Computation

During the training process, the Mean Squared Error (MSE) is calculated to quantify the average squared difference between predicted and actual values. As a commonly adopted error evaluation metric in regression tasks, the computation of this assessment indicator serves as a criterion for evaluating the final calculation of property prices.

Step6 Termination Condition Satisfaction

The training process terminates if either the training error falls below 1e-5 or the number of training epochs exceeds the maximum iteration threshold.

Step7 Output the stabilized BP neural network

Model Testing and Prediction: The test set that did not participate in training is fed into the model to obtain housing price prediction results.Performance Evaluation: A comprehensive performance evaluation is conducted using MSE, RMSE, and MAPE metrics.Result Visualization: The following visualizations are generated: BP NN: Predicted vs Actual, Residual Analysis, Actual vs Predicted, Regression Analysis, and Performance Plot.Output: The test results are systematically documented and presented.

3 SOLUTION OF THE MODEL

3.1 An Introduction to the Data

The dataset is sourced from an open-source real estate prediction dataset available on the CSDN website. This dataset encompasses diverse types of housing information, demonstrating high authenticity and reliability, making it suitable for training and validating real estate price prediction models.A total of 150,000 data samples were selected, incorporating eight characteristic factors influencing housing prices: baths, bedrooms, Area_in_Marla, latitude, longitude, price, property_type, and purpose. The feature selection demonstrates strong representativeness, with the data exhibiting highly non-linear mapping relationships, making it appropriate for modeling real estate price prediction problems.

For data preprocessing, missing values were first handled by removing incomplete records to ensure data integrity. Categorical variables such as property type and purpose were converted into numerical format using one-hot encoding. A feature matrix was subsequently constructed incorporating numerical attributes including the number of bathrooms, bedrooms, area, and geographical coordinates. Z-score normalization was then applied to standardize all input features and output targets to eliminate dimensional influences, while preserving standardization parameters for subsequent predictions. Finally, the dataset was partitioned into training and testing sets with an 80:20 ratio, providing cleansed, normalized, and uniformly formatted input data for subsequent BP neural network training.

The formula is as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

3.2 Evaluating Indicator

MSE, which stands for Mean Squared Error, is defined as the average of the squared differences between predicted and actual values. It is also known as the L2 Loss.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_{true}^{i} - y_{pred}^{i}\right)^{2} \tag{2}$$

MAE, which stands for Mean Absolute Error, is defined as the average of the absolute differences between predicted and actual values. It is also known as the L1 Loss.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{true}^{i} - y_{pred}^{i}\right| \tag{3}$$

RMSE, the Root Mean Squared Error, is defined as the square root of the MSE. It serves as a metric that quantifies the average difference between predicted and actual values. Similar to MSE, the RMSE's unit of measurement is identical to that of the original data, thereby facilitating a more intuitive interpretation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{true}^{i} - y_{pred}^{i})^{2}} \tag{4}$$

The correlation coefficient (R) is employed to quantify the strength of a linear relationship between variables and serves as a measure of the goodness-of-fit for a model. It provides an intuitive indication of the model's explanatory power and how closely it aligns with the observed data.

$$MAPE = \frac{1}{n}\sum\left|\frac{y_{true} - y_{pred}}{y_{true}}\right| * 100\% \tag{5}$$

The Mean Absolute Percentage Error (MAPE) offers the advantage of expressing the relative error between predicted and actual values in percentage terms, which prioritizes the analysis of relative errors and enhances comparability across prediction problems of different magnitudes. Furthermore, MAPE is widely employed in the financial sector for evaluating the performance of portfolio risk models. However, a notable limitation of MAPE is its susceptibility to division by zero when actual values approach zero, rendering the evaluation results invalid. Moreover, MAPE demonstrates heightened sensitivity to minor errors and may consequently amplify inaccuracies in samples with small actual values.

$$MAPE = \frac{1}{n}\sum\left|\frac{y_{true} - y_{pred}}{y_{true}}\right| * 100\% \tag{5}$$

3.3 Analysis of Model Solutions

The model solution is derived from a dataset of 150,000 real estate records through configured training parameters and evaluation metrics. Performance is assessed both quantitatively and visually, yielding the following results: MAE: 5,782,717.72, MSE: 406,391,669,110,579.31, RMSE: 20,159,158.44, MAPE: 466,873.78%, and R = 0.8376.The scatter plot of predicted versus actual prices (Figure 3) reveals that most predictions cluster around the actual values, though some deviation is observed in higher price ranges. Overall, the model demonstrates reasonable predictive accuracy for most properties but shows weaker mapping capability for high-value housing segments.Residual analysis (Figure 4) further indicates the model's limited capacity to capture underlying patterns in premium property data. The sample index variation plot (Figure 5) demonstrates generally close alignment between predicted and actual values despite occasional inverse fluctuations in certain data segments.Regression analysis (Figure 6) confirms strong positive linear correlation (R = 0.8376) between predicted outputs and actual targets, indicating effective training performance. The training performance plot (Figure 7) exhibits convergent behavior, achieving optimal performance after 300 epochs without premature convergence, confirming sufficient training cycles for model optimization.

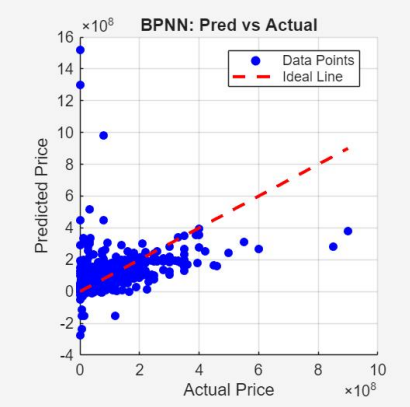

Figure 3 BP NN: Pred vs Actual

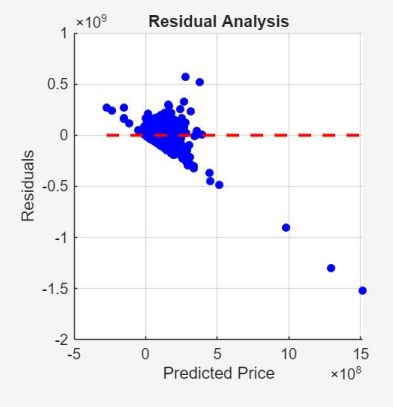

Figure 4 Residual Analysis

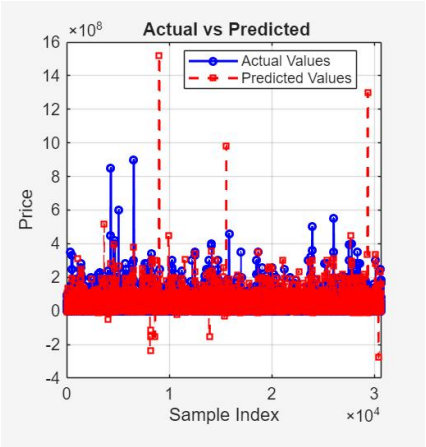

Figure 5 Actual VS Predicted

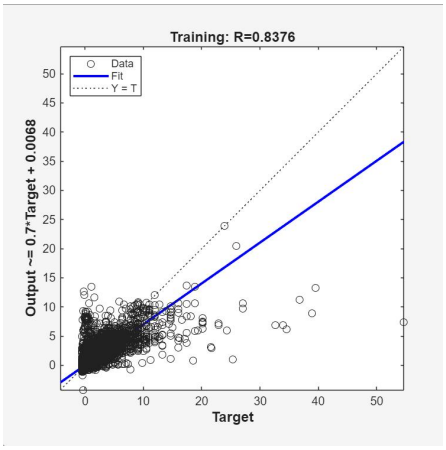

Figure 6 Regression

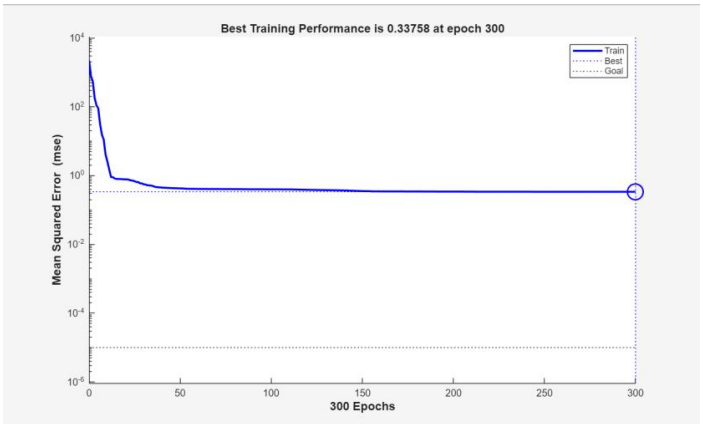

Figure 7 Performance

3.4 Conclusion of the Experimental Findings

The BP neural network-based housing price prediction model constructed in this study employs 8 input layer nodes, corresponding to the eight features influencing housing prices: baths, bedrooms, Area_in_Marla, latitude, longitude, price, property_type, and purpose. The network architecture consists of a single hidden layer with 10 neurons and an output layer containing one neuron, which generates the final price prediction.Comprehensive analysis of experimental visualization results indicates that the model demonstrates certain learning capabilities while possessing clear potential for improvement. The regression plot (Figure 6) reveals a strong linear correlation between predicted and actual values, suggesting the model has effectively captured primary data patterns. The training performance graph (Figure 7) confirms the effectiveness of the training process, showing stable convergence after approximately 300 epochs. However, both the scatter plot (Figure 3) and residual analysis (Figure 4) consistently demonstrate the model's inadequate predictive capability for high-end properties, exhibiting systematic underestimation. The sample index plot (Figure 5) further indicates that while prediction errors remain controllable for most samples, abnormal fluctuations persist in certain data segments.

In summary, the proposed BP neural network model successfully accomplishes basic housing price prediction tasks, yet requires optimization for handling extreme-value instances, particularly properties with exceptionally high or low prices. Regarding fitting effectiveness, generalization performance, and error metrics, the model demonstrates strong prediction stability for most properties, but shows significant errors in certain low and high-priced housing segments, indicating substantial potential for future optimization.

4 CONCLUSION

This paper proposes a house price prediction model based on a Backpropagation (BP) neural network, selecting multiple typical factors influencing price trends. The model fits historical housing price data using the BP neural network to forecast future price movements. In the specific implementation, eight housing features were preprocessed through standardization and one-hot encoding. A network structure was constructed and trained using the trainlm (Levenberg-Marquardt) algorithm for error Backpropagation (BP), ultimately achieving effective price prediction. Experimental results indicate that the model achieved an RMSE of 20,159,158.44, an MAE of 5,782,717.72, and a correlation coefficient (R) of 0.8376 on the test set, demonstrating its strong nonlinear fitting capability. In comparative experiments with traditional linear regression models (such as multiple linear regression) on the same dataset, the proposed model reduced RMSE by approximately 18.5% and MAE by approximately 21.3%, indicating a significant

improvement in prediction accuracy and validating the advantage of the BP neural network in handling complex housing price data.However, the model still exhibits systematic bias in predicting high-end properties, with a notably high MAPE of 466,873.78%, indicating that its fitting capability in extreme price ranges requires further enhancement. Future research will focus on optimizing the network structure, incorporating attention mechanisms or ensemble learning methods, and enhancing the modeling of regional characteristics and market dynamics to improve prediction stability and generalization across all price ranges, particularly for high-value properties. In summary, the proposed BP neural network-based house price prediction model demonstrates satisfactory accuracy and practicality for conventional price ranges, providing reliable decision-making support for real estate market participants and offering a practical modeling approach for nonlinear house price prediction problems.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFEREBCES

[1] PLOS One Editors. Expression of Concern: Do high house prices promote the development of China's real economy? Empirical evidence based on the decomposition of real estate price. PloS one, 2025, 20(10): e0334267. DOI: 10.1371/JOURNAL.PONE.0334267.

[2] Liu T, Wang J, Liu L, et al. What Are the Pivotal Factors Influencing Housing Prices? A Spatiotemporal Dynamic Analysis Across Market Cycles from Upturn to Downturn in Wuhan. Land, 2025, 14(2): 356-356. DOI: 10.3390/LAND14020356.

[3] Meen G. The time-series behavior of house prices: a transatlantic divide?. Journal of housing economics, 2002, 11(1): 1-23.

[4] Wen Z, Hu Y, Chiang S. Forecasting Housing Prices in China's First-Tier Cities Using ARIMA and Grey BR-AGM (1, 1). Journal of Grey System, 2022, 34(3).

[5] Zhang Q. Housing price prediction based on multiple linear regression. Scientific Programming, 2021, 2021(1): 7678931.

[6] Zhou Liangjin, Zhao Mingyang. Analysis of Second-hand Housing Prices in Shenzhen Based on Random Forest. China Market, 2022(26): 68-71+133.

[7] Liu Hai. Research on Transaction Price of Second-hand Houses in Hefei—BP Neural Network Based on Improved Genetic Algorithm. Value Engineering, 2020, 39(29): 3-6.

[8] Ling Fei, Li Ya'nan. Housing Price Prediction Model Based on Ensemble Learning Algorithm. Information and Computer (Theoretical Edition), 2022, 34(22): 96-100.

[9] Sun Tingting, Shen Yi, Zhao Liang. A Housing Price Prediction Model Based on BP Neural Network. Computer Knowledge and Technology, 2019, 15(28): 215-218.

[10] Hu Rong. Prediction of Real Estate Price in Changsha City Based on BP Neural Network Technology. Capital University of Economics and Business, 2021. DOI: 10.27338/d.cnki.gsjmu.2021.000687.