Housing Price Prediction Model and Impact Factors Analysis

Jingxuan Hao *

School of Computer Science and Information, Cardiff University, Cardiff, UK

* Corresponding author email: haojingxuan@yczhsyxs.com

Abstract. Housing price forecast is usually used in macroeconomic regulation, which can effectively avoid the housing price explosion brought by economic growth and promote better housing market control. Considering the rapid growth of housing prices in China in the past decade, the prediction and analysis of housing prices have become a top priority. In this paper, we establish the VAR model to understand the rule of housing prices between different cities and reveal some potential factors affecting housing prices through variance decomposition and impulse response. The results of the experiment show that the factors affecting the change in housing prices are different for different regions, but it is undeniable that housing prices are often affected by the prices of previous years. These results will effectively assess the housing market in these cities and help the government make decisions.

Keywords: House Price Prediction; VAR; Granger Causality.

1. Introduction

The Real Estate Investment Research Center of Zhejiang University and the Media Survey Laboratory of Tsinghua University conducted a nationwide survey on the living conditions of Chinese urban residents in 2012[1]. According to the survey, which covers 40 cities, more than 20 per cent of respondents do not own property, and the ratio of home ownership in economically developed provincial capitals or key cities is relatively low, especially in the three first-tier cities of Beijing, Shenzhen and Shanghai, which rank the last three among the 40 cities due to their higher housing prices. Therefore, exploring the main factors affecting housing prices has become one of the important conditions for the formulation of policies to curb the rapid rise of housing prices in the three cities [2].

Researchers have made forecasts and analyzed housing prices from different angles. Sonali Das et al. [3] used the large-section dynamic factor model framework of macroeconomic time series to forecast regional housing price inflation, and the results suggest that macroeconomic fundamentals are important in predicting house price inflation. Rangan Gupta et al. [4][5][6] used a 10-variable dynamic structure general equilibrium model to predict the real housing price index in the United States and its decline in the second quarter of 2006. Clapp et al. [7] used the autoregressive process to model the time series behavior of the city's housing price index and forecast individual real estate one quarter in advance. Kieran McQuinn [8] examined the relationship between household capital use costs and price-to-rent ratios in Ireland, using developments in the local Labour market to determine whether these expectations are determined by economic ‘fundamentals.

The paper establishes a housing price forecasting model based on VAR, which no longer starts from the perspective of the macroeconomy, but stands in the perspective of consumers to explore the main factors affecting housing prices and the following works have been done:

(1) Create a data set and preprocessed it that including data cleaning and attribute normalization.

(2) Granger causality test was used to determine the possible influencing factors of housing prices.

(3) VAR model was established according to these factors The model is then used to predict housing prices in three cities (Beijing, Shanghai and Shenzhen).

(4) The main factors affecting the housing price of the three cities were identified by the pulse corresponding and variance decomposition.

2. Method

2.1 VAR

VAR model is a kind of unstructured equation model, which adopts the form of multiple equations and constructs the model by taking each endogenous variable in the system as a function of the lag value of all endogenous variables, thus extending the univariate autoregressive model to an autoregressive model composed of multivariate time series variables. [9]

$$Y_i = C_i + \sum_{i=1}^{p} A_i Y_{I-1} + \epsilon_i \quad t = 1,2,\dots,T$$

The basic form of the model is the autoregressive expression of the weak stationary process [10], which describes several variables within the same sample period as linear functions of their past values. The Y Represents a k-dimensional endogenous variable column vector, C is the constant term, ε is the disturbance term, A is the coefficient matrix of the lag term to be estimated, P is the lag order, and T is the number of samples.

2.2 Granger Causality

The Granger causality test assumes that the information predicted for each variable y and x is fully contained in the time series of these variables, White noise u1t and u2t are assumed to be uncorrelated. For equation (1), its null hypothesis H0: α 1= α 2=... = alpha q = 0 and for Equation (2), its null hypothesis H0: δ1=δ2=... = the delta s = 0.

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{p} \beta_i x_{y-i} + u_{1t} \tag{1}$$

$$x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{i=1}^{p} \beta_i x_{t-i} + u_{2t} \tag{2}$$

Specifically, it is tested by examining whether the coefficients of the sequence lag term are all zero in the vector autoregressive model system [2]. A 2-d P-order stationary vector autoregressive model is taken as an example. To test the null hypothesis y2t is not the Granger cause of y1t, which should joint hypothesis is tested by the F test. If the test results reject the null hypothesis y_{2t} is not Granger reason for y_{1t}.

2.3 Impulse Response and Variance Decomposition

In the VAR model, the relationship between variables in the model can be better analyzed by constructing an impulse response function. L is the hysteresis operator and Θ is the impulse response function [10].

$$y_t = \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + \varepsilon_t$$

$$\left(I_k - \Phi_1 L - \cdots - \Phi_p L^p\right) y_t = \varepsilon_t$$

$$y_t = \left(I_k - \Phi_1 L - \cdots - \Phi_p L^p\right)^{-1} \varepsilon_t$$

$$y_t = \left(I_k + \Theta_1 L + \Theta_2 L^2 + \cdots\right) \varepsilon_t$$

In addition, the influence of different structures on endogenous variables can be obtained through variance decomposition, to rank the importance of these structures.

3. Experimental Results

3.1 Create a Data Set

The data studied in this paper comes from 'Lianjia' (China's largest real estate transaction website). The data includes more than 800,000 transaction records in three major Chinese cities (Beijing, Shanghai and Shenzhen) from 2010 to 2021, which include 14 attributes.

For the data set, we will create new features according to our requirements and conducted a series of preprocessing tasks to get better classification results:

(1) '*transactionTime*' feature is changed to only year base as '*transitionyear*'.

(2) Replacing names with the appropriate sorting code number in *'style'*, *'floor'* and *'distraction'* features.

(3) Create a Feature '*old'* showing the age of a house after construction.

(4) Dealing with missing values.

3.2 Granger Causality Test

After creating and cleaning the original data, set the dependent variable as *'uniprice'*, and the rest are 12 independent variable factors. Next, the Granger causality test is used to determine factors that significantly affect *'uniprice'*, and non-significant variables are removed to play a role in feature screening.

In Table1, it shows the Granger causality test for the three cities. For Beijing, it can be seen from the table that all the other factors except area and orientation are Granger reasons of *'uniprice'*, indicating that the impact of the area on the housing price in Beijing is only significant at the level of 10%, while the housing orientation will not affect the housing price in Beijing. In Shenzhen, only the height of the floor doesn't matter; In Shanghai, all factors will affect the housing price.

It can be seen that the factors affecting the housing price in different regions are different, but the factors that play a role at the same time are *'drawingroom'*, *'listprice'*, *'old'*, *'style'*, '*transactionprice'* and *'transationyear'*. In other words, these factors are universal factors affecting prices.

Table 1. Granger causality test by 3 cities

Cites	Beijing		Shanghai		Shenzhen	
Null Hypothesis	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.
AREA does not Granger Cause UNITPRICE	2.87	0.06	657.10	0.00	5.38	0.00
DIRECTION does not Granger Cause UNITPRICE	1.18	0.31	9.35	0.00	1434.03	0.00
DRAWINGROOM does not Granger Cause UNITPRICE	168.48	0.00	1740.04	0.00	18.15	0.00
FLOOR does not Granger Cause UNITPRICE	230.03	0.00	620.18	0.00	0.21	0.81
LISTEDPRICE does not Granger Cause UNITPRICE	5.44	0.00	429.22	0.00	4.72	0.01
LIVINGROOM does not Granger Cause UNITPRICE	92.43	0.00	3569.08	0.00	2076.35	0.00
AGE does not Granger Cause UNITPRICE	91.75	0.00	133.22	0.00	85.25	0.00
STYLE does not Granger Cause UNITPRICE	17.30	0.00	28.62	0.00	398.53	0.00
SUBWAY does not Granger Cause UNITPRICE	81.48	0.00	547.61	0.00	146.08	0.00
TRANSACTIONPRICE does not Granger Cause UNITPRICE	46.87	0.00	856.84	0.00	10.88	0.00
TRANSACTIONYEAR does not Granger Cause UNITPRICE	8.41	0.00	256.28	0.00	1916.68	0.00

3.3 Parameter Estimation of the VAR Model

In the problem of housing price prediction, housing price is affected by various complex factors, and it forms a system with its influencing factors. To study the interaction, we choose the vector autoregressive model (VAR) to study the dynamic correlation between them [11].

Table 2 shows the estimated values of each regression parameter in the VAR model. For the VAR model, in theory, the test results of individual parameters are not valued, but the overall interactive relationship of the model. However, the impulse response diagram and variance decomposition

diagram can still be drawn according to these coefficients, which analyze the impact of these variables on housing price changes.

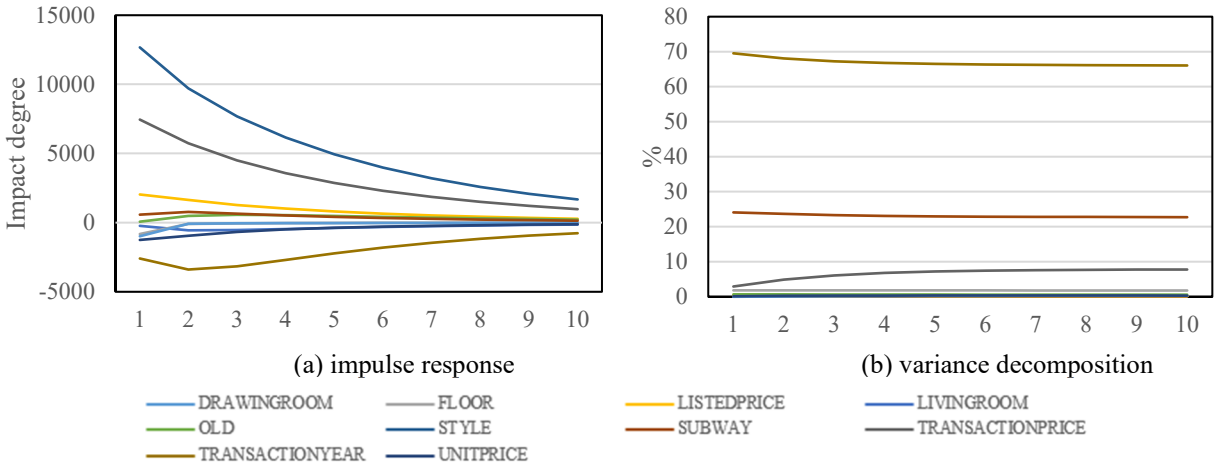
Table 2. Parameter estimation of the VAR model

VARIABLE	coefficient		
	Beijing	Shanghai	Shenzhen
AREA(-1)	--	27.98	36.26832
DIRECTION(-1)	--	-67.15	-1225.71
DRAWINGROOM(-1)	-1222.72	-951.13	-306.70
FLOOR(-1)	451.3766	436.60	--
LISTEDPRICE(-1)	0.004791	1.68	0.62
LIVINGROOM(-1)	554.1092	-139.42	-678.46
AGE(-1)	178.1162	217.33	330.8004
STYLE(-1)	-171.59	-121.69	10.03837
SUBWAY(-1)	1153.670	981.98	-164.51
TRANSACTIONPRICE(-1)	1.841889	-6.28	0.099141
TRANSACTIONYEAR(-1)	-175.84	497.45	1167.191
UNITPRICE(-1)	0.765236	0.72	0.642892
C	363040.9	-996549.10	-2336426.00
R-squared	0.637587	0.60	0.719460
Adj. R-squared	0.637579	0.60	0.719431
S.E. equation	15199.54	9794.22	12855.82
F-statistic	82127.57	21939.39	24426.24
Akaike AIC	22.09594	21.22	21.76110
Schwarz SC	22.09620	21.22	21.76219
Mean dependent	56836.12	45470.29	46335.40
S.D. dependent	25247.82	15442.21	24270.56

3.4 Influence Factor of Housing Price in 3 Cities

3.4.1 Beijing

The influence factors in Beijing can be seen from the impulse response that when '*transactionprice'*, *'transitionyear'* and *'uniprice'* have a one-unit impact, *'uniprice'* changes the most. However, when *'uniprice'* is impacted by other factors, its change is not obvious and it fluctuates around 0.



(a) impulse response (b) variance decomposition

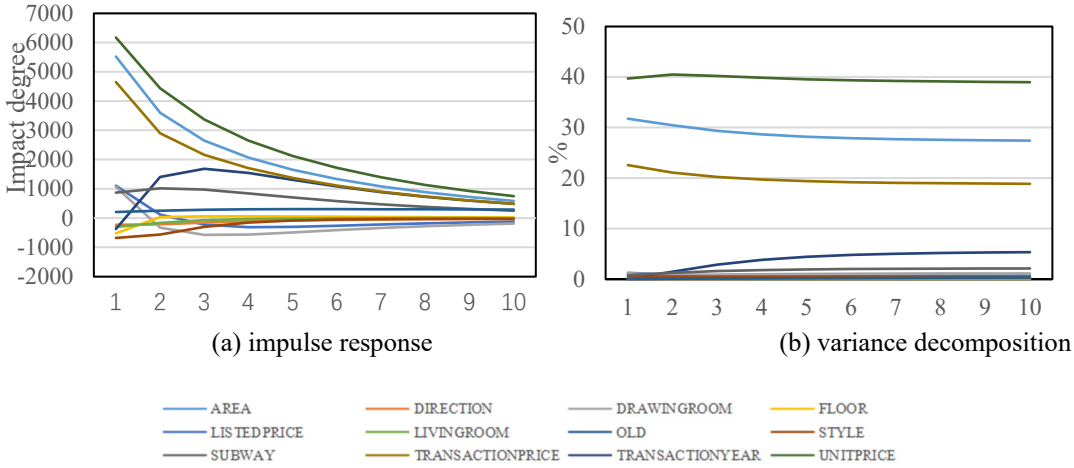
Fig 1. Influence factors of house price in Beijing

As shown in FIG1, the variance decomposition shows the changes of *'uniprice'* are mainly explained by the changes of *'Transactionprice'*, *'uniprice'* and *'transitionyear'*. The remaining variables contribute less to the change in price. Specifically, more than 70% of the variation of price

is explained by itself, while the total transaction price explains 20% of the fluctuation of housing price, and the remaining changes below 10% are explained by time of the trade.

3.4.2 Shanghai

For Shanghai, it can be seen in FIG2 from the pulse decomposition chart that the changes of *'uniprice'*, *'listprice'*, 'old' and *'Transactionprice'* have the greatest impact on the housing price of Shanghai. First of all, it can be found that *'uniprice'* of response to its impact is the largest. In the first period, the impact reaches the maximum and then decreases with time. When *'uniprice'* is impacted by *'listprice'* and '*Transactionprice'*, the overall *'uniprice'* presents a positive response, indicating that *'listprice'* and '*Transactionprice'* have a positive impact on *'uniprice'*. Finally, looking at the influence of *'old'* on the housing price in Shanghai, it can be found that in the first period, the response is negative, and then becomes positive, indicating that the influence of *'old'* on the housing price is non-linear, and the overall trend is positive.

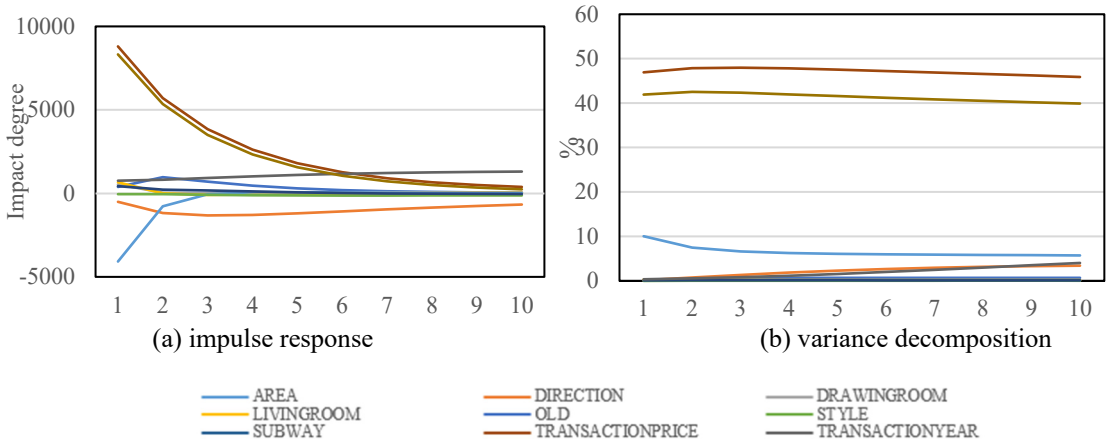


(a) impulse response (b) variance decomposition

Fig 2. Influence factors of house price in Shanghai

The variance decomposition shows that the changes of *'uniprice'* are mainly explained by the changes of *'listprice'*, *'uniprice'*, '*transactionprice'* and *'old'*. The remaining variables contribute less to the change in *'uniprice'*. To be specific, 40% of the variation of *'uniprice'* is explained by itself, while '*transactionprice'* explains 20% of the fluctuation of the housing price, *'listprice'* explains 30% of the fluctuation of the housing price, and the remaining changes below 10% are explained by the housing age. It shows that *'uniprice'* in Shanghai is most affected by its changes, and the influence degree of external factors on *'uniprice'* is *'listprice'*, '*transactionprice'* and *'old'* in order.

3.4.3 Shenzhen



(a) impulse response (b) variance decomposition

Fig 3. Influence factors of house price in Shenzhen

For Shenzhen the impulse response display that when '*transactionprice'*, *'old'* and *'uniprice'* themselves have a one-unit impact on *'uniprice'*, the response range of *'uniprice'* is the largest, and it has a positive impact, indicating that the housing price in Shenzhen is the most influenced by the previous housing price, as shown in FIG3. However, when *'uniprice'* is impacted by one unit of *'old'*, it will have a negative impact, indicating that *'old'* harms the housing price in Shenzhen.

The variance decomposition shows that nearly 40% of the variation of *'uniprice'* is explained by itself, while '*transactionprice'* explains nearly 50% of the fluctuation of housing price, and the remaining 10% or less is explained by the age of the house. This indicates that housing price in Shenzhen is most affected by their changes.

3.5 Predict

Finally, the predicted result of the three cities is compared, as can be seen in table3 that using the Mean Absolute Percentage Error (MAPE) [12], the prediction error of the housing price in Shanghai is the smallest, which is 18%, followed by Beijing and Shenzhen, which may be attributed to the Granger cause of the housing price in Shanghai due to all external factors [13]. The whole forecast effect is good and can provide targeted guidance for the regional housing price forecast.

Table 3. MAPE for 3 cities

City	MAPE
Beijing	0.25
Shanghai	0.18
Shenzhen	0.30

4. Conclusion

In this paper, we analyze the housing data of three major cities in China. The establishment of VAR based housing price forecast model to complete two tasks: (1) to judge the main factors affecting housing prices; (2) to Forecast the future housing price. For different regions, the factors affecting the change in housing prices are different, but there is no denying that housing prices tend to be influenced by prices in previous years, which means that if the government does not intervene in housing prices it can easily lead to a vicious cycle. In the future, we will build a data analysis platform to forecast and analyze the housing price data of other cities. In addition, we will explore more machine learning and deep learning models to get accurate results.

References

[1] Zaichao Du, Lin Zhang. "Home-purchase restriction, property tax and housing price in China: A counterfactual analysis", Journal of Econometrics 188(2015) 558-568. journal homepage: www.elsevier.com/ locate/jeconom.

[2] Alessia De Stefani. House price history, biased expectations, and credit cycles: The role of housing investors. Bloomington: Wiley Subscription Services, Inc Real estate economics, 2021, Vol.49 (4), p. 1238-1266.

[3] Sonali Das, Rangan Gupta, Alain Kabundi. Forecasting regional house price inflation: a comparison between dynamic factor models and vector autoregressive models. Chichester, UK: John Wiley & Sons, Ltd Journal of forecasting, 2011, Vol.30 (2), p.288-302.

[4] Rangan Gupta, Alain Kabundi, Stephen M Miller. Forecasting the US real house price index: Structural and non-structural models with and without fundamentals. Economic modelling, 2011, Vol.28 (4), p.2013-2021.

[5] Rangan Gupta, Stephen M Miller. The Time-Series Properties of House Prices: A Case Study of the Southern California Market. The journal of real estate finance and economics, 2012, Vol.44 (3), p.339-361.

[6] Rangan Gupta, Alain Kabundi, Stephen Miller. Using Large Data Sets to Forecast House Prices: A Case Study of Twenty U.S. States. Journal of Housing Research. Washington Vol. 20, Iss. 2, (2011): 161-190.

[7] John M Clapp, Carmelo Giaccotto. Evaluating house price forecasts. The Journal of Real Estate Research; Clemson Vol. 24, Iss. 1, (Jul/Aug 2002): 1-26.

[8] Kieran McQuinn, Teresa Monteiro. House Price Expectations, Labour Market Developments and the House Price to Rent Ratio: A User Cost of Capital Approach. The Journal of Real Estate Finance and Economics volume 62, pages25–47, 2021.

[9] Christian Pierdzioch, Jan Christoph Rulke, Georg Stadtmann. House Price Forecasts, Forecaster Herding, and the Recent Crisis. Int. J. Financ. Stud. 2013, 1, 16-29; doi:10.3390/ijfs1010016.

[10] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[11] Tim Bollerslev, Patton, Andrew Patton and Wenjing Wang (2016). Daily House Price Indices: Construction, Modeling, and Longer-run Predictions. Journal of applied econometrics (Chichester, England), 2016, Vol.31 (6), p.1005-1025.

[12] Gregory H Bauer. International house price cycles, monetary policy and credit. Journal of international money and finance, 2017, Vol.74, p.88-114.

[13] Christian Pierdzioch, Jan-Christoph Rülke, Georg Stadtmann. Forecasting Changes in House Prices Under Asymmetric Loss: Evidence from the WSJ Forecast Poll. Credit and capital markets (Berlin), 2013, Vol.46 (4), p.495-521.