Research on the Price Prediction of Commercial Housing in Beijing

Yiyang Zhou[1], Xinping Liu[2,*] and Mingju Sun[3]

[1]College of International, Zhengzhou University, Henan, 450000, China

[2]College of Art and Science, The Ohio State University, Columbus, 43210, America

[3]ZhuCheng No.1 High School, Shandong, 262200, China

*Corresponding author: liu.11429@buckeyemail.osu.edu

Abstract. In order to determine the best appropriate housing price prediction model, this study constructs a linear regression model, a BP neural network model, and a time series model based on data on commercial housing prices and their affecting variables in Beijing during the past 23 years. A set of 12 independent factors that impact the cost of commercial housing is selected first. Then, three models are created, and their fitting effects are compared and analyzed to forecast the average cost of commercial real estate in Shandong Province. It is shown that the multiple linear regression model is more appropriate for long-term prediction and that the time series model predicts data more recently than the BP neural network model; Therefore, it can be used if the prediction period is longer. Although home prices can be predicted theoretically using these three approaches, more work has to be done to fine-tune the applicable models.

Keywords: House prices prediction; ARIMA model; Back Propagation neural network; Beijing; multiple linear regression model.

1. Introduction

The cost of commercial real estate in China has generally increased in recent years, however there has been a little drop lately. Nonetheless, there appears to be a continued increase in the demand for homes, a phenomenon known as "buying and selling" in the real estate market, and a number of issues that have led to the real estate industry's development becoming increasingly deviant [1]. The price of commercial housing has always had an important impact on the country's economy, social stability, and people's quality of life. Shandong coastal cities are more, the geographical location is superior, is China's population of hundreds of millions of economic provinces, the level of housing prices in China is relatively high, and other provinces, after experiencing the tide of the property market in previous years, the trend of housing prices in Shandong Province has also begun to appear to a certain extent [2]. Therefore, it is particularly important for people who have the need to buy a house to accurately understand the development trend of housing prices and find a more suitable method for housing price forecasting.

As for the problem of housing price prediction, many scholars have conducted corresponding research. The results demonstrate that the gray Markov model has high fitting accuracy and is very suitable for the prediction of housing price index. Cheng established a gray and gray Markov prediction model based on the price of real estate in Wuhan to rank the factors affecting the housing price in Wuhan. This was done using a combination of qualitative and quantitative, theoretical and empirical methods [1]. Sun et al. used regression analysis to statistically learn the typical factors affecting real estate prices [3]. Then, the BP neural network model is constructed based on the learned elements, and the network parameters are trained and verified by historical data to predict the future trend of housing prices [4, 5]. Li and Zhang selected China's housing prices and their main influencing factors as experimental data for simulation training, and compared the prediction effect of the model [6]. Experimental results show that the BP neural network improved by genetic algorithm has higher prediction accuracy and faster convergence speed, which can avoid falling into the trap of local optimum [7, 8]. Zheng and Liu used the ARIMA model to make a rolling prediction of future housing prices, and used RMSE to judge the prediction accuracy [9]. The results show that the model can

continuously predict the average price of second-hand houses with high prediction accuracy, which can provide reference for house buyers and sellers [10]. When Qiu and Yu employed the principle component analysis-based BP neural network model to forecast housing prices, they discovered that the developed model's prediction error was very low, making it a useful tool for housing price prediction. Gao and Zhang proposed a BP neural network model optimized by genetic algorithm for housing price prediction. Wang and Gao verified that the BP neural network model has a strong validity in the prediction of housing prices in Chongqing.

This paper collects and sorts out the relevant data of the average price of commercial housing in Beijing from 1999~2021 and some important influencing factors, reduces the dimensionality of the data, selects the variables that have a relatively important impact on housing prices, and then establishes three different types of models to predict the prices of commercial housing in Beijing in 2022, in order to select a suitable housing price prediction model, so as to make more accurate prediction of future housing price trends and provide some effective price references for future house purchases.

2. Methods

2.1. Data Resource

The data from the National Statistical Yearbook and the National Bureau of Statistics, Select the relevant data of 1999-2021, the Beijing commodity houses. The following is the data selection website: http://www.stats.gov.cn/sj/ndsj/, http://www.stats.gov.cn/zs/tjwh/tjkw/tjzl/.

The following are the symbols and corresponding meanings of the selected variables: Y is the price of commodity housing in Beijing (CNY/m2). The explained variables are shown in the table below (Table 1):

Table 1. Descriptive of the explained variables

variables	Description
X1	Investment in real estate development (108CNY)
X2	The per capita disposable income of urban residents in Beijing (CNY)
X3	The GDP of Beijing (108CNY)
X4	The GDP growth rate of Beijing (%)
X5	The sales area of commercial housing (104m2)
X6	The natural population growth rate (%)
X7	The area of real estate development enterprises completed (104m2)
X8	The cost of the completed housing of real estate enterprises (CNY/m2)
X9	The GDP (108CNY)
X10	The real estate tax (108CNY)
X11	The total consumer price index of Beijing (%)
X12	The population of Beijing at the end of the year (104)

2.2. Variable Selection

To make the model more concise and efficient, first select the variables. Variable selection can help us select the optimal variable from more independent variables to build a more simple and effective prediction model, so as to save the measurement cost of independent variables and improve the accuracy of the model. Plot a thermodynamic chart of the correlation of each variable (Fig. 1).

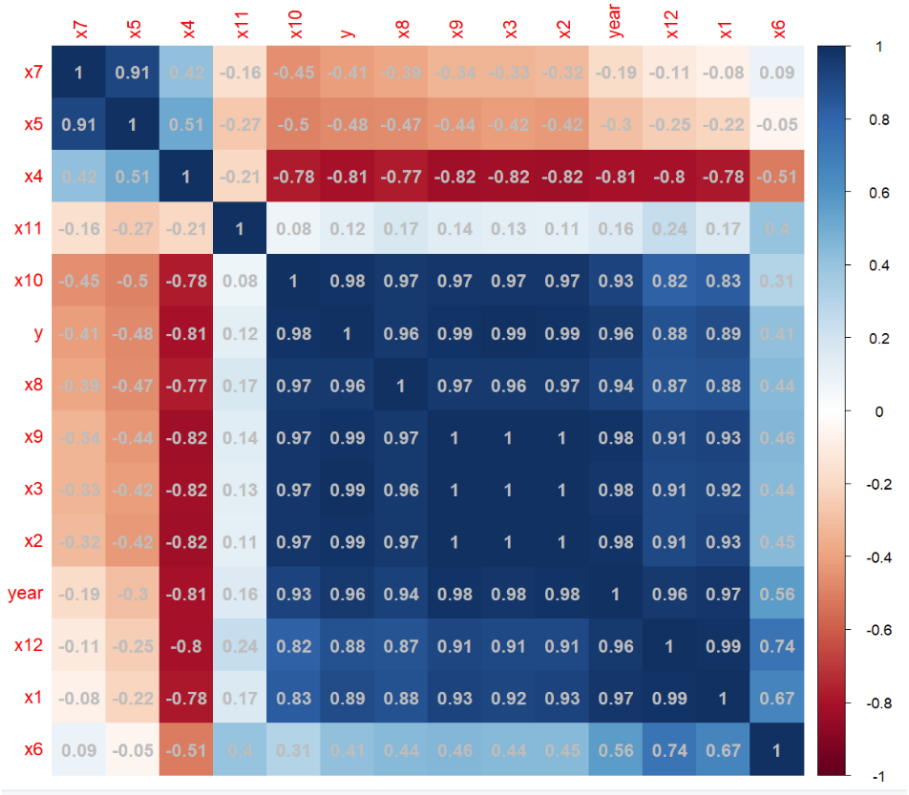

Fig. 1 Thermodynamic chart

According to the scatter plot of the correlation relationship, it can be seen that the 12 independent variables we selected have multicollinearity to prevent the accuracy of the subsequent prediction model from being affected. After the stepwise regression is completed, six significant variables, X1, X2, X3, X5, and X8, are selected as independent variables to establish the model. Plot a scatter plot of the correlation of each variable (Fig. 2).

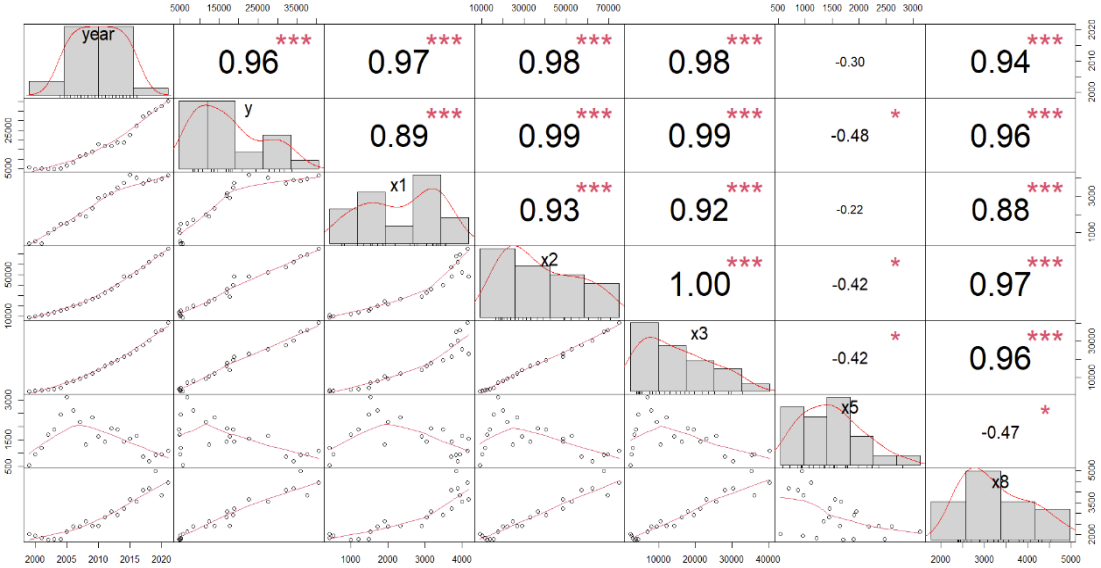

Fig. 2 Scatter plot of variable correlations

2.3. Model Selection

In this paper, due to the study of the price of commodity housing in Beijing, based on the influencing factors, multiple linear regression model (Mul-linear), BP neural network model (BP-NN), and ARIMA model are established respectively, and the fitting effect of three different models on the price prediction of commercial housing in Beijing is compared and analyzed.

Since the purpose of this paper is to predict housing prices in Beijing, according to previous literature and surveys, it is found that housing prices are related to a variety of factors, so we establish a multiple regression model with X1, X2, X3, X5, and X8 as variables, so as to further analyze the changes in housing prices.

BP neural networks are multi-layered feedforward neural networks that are suitable for solving complex nonlinear problems. Firstly, the data is normalized using range normalization, so as to effectively eliminate the difference in units of measurement. Next, the price of commercial housing from 99 to 21 years is used as the training set. Then the accuracy of the model is reversed and tested by using the price of commercial housing in Beijing in 2022 as the test set, so as to judge.

The ARIMA model treats a sequence of data formed by the predicted object over a period of time as a random sequence, makes a difference to stabilize the sequence, and then uses a mathematical model to approximate the sequence. Firstly, the stationarity of the data is tested. The white noise test is performed on the original data to determine whether there is a trend in the original data. The unstationary data is differentially processed to perform the model order, and finally, the model significance test and related prediction are carried out.

3. Results and Discussion

3.1. Multiple Linear Regression Model

The Mul-linear regression model can be written as:

$$\boldsymbol{y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon} \tag{1}$$

Where $X_1, \ldots, X_p$ are independent variables; $\beta_0, \ldots, \beta_p$ are regression parameters; ε indicates an error. The R language statistical software is used to process the data, and the OLS regression model is obtained as follows:

$$\boldsymbol{y = 2.05X_1 - 0.45X_2 + 1.37X_3 - 1.37X_5 - 1.98X_8 + 64.06X_{10} + 9952.48} \tag{2}$$

Table 2. Model Results

Model	β	Std. Error	T	P
Constant	9952.477	2632.192	3.781061	0.0016
X_1	2.052692	0.889098	2.308734	0.0346
X_2	-0.445611	0.184745	-2.412035	0.0282
X_3	1.092096	0.281456	3.880164	0.0013
X_5	-1.370460	0.505747	-2.709773	0.0155
X_8	-1.978998	1.206153	-1.640753	0.1204
X_{10}	64.06400	17.78127	3.602892	0.0024
F	F statistic=367.4962, P=0.000			

The significance result test results are as follows. R^2 is 0.993, indicating that the model fits the sample very well, the F statistic is 367.4962, and the F test result is significant, p<2.871e-15, which indicates that the established model is appropriate.

Calculate the regression value and residuals of the model: The model passes the residual test, and the residuals roughly obey the normal distribution, so it is concluded that the multiple linear regression model established can be used for housing price prediction.

3.2. BP Neural Network Model

The data from 1999~2021 are used as the training set, and the independent variable data in 2022 are used as the test set to divide the original data, and the neural network is used to process the training data.

The accuracy of the BP neural network model established by the test set is tested several times, and the results show that when the threshold and weight are still the default values of the system, and the number of neurons is 8.

The RMSE reaches 0.0235, indicating that the model prediction effect is relatively good, and the model fitting effect and its performance are improved. R^2 is 0.994, indicating that the model fits very well. The MAPE is 4.4509%, indicating better prediction performance.

3.3. ARIMA Model

According to the time series diagram, it can be seen that the fluctuation range of the data is unbounded and has an increasing trend, showing the characteristics of unstable. The ADF test of the time series yields a p-value of 0.6533, which indicates that the series is a non-stationary series, so it needs to be differentially processed to make it stationary (Fig.3).

After several attempts, it was finally found that after using four first-order differential processes, the p-value of the sequence was 0.015, which was less than 0.05, which was determined to be significantly stable at this time (Fig.4).

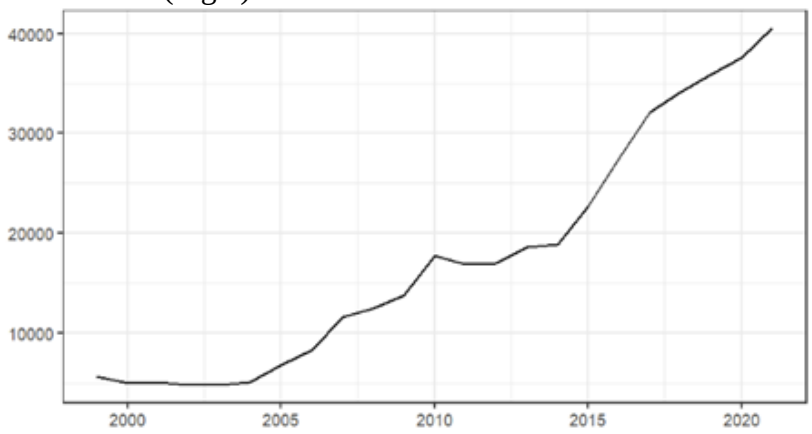

Fig. 3 Original Sequence-Diagram

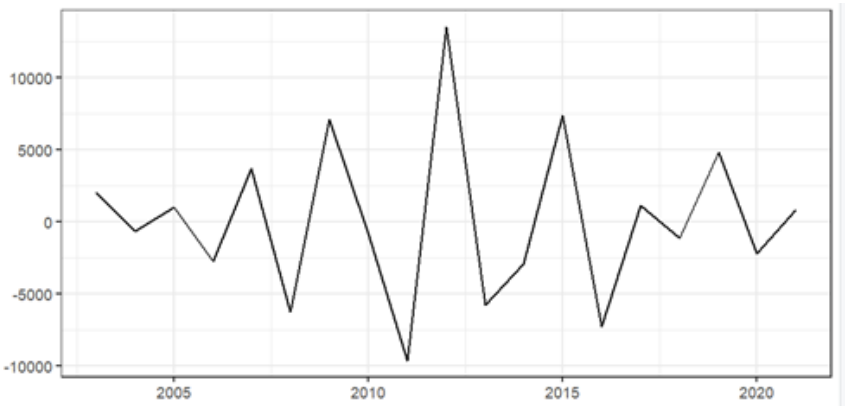

Fig. 4 After-Diff Sequence-Diagram

After determining the difference order, the ARIMA(2,4,2) model was obtained by the auto. arima function and the p-value was finally determined to be 0.71 by using the Ljung-Box test, greater than 0.05, the results show that the residuals are a white noise sequence, and the sequence does not care, and the ARIMA(2,4,2) model is significantly effective. Plot both the Autocorrelation Function and the QQ Plot for a more pronounced view of the sequence (Fig.5).

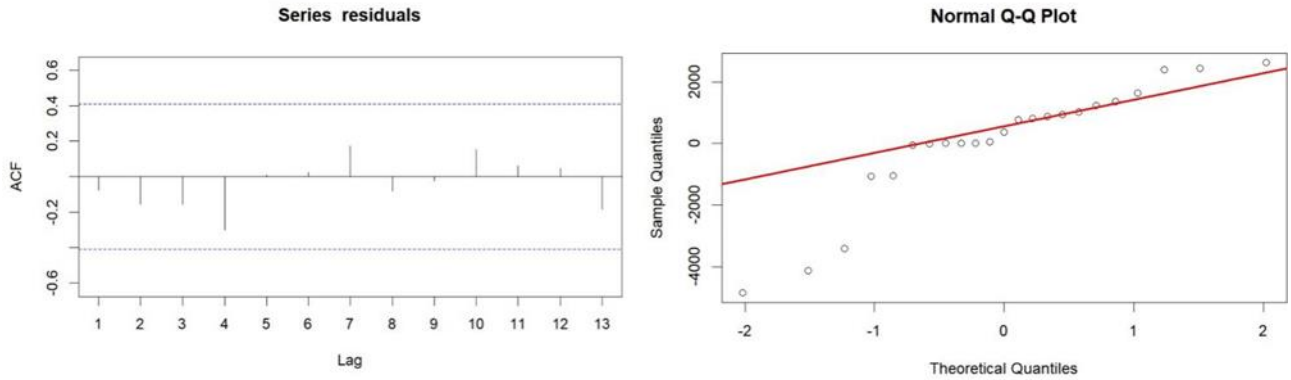

Fig. 5 Autocorrelation Function & Q-Q Plot

In Fig.5, the left figure shows the autocorrelation diagram of the residual sequence, which clearly shows that the residual sequence of ARIMA(2,4,2) is a white noise sequence, and the model passes the test; the graph on the right shows the normality test of the residual sequence. The points on the graph are densely distributed in the red line (symmetric line), indicating that the sequence follows the normal distribution, which indicates that the model is significant.

3.4. Comparison of Model Results

In this paper, multiple linear regression model, BP neural network model and ARIMA model are established. The following table (Table 3) is a comparison of the results of the model test.

Table 3. Prediction accuracy metrics

Model	Multiple linear regression	BP Neural Network	ARIMA
RMSE	919.2797	842.0623	2378.212
R^2	0.9937401	0.9947475	0.989921
Average error	0.06668255	0.044509	5.773505

From the comparison results in Table 2, the RMS error of BP-NN is small, indicating that the deviation between the results of BP-NN and the actual value is minimal. From the perspective of goodness of fit, the R2 of the three models is very different, but the R2 value of BP-NN is the largest, indicating that the model has the best fitting effect. The average relative error of the BP neural network model is the smallest, which indicates that the prediction data is the most accurate. To sum up, the prediction effect of the BP neural network model should be the most stable, and the fitting effect of the ARIMA model is the worst.

The above three models are respectively used to forecast the average price of commodity housing in Beijing in 2022, and the comparison is made with the real value data, as shown in Table 4:

Table 4. True value and model-predicted value

Year	True Value	Multiple linear regression	BP Neural Network	ARIMA
2022	41191.83	44535.76	45152.36	43570.04

From the prediction results of the three models, it can be found that the prediction data of the ARIMA model is closest to the true value and has the highest prediction accuracy, followed by the Mul-linear. Although the BP-NN has the best fitting value, the predicted value is relatively high, and the deviation from the true value is large.

After discussion, it was found that due to Covid-19 and related policy reasons, there may be certain difficulties in the sales process of commercial housing, the income growth of domestic residents has slowed down, the willingness to buy houses has declined, and the oversupply of real estate in various cities has continued to shrink. As a result, the sales price of commercial housing in Beijing in 2022 is lower than predicted by the model based on previous data.

4. Conclusion

In the paper, the Mul-linear model, BP-NN model, time series model are established to predict the typical cost of business properties in Beijing, and it is found that the multiple linear regression model is more suitable for long-term prediction, and this model can be used if the prediction period is longer, the BP-NN model and the time series model are more suitable for predicting recent data, and the forecasting impact of the time series model is slightly worse by comparison. This may be related to the fact that the model itself has strict requirements for historical data, and the model's predictive accuracy is susceptible to external influences.

At the same time, due to the effects of the novel coronavirus outbreak and related policies, the selling price of commercial housing in 2022 is lower than the forecast value, but the results are still within the forecast range. At the same time, it is necessary to take into account the relevant local policies and economic development and other related factors to further process the forecast results, so that these three methods can theoretically be used to predict housing prices, but it remains crucial to meticulously consider the specifics and continually enhance the pertinent models.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Cheng Songlin. Price prediction and analysis of commercial housing in Wuhan based on grey theory. Central China Normal University, 2008.

[2] Housing price prediction based on BP neural network. Chengdu: Southwest Petroleum University, 2011.

[3] 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.

[4] Wang Xiaoxin, Gao Pan. Verification and prediction of housing prices in Chongqing based on the BP neural network. Journal of Chongqing University of Technology (Social Sciences), 2016.

[5] Qiu Qirong, Yu-Ting. Research on prediction of housing price by BP neural network based on principal component analysis. Journal of Hunan University of Arts and Sciences (Natural Science Edition), 2011, 23(03): 24-26+36.

[6] Li Chunsheng, Li Xiaoye, Zhang Kejia. Housing price prediction analysis of BP neural network based on genetic algorithm. Computer Technology and Development,2018, 28(08): 144-147+151.

[7] Gao Yuming, Zhang Renjin. Housing Price Prediction Analysis Based on Genetic Algorithm and BP Neural Network. Computer Engineering, 2014, 40(04): 187-191.

[8] Chen Shipeng, Jin Shengping. Housing price prediction based on random forest model. Science and Technology Innovation and Application, 2016, 4: 52.

[9] Wu Xiuli, Zhang Feng. Application of time series analysis method in housing price forecasting: A case study of Guangzhou data. Science Technology and Engineering, 2007, 21: 5631-5635.

[10] Zheng Yongkun, Liu Chun. Price prediction of second-hand housing based on the ARIMA model. Computer and Modernization, 2018, 4: 122-126.