Advances in Social Science, Education and Humanities Research, Volume 496
Proceedings of the 2020 3rd International Conference on Humanities Education and Social Sciences (ICHESS 2020)

A geographic feature integrated multivariate linear regression method for house price prediction

Yuhang Mao[1], Ruili Yao[2]

[1]*Mollyyaola@gmail.com，Highlander technology company*

[2]*yuhangm2@illinois.edu，University of Illinois at Urbana-Champaign*

Abstract
Housing price prediction is of great significance in financial real estate investment and urban construction planning. Multiple linear regression models are commonly used for housing price prediction. However, traditional methods are mostly focused on the characteristics of the houses themselves, without or little considering the features of the surroundings. The features of the surroundings are also important for house price prediction. Motivated by these, we propose a geographic feature integrated multivariate linear regression method for house price prediction. Especially, the Zip Code is chosen as the additional geographic feature for its convenience to obtain. Then the integrated features are used to learn the multivariate linear regression model. We conduct an extensive experiment on the real-world case of the King County area and compare our method linear regressions. The results verified the effectiveness and superiority of our model.
Keywords: *House price prediction, Multiple linear regression models, Geographic feature integrated multivariate linear regression method, Real-world case*

1. Introduction

As a long-standing research content, the housing price is embracing more and more meanings with the development of society. Apart from satisfying the most fundamental housing demands, real estate has also become an important financial product. In some respects, the housing market dynamic is a barometer of the process of urbanization, measuring the comprehensive development and the competitiveness of one region which is instructive for politicians to improve public livelihood and urban planning. Multiple linear regression has been well studied and widely used for housing price prediction[3][8][1]. It reveals linear relationships between house prices and some of their key factors. For example, with other conditions fixed, house price raises with the increment of the size of the house. In addition to the elements belonging to a house itself, characters that relate to its surrounding environment or facilities can also affect its value, which has naturally become a research object for scholars. The hedonic price model proposed by Griliches Z.[4] and Rosen S.[12] is to measure the relationships between house prices and environmental attributes around the asset, such as air quality, water environment quality, noise pollution, transportation facilities, etc. This model is active in many real estate price studies up to now [10][14][7]. Besides, a method called Kriging originated from geostatistics is also applied to house price estimation to integrate more complex and detailed geographic information. Empirical works show in areas like Milwaukee in Wisconsin[9], the metropolitan area of Vienna[6], and Fukushima[2]. In fact, great results achieved by these methods are relying on the thorough consideration of features relating to the surrounding environments that involve in housing valuation. While the acquirement of these features may require detailed investigation or technical support like remote sensing, which is not easy for normal individuals or enterprises.

In this article, we construct a multiple linear regression based method for house price prediction. Different from deeply peeping the effects surrounding factors or geographic information induced to house prices, we look for an agent variable to roughly integrate the information of these external features. Zip Code, for instance, in the empirical analysis on house prices in King County, Washington, is the proxy variable for geographic information. The empirical test corroborates this method can learn the data set well and obtain a high prediction accuracy outperforming linear regressions.

This article is organized as follows: In Section 2, we briefly review the linear regression theory. In Section 3, we introduce our method and show the empirical results of the King County data set. In Section 4, we summarize our work and propose some potential directions for future improvement.

2. Introduction on linear regression

Assume in one data set, we have n observed data $x_{i1}, x_{i2}, \ldots, x_{ip}, y_i (i = 1,2, \ldots, n)$, then the linear regression model can be expressed as:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \epsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \epsilon_2 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \epsilon_n \end{cases} \quad (1)$$

Here, $\beta_0, \beta_1, \ldots, \beta_p$ are parameters, ϵ are random noises. The matrix format of linear regression is:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \quad (2)$$

Here,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (3)$$

When $p \geq 2$, Equation2 is a multivariate linear regression problem. By minimizing the 2nd norm of residuals ϵ, the parameters are solved as:

$$\beta = (X^T X)^{-1} X^T Y \quad (4)$$

3. Empirical Analysis on House Sales in King County

3.1 Proposed approach

Our method can generally be summarized as two folds. First, find a proxy to contain the geographic information possibly related to the housing prices. This proxy is an index for classification to divide the data set into different classes. Then, for each class, we perform regressions to depict the effect of features on house prices. We use the proxy as the classification standard to take into account the different relationships between housing prices and potential features in different regions. For example, housing prices in the city center could be more affected by traffic factors than housing prices in the suburbs, while as to environmental factors, things may be the opposite. The flowchart in figure 1 shows the process of this method being implemented in the empirical analysis of King County. In the experiment, we pick the Zip Code as the proxy variable. Then for house sale records in each Zip Code class, we apply the forward step linear regression via adjusted R^2 as the criterion for feature selection. This method is extensible. The proxy variable can be selected from existing features or can be extracted from geographic information and surrounding elements. Also, the linear regression for each class can be extended to nonlinear forms, which could be explored in further study.

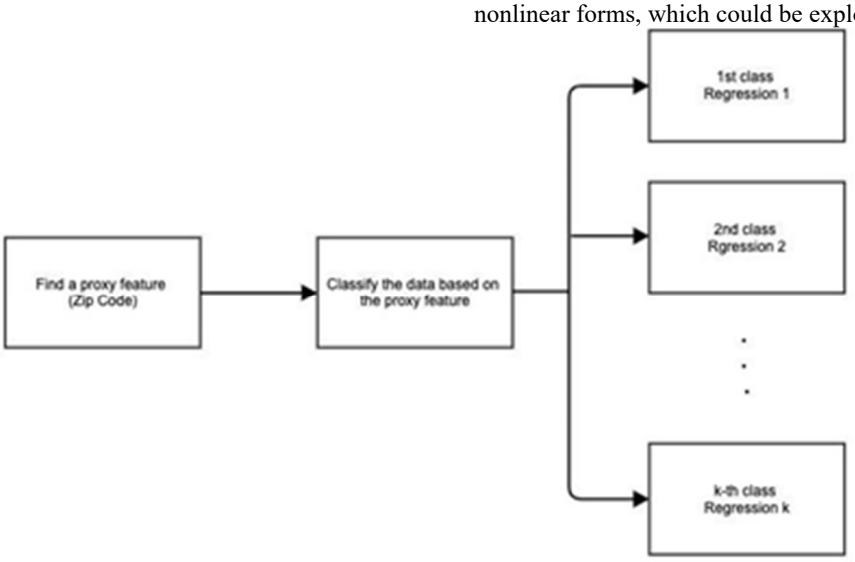

Fig. 1 Method schematic

Next part we show the prediction result of this method, along with the results of linear regressions as a comparison. To avoid the right-skewed distribution of the house prices that violate the requirement of Gauss–Markov theorem[15], in data preprocessing, we took the logarithm of prices to make them close to the normal distribution. Thus, the method in this case is defined as:

$$\ln(price) = \sum_{i=1}^{\#zipcode} 1_i \cdot y_i. \quad (5)$$

i is the number of 'zipcode', y_i is the regression function in region noted by i.

In the case study, R^2 is applied as the criterion to measure the performance of models(formula 6). The data set is split into the training set and test set at a ratio of 7:3. Also, we compare the prediction accuracy of 10-fold cross-validation for each model.

$$\mathbf{R}^2 = 1 \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2} \quad (6)$$

3.2 Case Study

The King County Houses Sales data set has 21613 house sales records between May 2014 to May 2015. It provides prices and some other potentially related features(See table 1).

Table 1 Features and Descriptions

Feature	Meaning
price	price for each house sold
bedrooms	number of bedrooms
bathrooms	number of bathrooms, where 0.5 counts for a room with a toilet but no shower
sqft_living	square footage of the apartments interior living space
sqft_lot	square footage of the land space
floors	number of floors
waterfront	a dummy variable for whether the apartment was overlooking the waterfront
view	An index from 0 to 4 of how good the view of property was
condition	an index from 1 to 5 on the condition of the apartment
grade	an index from 1 to 13 on the grade of building construction and design
sqft_above	square footage of the interior housing space above the ground level
sqft_basement	square footage of the interior housing space below the ground level
yr_built	year the house was built
yr_renovated	year the house was last renovated
zipcode	the zipcode area the house was in
lat	longitude
long	latitude
sqft_living15	square footage of interior living space for the nearest 15 neighbors
sqft_lot15	square footage of the land lots for the nearest 15 neighbors

We trained and tested five models respectively in this data set, which is: linear regression with all features, linear regression with dummy variable 'zipcode', the geographic classified regressions with all features, and the geographic classified regressions simplified by the forward stepwise regression. These two geographic classified regressions are the methods following the structure of formula 2. The results of prediction accuracy are summarized in table 2.

Table 2 R^2 of regression models

Model	Training Set	Test Set	Cross-validation
linear regression with all features	77.1%	77.0%	76.6%
linear regression with dummy variable 'zipcode'	87.7%	87.3%	87.4%
'zipcode' classified regressions with all features	89.1%	90.4%	88.2%
'zipcode' classified forward stepwise regressions	89.2%	89.4%	88.5%

As we can see in table 2, having the dummy variable 'zipcode' brings a huge rise in R^2 when comparing the second linear regression to the first one. This increase corroborates the effects of geographic information on house prices, that with all other conditions fixed, houses located in different regions (have different values of 'zipcode') would have different fundamental valuation. We can also visualize this influence by plotting the distribution of log prices for house sales in different regions (figure 2).

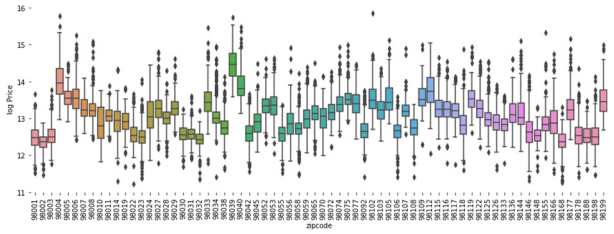

Fig. 2 Boxplot of log prices in different zipcode

However, the second linear regression still dismisses the relations between locations and other features, which belies the fact that housing prices in different regions may have different influencing factors, and the same influencing factor could have a different impact on housing prices in different regions. Therefore, when we apply the geographic information integrated method, that is, regressing house prices in each region individually like the last two models in Table 2, the prediction accuracy could acquire further improvement and can outperform the single linear regressions.

The last model in Table 2 is an improvement based on the third one. Instead of using the same features for each linear regression, we introduce forward stepwise regression to perform feature reduction and eliminate the influence of irrelevant variables on the fitting results. Table 3 counts the number of features of each linear regression in this simplified model. It shows that more than half of the features of almost all the linear equations fitted get pruned (the total number of features is 18). These features are insignificant to house prices and after being eliminated, the cross-validation accuracy of the new model has also improved as shown in Table 2.

Table 3 Frequency of regression equation with specific number of features

Number of features	2	3	4	5	6	7	8	9	10	11	12	13
Frequency of linear equations	1	2	3	10	7	8	6	11	8	9	4	1

Additionally, in Table 4, we count the frequency of features that appeared in the regression equations of this new model. Results corroborate that for real estates in different locations, features that matter to their values could be different. There are 70 different regions divided by 'zipcode' in the King County data set. While the most frequent variables – 'sqft_living' and 'grade' - exist only in 61 of the regression equations. Coefficients of specific features fitted also vary in equations of different regions.

Table 4 Frequency of features

Feature	**Frequency**	**Feature**	**Frequency**
sqft_living	61	bathrooms	32
grade	61	yr_built	31
waterfront	47	sqft_lot	30
condition	46	yr_renovated	29
sqft_living15	44	sqft_basement	24
view	42	floors	21
sqft_above	36	bedrooms	18
sqft_lot15	33		

Overall, the empirical analysis of the King County data set shows this method outperforms linear regressions for house price prediction. 'zipcode' in this case, is a good proxy for geographically related information. After classifying data rely on 'zipcode', regressions of different regions can reveal the relationships of features to house prices more closely to the reality compared to a single linear model.

4. Conclusion

This article proposes a geographic feature integrated linear regression method to solve the problem of predicting real estate prices. The framework of this method can easily be extended by adjusting the proxy variable for geographic information as well as applying other forms of regression. In this article, we compared the prediction accuracy among four regressions in the King County house sales data set, two of them are linear regressions, the other two are built followed by this method. Results show our method can outperform the linear regressions with higher accuracy on both test set and 10-fold cross-validation. This method can also reflect the important influence of location on housing prices: not only in the differences of important features affecting house prices in different areas but also on the different effects of the same variable on house prices in different regions.

For future studies, we may excavate nonlinear relationships by extending the linear regressions in this method. Since now there are various researches on studying housing prices via machine learning[5][13][11], efforts can be made to incorporate these machine learning algorithms into this framework. Moreover, 'zipcode' is proved as a great proxy in King County's case, while things may not be the same for other data set. Hence, to find a method that could generate a qualified index to integrate geographic information should be another direction to explore.

References

[1] P Boye, D Mireku-Gyimah, and CA Okpoti. Multiple linear regression model for estimating the price of a housing unit. *Ghana Mining Journal*, 17(2):66–77, 2017.

[2] Ahmed Derdouri and Yuji Murayama. A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across fukushima prefecture, japan. *Journal of Geographical Sciences*, 30:794–822, 2020.

[3] Nehal N Ghosalkar and Sudhir N Dhage. Real estate value prediction using linear regression. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–5. IEEE, 2018.

[4] Zvi Griliches. *Price indexes and quality change: Studies in new methods of measurement*. Harvard University Press, 1971.

[5] Yuhao Kang, Fan Zhang, Wenzhe Peng, Song Gao, Jinmeng Rao, Fabio Duarte, and Carlo Ratti. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, page 104919, 2020.

[6] Michael Kuntz and Marco Helbich. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *International Journal of Geographical Information Science*, 28(9):1904–1921, 2014.

[7] Tian Liu, Weiyan Hu, Yan Song, and Anlu Zhang. Exploring spillover effects of ecological lands: A spatial multilevel hedonic price model of the housing market in wuhan, china. *Ecological Economics*, 170:106568, 2020.

[8] Tiantian Li and Pengyan Li. Analysis on real estate price influence factors in china based on the multivariate linear regression model. In *ICCREM 2016: BIM Application and Off-Site Construction*, pages 998–1004. American Society of Civil Engineers Reston, VA, 2017.

[9] Jun Luo and Yehua Dennis Wei. A geostatistical modeling of urban land values in milwaukee, wisconsin. *Geographic Information Sciences*, 10(1):49–57, 2004.

[10] Yingdan Mei, Diane Hite, and Brent Sohngen. Demand for urban tree cover: A two-stage hedonic price analysis in california. *Forest Policy and Economics*, 83:29–35, 2017.

[11] Zhen Peng, Qiang Huang, and Yincheng Han. Model research on forecast of second-hand house price in chengdu based on xgboost algorithm. In *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)*, pages 168–172. IEEE, 2019.

[12] Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55, 1974.

[13] Feng Wang, Yang Zou, Haoyu Zhang, and Haodong Shi. House price prediction approach based on deep learning and arima model. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pages 303–307. IEEE, 2019.

[14] Yonglin Zhang and Rencai Dong. Impacts of street-visible greenery on housing prices: Evidence from a hedonic price model and a massive street view image dataset in beijing. *ISPRS International Journal of Geo-Information*, 7(3):104, 2018.

[15] *Gauss–Markov Theorem*, pages 217–218. Springer New York, New York, NY, 2008.