Analysis and Prediction of Influencing Factors of American Real Estate Price Based on Machine Learning

Tingxi He[1,a,*]

[1]Department of Statistics and Mathematics, Zhejiang Gongshang University, Zhejiang, 310018, China
a. 405953238@qq.com
**corresponding author*

Abstract: The American real estate market has always been a subject of widespread interest, with its price fluctuations impacting not only the lives of millions of American households but also exerting profound effects on the nation's overall economic health. This paper focuses on analyzing which factors are more significant in influencing the American real estate market in order to gain a better understanding of the factors influencing this market and to offer insightful information for stakeholders, policymakers, and investors navigating this dynamic market. We apply linear regression model, lasso regression model and random forest regression model to United States real estate data to explore variables that influence the U.S. housing price and find that PPI, M3, and population are the critical variables impacting the housing price most. By comparing the performance of three regression models, we identify the random forest regression model as the optimal model for predicting real estate prices.

Keywords: house price prediction, real estate, machine learning, random forest regression

1. Introduction

The real estate sector plays a crucial role in a country's progress, serving as a fundamental need for its people and a critical component of its economy. Its progress is essential for economic growth, employment opportunities, and effective urban planning. Given the United States' status as a major global economy, its real estate industry significantly impacts the world economy. As such, it is crucial to monitor its present movements and trends.

In the United States, the residential real estate industry plays a substantial role in the nation's gross domestic product, contributing 13% to the overall economic growth in 2020. The housing market is the dominant force, with consumption outweighing investment. American real estate is highly financialized, with mortgage-backed securities playing a crucial role in the securities market. Real estate represents 23% of residents' assets, with mortgage loans being the primary financing method for purchasing homes. Following the financial crisis, mortgage loans decreased, and real estate assets became more prevalent among the "three low groups (low net worth, low income, low education)." [1].

In general, the US real estate market has long periods of boom and short periods of recession. House prices and house sales are synchronized movements, and both can show apparent periodicity. Regarding the long cycle of real estate sales, the upward stage of real estate in the United States lasts an average of 5.7 years, the downward stage lasts an average of 3.4 years, and the complete cycle is about ten years.

Among the theories of real estate market pricing, the supply and demand equilibrium theory stand as one of the most fundamental, positing that prices are influenced by market supply and demand dynamics. When demand exceeds supply, prices rise, and conversely, prices fall when supply surpasses demand. Another vital theory is the geography theory, which asserts that the location is a primary determinant of real estate prices. The attractiveness of specific locations, infrastructure investments, and the development of surrounding communities can significantly impact property values. Furthermore, the financialization theory highlights the influence of financial products and tools on real estate prices. With continuous innovation in financial markets, the development of real estate derivatives and investment instruments has become a significant factor in price fluctuations. And the investor sentiment theory suggests that market participants' confidence and emotions can influence prices, which underscores the non-rational factors in the market that can cause price fluctuations beyond fundamental factors.

Many scholars have applied different models and empirical methods to study real estate price changes, and the pertinent researches can be categorized into four dimensions: the impact of cost factors, economic factors, social factors, and macro-control factors on housing prices. American scholars studying the impact of cost factors on housing prices mainly focus on the impact of land prices on real estate prices. Glaeser and Gyourko [2] conducted a comprehensive analysis on the relationship between land zoning and house prices. Their findings revealed that land resource allocation had a significant role in influencing fluctuations in housing prices. Gregory and Kisunko's [3] research showed that the rise of land prices was an important factor leading to the rise of real estate prices, and there was a significant positive correlation between land prices and real estate prices.

In studying how economic factors affect housing prices, Baffoe-Bonnie [4] conducted an empirical analysis on the association between five key factors: the unemployment rate, inflation rate, money supply, housing loan interest rate, and residential real estate price. This analysis was performed using the vector auto-regression model. The researcher's findings indicated a notable correlation between the unemployment rate and its influence on the US real estate market. Dennis R. Charlotte et al. [5] conducted a study on 62 metropolitan areas, finding that city size, real income growth, and population growth were the primary factors influencing housing prices. They estimated serial correlation and average regression coefficients based on their data analysis. Min Hwang and John M. Quigley [6] applied the real estate supply and demand model to analyze and concluded that the city's economic situation, the level of household income and the unemployment rate would have a significant impact on the housing price.

In studying the influence of social and macro-control factors on housing prices, according to Mankiw and Weil's [7] research, the "baby boom" generation had a significant impact on changes in housing prices. They developed a model that linked population factors and housing prices, focusing on age as a key factor. The results of their study indicated that population factors were the most crucial drivers of housing price changes in the United States during the 1970s. Dag H and E. Aug's [8] study utilized least squares regression to examine the factors that influence housing prices. They discovered that interest rates, housing construction, unemployment, and household income were the primary factors that explained the rise in housing prices. Among these factors, the interest rate had the most significant effect on housing prices. Otto [9] found that variable mortgage interest rates significantly impacted urban housing price growth rates. Through empirical analysis, Hoyt. Coomes and Biehl [10] found that the imposition of real estate tax significantly impacted housing prices. Ting Zhang Dan Gerlowski [11] studied the real estate market of 20 American cities, and the results showed a correlation between the loan interest rate and the local real estate market trend.

To sum up, real estate prices have been extensively studied by researchers across the globe, resulting in a wealth of research findings. However, these findings vary due to differences in model construction, research focus, and specific influencing indicators. In this paper, we delve into the United States real estate market and apply machine learning techniques to analyze price trends and make predictions. We choose US real estate data from 1998 to 2021, which includes eleven socio-economic indicators, and split it into training and test sets. We employ linear regression, lasso regression, and random forest regression models to train the data in the training set. By utilizing these models with US real estate prices as the dependent variable, we identify the core factors that influence these prices. Finally, we predict future real estate prices using the test set and compared the accuracy of our model predictions to determine the optimal model for analyzing future US real estate market trends.

2. Method

In this research, we analyze and predict the datasets through three models: Linear Regression, Lasso regression, and Random Forest Regression. We first downloaded the US real estate data on the open website -- Kaggle. In this dataset, there are 116 sets of data and 11 socio-economic indicators; one of the features is the target variable. Then, we split the test and train samples; we set the test sample to 86 in the empirical analysis. The three models we used to classify are as follows.

2.1. Linear Regression

Linear regression is a statistical model used to predict how independent variables influence a dependent variable, assuming a linear relationship between them.

The linear regression model is established on the following principles: Linear Assumption: Linear regression assumes that the relationship between the dependent variable (Y) and the independent variables ($X_1, X_2, \dots, X_n$) is linear, meaning it can be described by a straight line. Least Squares Method: The goal of linear regression is to find a straight line that minimizes the sum of the squared vertical distances between the data points on the line and the actual data points. This is known as the least squares method, and it aims to fit the best line.

The linear regression model is typically represented by the following Equation (1):

$$y = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t, t = 1,2,\dots,n(n > k+1) \quad (1)$$

In this formula, x and y denotes the independent variable and dependent variable, k is the number of independent variables, n is the number of groups of observations, β_0 represents the intercept term, meaning the predicted value when all independent variables are zero, $\beta_1, \beta_2, \dots, \beta_k$ represent slope, meaning the impact of a one-unit change in the independent variable on the dependent variable, and ε_t is the error term, representing random errors that cannot be explained by the model.

The parameters β_0 and $\beta_1, \beta_2, \dots, \beta_k$ are estimated using the least squares method to fit the optimal line that best approximates the data. Once the best-fitting line is obtained, the model can be used to make predictions for new data points.

2.2. Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression is a variation of linear regression that incorporates L1 regularization for feature selection and model complexity control. The primary principle of Lasso regression is to minimize the following Equation (2) :

$$LassoLossFunction = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

In this formula, n is the number of observed samples, p is the number of features, y_i and $\hat{y}_i$ represents the actual observed values and predicted values respectively, β_j denotes the regression coefficients, representing the weights of the features, λ is the regularization parameter used to control the strength of regularization. A larger λ leads to more feature coefficients being shrunk to zero or reduced, achieving feature selection.

Lasso regression accomplishes feature selection by adding L1 regularization term $\lambda \sum_{j=1}^{p} |\beta_j|$ to the loss function, causing some feature coefficients to become zero, thus reducing model complexity and improving generalization.

2.3. Random Forest Regression

Random forest regression is a machine learning technique that uses multiple decision trees to perform regression analysis. Its model is established on the following principles: Ensemble of Decision Trees: random forest regression consists of multiple decision trees, typically hundreds or even thousands. Each tree serves as a base learner to fit and predict the data. Injecting Randomness: when constructing each decision tree, random forest introduces two main sources of randomness. Firstly, it uses Bootstrap Sampling to randomly select a subset of data from the original dataset for building each tree. Secondly, at each node, Random forest considers only a subset of features for splitting, rather than using all features. These sources of randomness enhance the model's generalization and reduce the risk of overfitting. Ensemble Voting: once all the decision trees are constructed, Random forest employs a Voting mechanism for regression analysis. Each tree provides a prediction for a given input data point, and the final regression result is the average of all tree predictions.

Given a training dataset, multiple decision trees are constructed using Bootstrap sampling and random feature selection. For each tree, the decision tree regression algorithm is applied to split the data into leaf nodes, and a numeric value is assigned to each leaf node (typically the average of all sample labels in that leaf node). For a new input data point, it is fed into each tree, resulting in predictions from all trees. The final random forest regression result is the average of predictions from all trees, represented as Equation (3):

$$Y_{RF} = \frac{1}{N} \sum\nolimits_{i=1}^{N} Y_i \tag{3}$$

In this formula, Y_{RF} represents the prediction from random forest regression, N is the number of decision trees in the random forest, Y_i denotes the prediction from the i - th decision tree.

Random forest regression excels in its robustness, high generalization capability, and capacity to handle large datasets and numerous features. It is commonly used for regression tasks such as house price prediction and sales forecasting.

3. Result

When analyzing this data set, we first standardized the original data and then used Linear Regression, Lasso Regression, and Random Forest Regression models to analyze and process our samples. The following are the results obtained from our analysis of different models.

3.1. Heat Map--the Correlation between Variables

Heat Map, also known as Hot Spot Map, is a form of representing data changes. By observing the heat map, relationships between data can be discovered, and the similarity and correlation between data can be expressed through gradients of different colors, facilitating further data analysis. In our analysis process, we draw a heat map as shown in figure 1 to represent the degree of correlation between various variables.

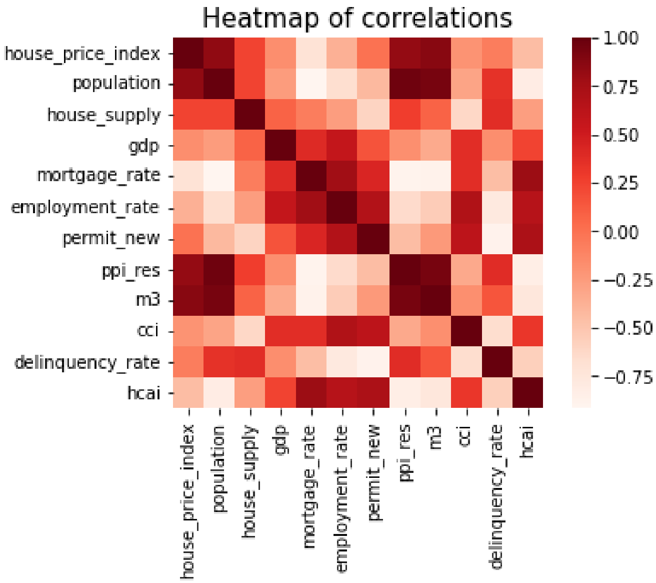

Figure 1: The degree of correlation between different variables

In figure 1, by observing the depth of colors, we can obtain correlations between different variables. As is shown, house price has a very strong positive linear relationship with population, PPI (producer price index) and M3 (money supply). It also has a strong negative linear relationship with mortgage rate.

3.2. Construction of Linear Regression Model

Before establishing the linear regression model, we first draw a scatter plot for different pairwise variables, which can be used to represent the relationship between two variables. The diagonal of this graph as shown in figure 2 displays a bar chart of the data.

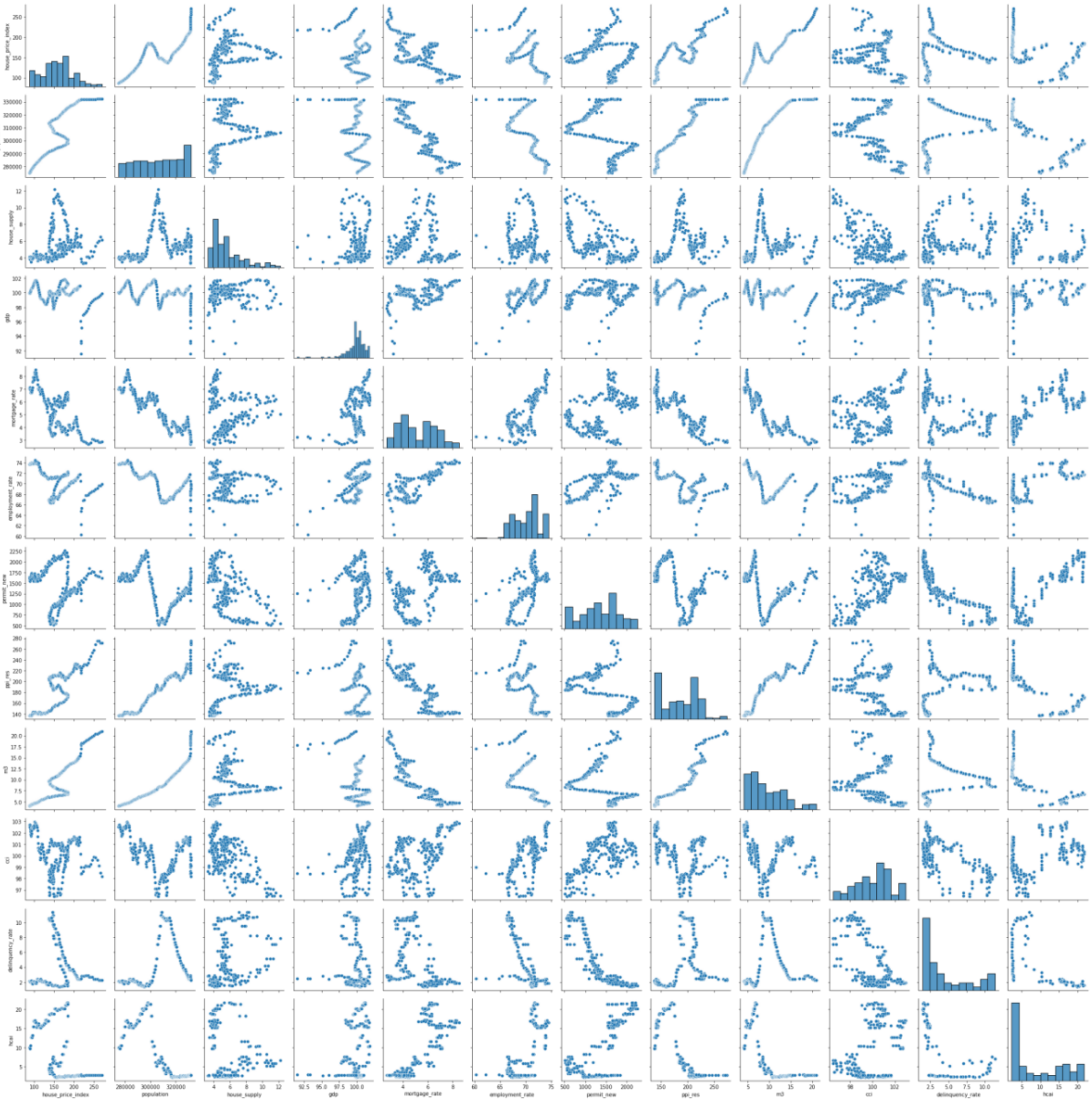

Figure 2: scatter plots between two variables and bar plots of a single variable

By observing the scatter plot, we can discover the correlation between variables. For example, there is a linear relationship between population and m3. The supply of house has no linear relationship with all the variables in this data.

Based on the least square method, we use the training set data to build a multiple linear regression model with house price as the dependent variable. The specific model is shown in Equation (4) below.

$$\begin{aligned} house\_price &= 0.2579*population + 0.2772*house\_supply + 0.1324*gdp \\ &- 0.1059*mortgage\_rate - 0.2349*employment\_rate + 0.1444*permit\_new \\ &+ 0.3119*ppi\_res + 0.3027*m3 - 0.0261*cci - 0.2179*delinquency\_rate + 0.2350*hcai \end{aligned} \tag{4}$$

3.3. Construction of Lasso Regression Model

In analyzing the dataset, we use Lasso regression, find the optimal parameter of 0.001 by using cross-validation, then bring the parameters into the fitting and output the model variable coefficient. Finally, we output the MSE value under this model coefficient to judge the model's goodness. The specific model is shown in Equation (5) below.

$$\begin{aligned} house\_price &= 0.4161 * \text{population} + 0.2213 * \text{house\_supply} + 0.1884 * \text{permit\_new} \\ &+ 0.2922 * \text{ppi\_res} + 0.2259 * \text{m3} - 0.0441 * \text{cci} - 0.1069 * \text{delinquency\_rate} + 0.1937 * \text{hcai} \end{aligned} \tag{5}$$

Since the regression coefficients of variables GDP, mortgage rate, and employment rate are insignificant, they are not included in the model.

3.4. Construction of Random Forest Regression Model

We first fit the random forest regression model. Then, we conduct a statistical analysis of the importance of variables based on this model, which refers to the degree to which features impact the target variable, that is, the importance of features in the model.

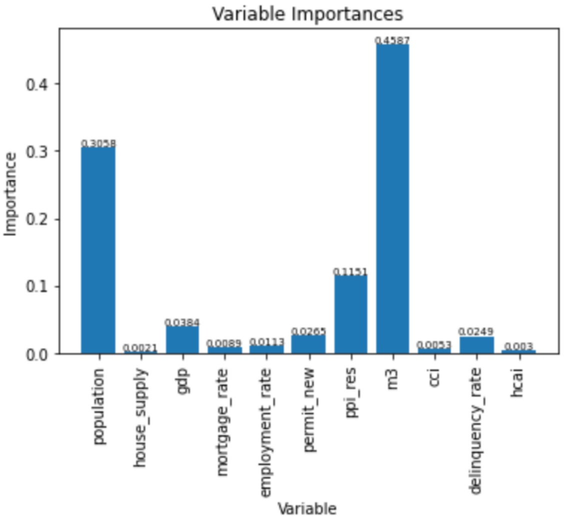

Figure 3: The importance of variables

As figure 3 shows, the variable with the greatest impact is M3 (the supply of money), which indicates that among these features, the variable M3 has the greatest impact on the prediction of results, and its impact on the group of data is as important as about 0.45. In addition to this variable, the feature variable population also has a relatively high degree of influence.

3.5. Comparison of Three Models

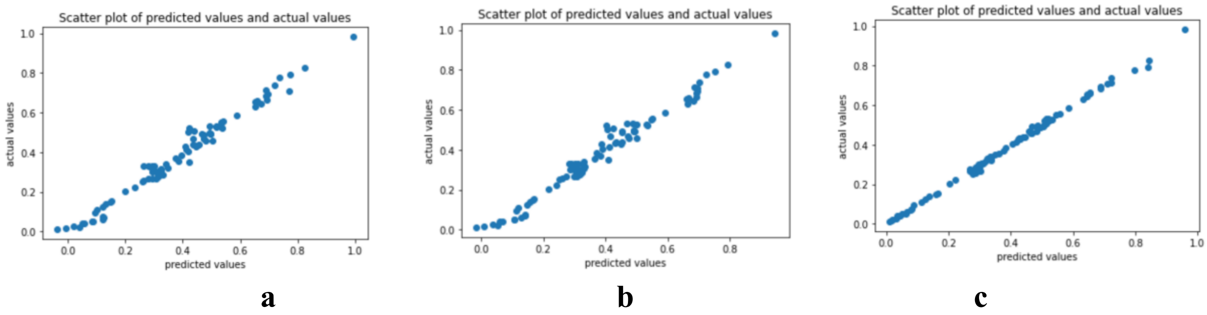

Figure 4: The scatter plots of predicted and actual values in the three models. a: Linear Regression; b: Lasso Regression; c:Random Forest Regression

Through the analysis in the figure 4, it can be seen that the scatter point formed by the actual value and the predicted value of the random forest regression model is closest to a diagonal line, that is, the model has the best fitting and prediction effect.

We compared the scores and MSE values of three models. It is known that higher scores indicate better models, while smaller MSE values also signify better performance. Based on the graph analysis, it is evident that the random forest model performed the best among the three.

Table 1: The score and the MSE of three models

	Linear Regression	Lasso Regression	Random Forest Regression
score	0.97985167	0.973919775	0.9974469550
MSE	0.00094147	0.001218651	0.0001192962

After analyzing the parameters presented in the table 1, it is apparent that Random Forest achieved the highest score of 0.997, whereas Lasso Regression yielded the lowest score. Furthermore, upon examining the MSE values for each model, it is evident that Lasso Regression produced the largest MSE value. Consequently, Lasso Regression is not the ideal candidate for selecting a model from these three. As a result, we strongly suggest taking advantage of the Random Forest Regression model for analysis of house price prediction.

4. Conclusion

This paper conducted a thorough analysis of United State house price data using three different models: linear regression, lasso regression, and random forest regression. Prior to applying these models, we standardized the raw data and utilized visualizations to identify correlations. After careful evaluation, we determined that the random forest regression model produced the most optimal results with the highest model score and smallest MSE. To identify the best parameters for lasso regression, we utilized cross-validation and calculated MSE values for comparison. It is important to note that each model has its own unique set of advantages and disadvantages.

Lasso Regression has the advantage of reducing model complexity through feature selection using L1 regularization. However, selecting the regularization coefficient can be more complex. If the coefficient is too large, the model may underfit, and if it is too small, it cannot limit the model.

Multiple linear regression has several advantages, including ease of use, high accuracy, and widespread use. It is simple to understand, practical for small data volumes with simple relationships, and provides a basis for many powerful nonlinear models. The results are easy to interpret, and can help with decision analysis. They are also useful for solving regression problems. However, it may not be suitable for nonlinear data, and its ability to express highly complex data is limited. Additionally, it may experience low prediction accuracy and over fitting.

Random Forest Regression is capable of determining essential features after training. It utilizes unbiased estimation for generalization error, providing strong model generalization ability with a relatively simple implementation. However, it may not perform as well as in classification when solving regression problems because it does not provide a continuous output.

When examining housing data, the random forest model has proven itself to be highly effective in many areas. We can now utilize the technique of AI and machine learning to predict house prices, analyze social development trends, and assess the direction of housing prices. Although we have focused on macroeconomic and social factors, we have not yet considered market participants' confidence and emotions, or other investor sentiment factors. Additionally, we have not factored in geographical considerations, which can also have a significant impact on the real estate market. Given that the real estate sector is a vital component of the US economy, it is crucial to address the pressing issue of housing supply and predict potential development and sales problems using machine learning methods. The government has already implemented measures to ensure housing market stability and steady social economic development. Moving forward, we must expand our scope to fully reflect the data and comprehensively list the influencing factors in order to improve prediction rates, promote social and economic development, and better meet the housing needs of residents.

References

[1] *http://finance.sina.com.cn/stock/stockzmt/2021-05-02/doc-ikmyaawc3015218.shtml*

[2] *Glaeser Edward L. Gyourko Joseph. The Impact of Zoning on Housing Affordability [J]. NBER Working Paper No. 8835,2002.*

[3] *Gregory Kisunko. The Important of the Context and the Level of Analysis: Authors Response[J]. Housing, Theory and Society,2003(20):134-136.*

[4] *Baffoe-Bonnie J. The dynamic impact of macroecomic aggregates on housing prices and stock of houises: a national and regiona lanalysis [J]. The Journal of Real Estate Finance and Economics ,1998,17(2):179-197.*

[5] *Dennis R. Charlotte Mack. Deterinants of Real House Price Dynamics [J]. NBER Working Paper, No.9262,2002.*

[6] *Min Hwang and Quigley, John M. Economic fundamentals in local housing markets: evidence from U.S. metropolitan regions[J]. Journal of Regional Science, 2006(46): 425- 453.*

[7] *Mankiw.NG, DN Weil. The baby boom, the baby bust and the housing market, Regional Science and Urban Economics,1989,235-258.*

[8] *Dag. H and E. Naug. What drive house price? Economic Bulletin, 2005,Vol. 1: 29-41.*

[9] *Otto Glenn. The growth of house prices in Australian capital cities: what do economic fundamentals explain [J]. The Australian Economic Review,2007,40(3): 225-238*

[10] *Hoyt, William H. et al. “Tax Limits and Housing Markets: Some Evidence at the State Level.” ERN: Urban Economics & Public Policy (Topic) (2011): n. pag.*

[11] *Ting Zhang, Dan Gerlowski. Housing price variability: national and local impacts[J]. Applied Economics,2014,46 (28):3494-3502*