Spatial Dependency and Hedonic Housing Regression Model

Conference Paper · December 2016

DOI: 10.1109/ICMLA.2016.0097

2 authors:

Timothy Oladunni
University of the District of Columbia
23 PUBLICATIONS **106** CITATIONS
SEE PROFILE

Sharad Sharma
Bowie State University
89 PUBLICATIONS **691** CITATIONS
SEE PROFILE

2016 15th IEEE International Conference on Machine Learning and Applications

Spatial Dependency and Hedonic Housing Regression Model

Timothy Oladunni
Department of Computer Science
Bowie state University
Bowie, Maryland, 20715, USA
oladunnit0423@students.bowiestate.edu

Sharad Sharma
Department of Computer Science
Bowie state University
Bowie, Maryland, 20715, USA
ssharma@bowiestate.edu

Abstract

The location of a real estate property has a considerable impact on its appraised value. Accounting for geographical information eliminates some reducible errors in the accuracy of a hedonic housing regression model. An improved performance will benefit home buyers, sellers, government and real estate professionals. This paper investigates the spatial dependency and substitutability of submarket and geospatial attributes in a hedonic housing regression model using mutual information (MI) and variance inflation factor (VIF). Best subset linear regression and regression tree predictive models were built as learning algorithms. Bayesian Information Criterion (BIC) and Residual Mean Deviance (RDM) measured the performance of the linear regression and regression trees respectively. The BIC of the linear regression model indicated a best fit at 14 and 11 variables for submarket and geospatial models respectively. Optimization of the submarket tree was attained with 9 parameters comprising of 15 terminal nodes, while 7 parameters comprising of 13 terminal nodes achieved optimization in the geospatial tree. While geospatial models have a slight edge over the submarket model, the experiment suggested the substitutability of the models. The dataset consisted of single family's homes in 8 counties between January and December 2006 extracted from the Multiple Listing Service repository.

Keywords: real estate prediction, decision tree, best subset linear regression, housing prices prediction, variance inflation factor, mutual information, hedonic theory.

1. INTRODUCTION

The housing bubble and burst experienced recently in the United States and some other parts of the developed worlds affected the government, investors and homeowners. Government budgets that were dependent on property taxes took a hit. Investment at different levels of real estate transactions were cut unawares. The most affected groups were the homeowners who saw the equity of their homes disappeared. (Equity of a home is the difference between the mortgage and the value of a property). Thus, the stock market sent a ripple effect that reverberated to other sectors of the economy creating panic and economic insecurity. While there has been much soul searching into the different reasons as to why the world was caught unaware about the bubble and burst which was blamed primarily on sub-prime mortgages, there has not been adequate research into the role of the geographical locations of the properties. In most jurisdictions, the estimated value of properties are provided by appraisals. Due to manipulations by some major players in the real estate world, appraisers often provide the wrong estimates.

The question here is; *does the geographical location have any effect on the estimated value of properties*? In other words, if the same builder constructs the same type of properties in two different geographical locations with same cost, are the values the same? If the answer is no, then *how best do we account for the geographical location in predicting the value of the properties?* Government, investors and homeowners should have a satisfactory answer to this question. Such an answer will help the government to determine how best to project the expected income from properties as it affects their budget. Knowledge about how geographical location affects property values will help banks and mortgagees in red-tapping vulnerable areas. Lastly, homeowners will know the implications of the choice of location on the values of their homes.

We will discuss related works about real estate prediction and estimates in section two, and our experiment will be discussed in section three. In section four, we will discuss the result and analysis of our experiment. Chapter five will be about our future work. We will conclude the paper in section six.

2. LITERATURE REVIEW

2.1 Related Work

The real estate market has been a crucial factor in the economy of many developed countries. Inadequate analysis of values of properties have been blamed for real estate crisis in different countries. Therefore, economists, urban and regional town planners, geographers, statisticians and computer scientists have done some tremendous jobs on the real estate pricing and estimates.

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.161

Sotirios *et.al*, argued that time as well as location have direct impact on the pricing of real estate properties. According to the study, omission of time in forecasting model contributed to the bubbling and bursting of housing market [1]. Economic and computer science researchers in NYU investigated the error pattern of housing pricing in real estate transactions. Using dataset from L.A, they demonstrated that error in turn over time is associated with under-estimating geographical location factors. The study further suggested that in forecasting properties values over a period of time, high initially priced properties have been over estimated, while others have been under estimated. According to the researchers, real estate value prediction has a ripple effect on mortgage defaults [2]. Other research has demonstrated spatial autocorrelation of housing markets using Moran's I statistics. Spatial data were used to establish the interwoven relationship between housing price, income, population, consumer pricing and urbanization [3]. Yang *et.al*, investigated the relationship between the housing price and relative housing price. Using datasets from Chinese urban housing market, the study analyzed the spatial heterogeneity, differentiation pattern, and overall trend of housing market in the urban areas of China [4]. Using time-varying model, Lei and Xueqin investigated the values and burbles of housing market in China [5]. Also in China, some researchers investigated the spatial variation in housing prices in Xiamen Island, China. The study shows the uneven spatial distribution of housing prices in the investigated area. They argued that factors such as public service facility allocation and urban development history have a direct effect on housing prices [6]. Another study shows that understanding the economic spatial structure of China will guide the government in the development of real estate properties [7].

Xiaolu *et al*., investigated the relationship between spatial pattern and spatial features of housing prices. According to the study, accounting for this relationship increases the accuracy of hedonic regression models. Global regression residual and geographically weighted regression models were analyzed. The outcome of the study shows that both models are complimentary [8]. Tom and Thomas investigated the predictability of real estate pricing using rent-price ratio as explanatory parameters. Findings were based on datasets from 18 OECD countries [9]. Some researchers studied the spatial effects of liquidity factors on the performance of real estate market. Their findings showed that there is a correlation between housing liquidity level and expected returns on their values [10]. Zhong *et al.,* proposed a SVM model for predicting housing values and compared the outcome with a neural network model [11]. Using hedonic approach, Eva and Martin argued that a combination of geostatistical and disaggregated submarket variables is the best predictive model of real estate prices [12]. Steven *et al.,* compared the performance of a submarket-OLS with geostatistical predictive variables. The experiment showed that submarket model performed better than the geostatistical model [13]. Timothy and Sharma developed a real estate predictive application software using linear regression and model view controller architecture. According to the study, stakeholders in the real estate world can perform real estate transaction and verify the estimated values of properties [14].

2.2 Hedonic Regression

The hedonic theory has been widely used in different areas of demand and supply of commodities. It was introduced into real estate market in 1968. The American Real Estate Price Index is based on this theory [15]. According to hedonic theory of demand, the price of a commodity is a function of its attributes. In other word, if a composite commodity can be decomposed to its heterogeneous features, then we can estimate its market value. Wen *et al.,* in a study argued that, just like any other heterogeneous goods, the features of a real estate property determine its utility [16]. Therefore, using the hedonic theory, the price (response variable) of a real estate property can be computed from its features (explanatory variables).

2.3 Linear Regression

Given an explanatory variable x and a response variable y, a simple regression line regresses y on x. In other words, for every value of x, there is a correspondent value of y, provided y and x are linearly related [17]. If x has more than one variable, then it is considered as multi-variate. Mathematically, we represent a multiple regression equation as;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon. \quad (1)$$

Where; Y is the response parameter, β_0 the bias known as the intercept term, $\beta_1 \ldots \beta_p$, the coefficient parameters, $X_1 \ldots X_p$, the explanatory parameters and ε the error of prediction.

2.4 Regression Tree

Given a response variable Y and inputs $X1$ and $X2$. Using binary tree, we can recursively map the value of Y with each input $X1$ and $X2$ taking values at unit intervals. Splitting starts at the root node, with conditions of either to assign observations to either the left or right side of the tree. The process continues on each node until a predetermined condition is realized. The terminal nodes are referred to as the leaf. In general, given two regions R_1 and R_2 with splitting variable j and splitting point s, splitting into regions can be achieved using the following criteria;

$$R_1(j, s) = \{X|X_j \leq s\} \text{ and } R_2(j, s) = \{X|X_j > s\}. \quad (2)$$

While decision trees may not produce the best model for inferential purposes, it has the interpretability advantage [18].

3. EXPERIMENT

For this experiment, four thousand datasets of single family properties sold on the Multiple Listing Service (MLS) between January and December 2006 were extracted and processed. Only properties with sold status were used for the experiment. The status of sold is an indication that their values were determined by licensed appraisers. Appraisers give professional advice based on established government regulations. Datasets were considered complete and accurate because the MLS is the official repository of real estate transactions in a metropolitan area. In most cases, each metropolitan area has its own MLS. The Washington DC metropolitan area of the United States was considered for the study [19]. Eight counties were considered; each county was defined as a submarket boundary. Thirty-nine parameters were investigated. We pre-processed the datasets to remove extraneous information, detect outliers and missing rows.

The question here is; *is there a statistical relationship between the geographical location and the value of properties?* Visualization of the spread of the distribution was done using Rapid Miner - a data analytical tool. Using some mechanism capable of sniffing through the large piles of datasets, our analytical tool provides a means of visually observing some useful and vital patterns. Figure 1 shows a bar chart comparing two categorical variables using the heights of a bar. It provides information about the absolute and relative variables. [20]. The figure shows the bar chart of the average price of properties in each of the counties.

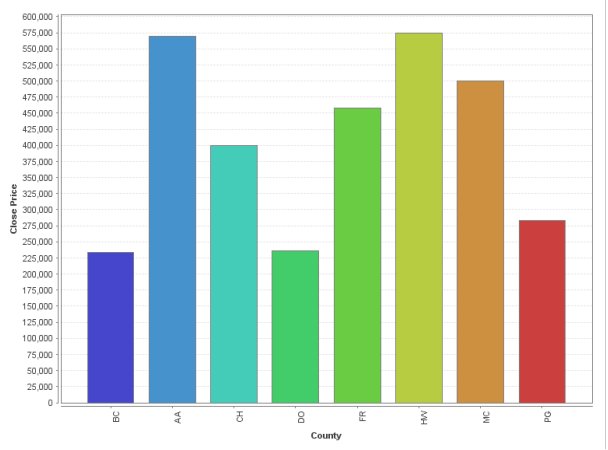

Figure 1. A bar chart of the median values of properties in Baltimore, Anne Arundel, Charles, Dorchester, Fredrick, Howard, Montgomery and Prince Georges counties. Each bar represents a county.

The figure shows that Baltimore and Dorchester counties have the lowest property values of approximately $225,000.00, while Anne Arundel and Howard counties have the most expensive properties with an approximate mean value of $575,000.00. The figure suggests that while two properties may have the same structural cost, the geographical location is crucial to their values. In other words, keeping all other factors constant, the construction cost of same style of properties may be the same in Baltimore and Howard counties, however, their values may not be same. Intuitively, the figure suggests that Baltimore and Dorchester counties may have more low income earners than Howard and Annie Arundel counties.

We further examined the datasets using a scatter plot. Figure 2 shows a scatter plot of Closing Price vs Latitude. With the exeption of a few outliers, the figure suggests that there is a positive correlation bewteen 30.2º and 39.0º vs prices, while there is a negative correlation between 39.0º and 39.7º vs prices. A postive relationship suggets that prices and latitude rises and falls in the same trajectory while negative relationship shows that values of propeties decreases with an increase in the latitude.

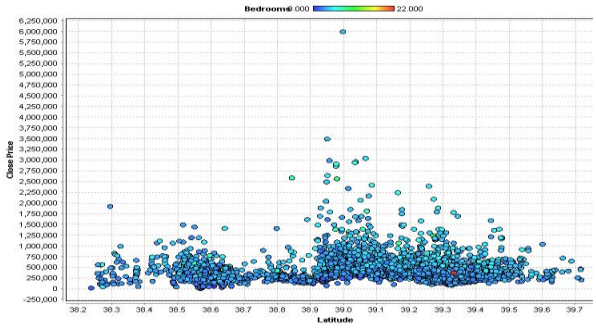

Figure 2. A scatter plot of Close Price vs Latitude in the eight counties.

3.1 Information Redundancy

We hyopthezised that the presecence of *both geospatial and submarket parameters in a dataset do not constitute redundancy.* Two parameters are said to be colinear or redundant if one is a substitute of the other. This implies that they provide the same information in the dataset. If there are more than two varibles, then we say that there is multi-colinearity. We investigated colinearity by conducting a mutual information (MI) and variance inflation factor (VIF) tests.

I. Mutual Information

MI matrix has been widely used in measuring the strength of statistical relationship between two or more variables. The most popular choice of statistical relationship is the correlation matrix, however it beocmes unreliable when measuring non-linear relationship. MI on the other hand has the capability of measuring non-linear relationships [21]. It is considered a *Kullback-Leibler divergence* case. Given two variables *X* and *Y,* the mutual statistical relationship between the two variables is given mathematically as:

$$I_{xy} = E\{log \frac{p(X,Y)}{px(X)py(Y)}\}. \quad (3)$$

We computed the MI of submarket boundaries and the geospatial parameters of our datasets with Table 1 showing

the results. The parameters were normalized prior to the computation. Zero MI means that the two variables are independent, while an MI of more than one is considered evidence against the hypothesis (the two variables are predictive of each other).

Table 1. Mutual Information Table

	Longitude	Latitude
County	1.709	1.704

The MI table shows that there is statistical evidence that there is colinearity between submarket boundaries and geospatial parameters.

II. Variance Inflation Factor

We further examined colinearity between the two parameters using Variance Inflation Factor (VIF). VIF measures the degree of 'largeness' of the variance of a coeffiecient considering its collinearity. A VIF of one means that the variable cannot be influenced by other variable in the predictive model (abscense of collinearity) [22]. We computed the VIF of the variables using the following Mathematical relationship:

$$VIF(\beta_j) = \frac{1}{1 - R^2_{X_j | X_{-j}}} . \qquad \text{equation (4)}$$

From the equation, $R^2_{X_j | X_{-j}}$ is the R-square of one variable to other variable in the predictive model, while β_j is the coefficient value of the parameter. A VIF that is more than 5 is considered an evidence against the hypothesis. Table 2 shows the variance inflation factor table.

Table 2. Variance Inflation Factor Table

Variables	DF	VIF	VIF minus Submarket	VIF minus Geospatial
County	7	454.43172	N/A	3.387471
Latitude	1	16.072175	1.542319	N/A
Longitude	1	17.049822	1.507381	N/A

The degree of freedom of county is seven because there are eight possible values between 0 and 7 (eight counties). Based on our hypothesis, County, Latitude and Longitude are considered collinear with values of 454.431722, 16.0721175 and 17.049822 respectively. This suggests that they are measuring the same feature - geographical location of the property under consideration. Thus, there exists a strong evidence of multi-collinearity between submarket boundary and geospatial parameters. While removing each of the parameters with a high VIF one at a time, we re-calculated the VIF. The result is shown in fourth and fifth columns

3.2 Learning Algorithms

We have demonstrated via bar-chart in figure1 the relationship between the mean price of properties and submarket boundaries. Figure 2 demonstrated relationship between the price of a property and its geospatial variables. Also, we demonstrated that submarket boundaries and geospatial parameters exhibited a collinear relationship when combined in the same model. Thus, we separated the two parameters, built two models and compared their performances. The two models considered the same number of datasets (four thousand) and same number of explanatory parameters (bedrooms, baths. full, baths. half, levels, fireplaces, basement, lot. sqft, year.built, DOMM, DOMP, Seller.Subsidy and geographical location). With the same number of datasets and parameters, the difference between the two models was the geographical information. While the geographical information of the geospatial model was based on geospatial parameters (longitude and latitude), the submarket model was based on submarket boundaries (counties). Linear regression and regression trees were used to implement the hedonic pricing model. Thus, our response variable was the price, while other parameters were the predictors. R was used for computation and analysis. A comparison analysis was made on the performances of the models.

I. Linear Regression Model

We implemented the hedonic housing theory using linear regression to build submarket and geospatial models. A linear regression model has the capability of predicting the value of a response variable, given an explanatory variable. There are different types of linear regression models based on different applications. The *best subset selectio*n (BSS) multiple linear regression model was chosen as our learning algorithm. A BSS model has the advantage of improving the accuracy of an ordinary least square while preserving its inferentially and interpretability. It also considers all possible combinations of the explanatory parameters. Cross validation was used to obtain the best fit. BIC (Bayesian information criterion) measured the performance of the model. Using a likelihood criterion, the BIC is penalized by the complexity of the model [23].

II. Regression Tree Model

We further implemented the hedonic housing regression model using regression tree to build geospatial and submarket models. *Recursive binary splitting* was used to grow a decision tree. This is because it is computationally infeasible to consider every possible partition into the feature space. The most important parameter in determining the price of properties were chosen as the root. The decision is *greedy* because splitting decision was based on the particular node under consideration rather than looking ahead. Furthermore, splitting decision began at the 'top' of the tree which counter intuitively is the root. The splitting continues recursively until optimization was achieved on

the maximum amount of parameters necessary for our prediction. Residual mean deviance was used to for performance evaluation.

4. RESULT AND ANALYSIS

We analyzed the result of each model and discussed its implication.

I. Linear Regression Outcome

Performance evaluation was done using the Bayesian information criterion (BIC). The figures below show the graphical representation of the best subset selection linear regression model. Figure 3a shows the performance of the model when submarket boundary was considered as the geographical parameters.

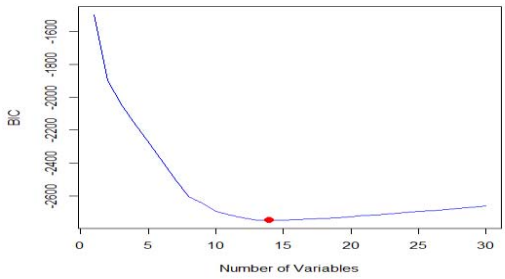

Figure 3a Graph of BIC vs Number of Variables for the Submarket model

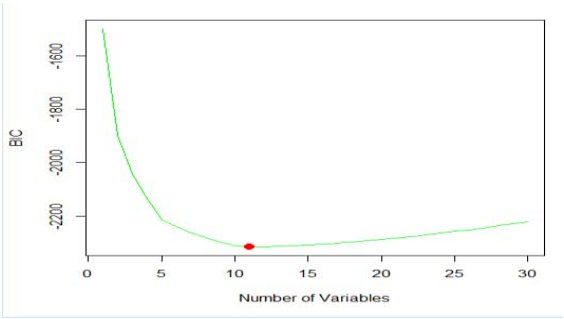

Figure 3b. Graph of BIC vs Number of Variables for the Geospatial Model

Figure 3b shows the performance of the model when geospatial attributes were considered as the geographical parameter. As shown in figures 3a and 3b, submarket model optimized with 14 variables, while the geospatial model optimized with 11 variables.

I. Regression Tree Outcome

Performance measurement was done using residual mean deviance. Optimization of the submarket tree was attained with 9 parameters comprising of 15 terminal nodes, while 7 parameters comprising of 13 terminal nodes achieved optimization in the geospatial tree. Residual mean deviance of 2.562e+10 and 2.451e+10 for the submarket and geospatial trees respectively. Complexity pruning was done at level 5, for easy comparison of the performances of the models. Figure 4a shows the submarket model, while figure 4b shows the geospatial model. Both trees consist of a series of binary splitting starting at the root.

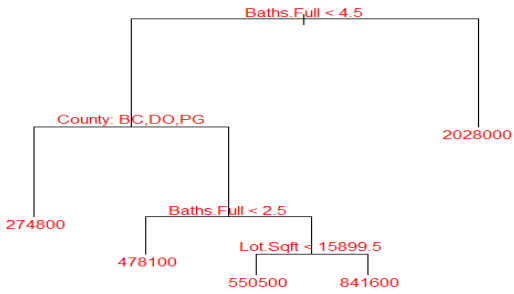

Figure 4a. Submarket Decision Tree Model

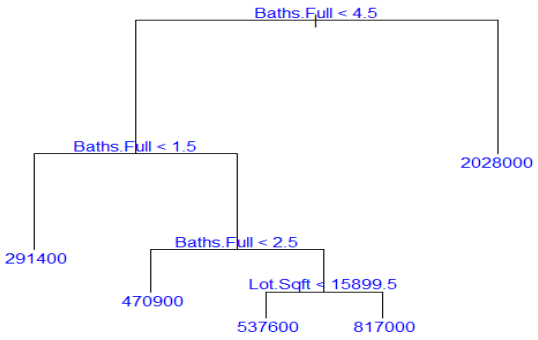

Figure 4b. Geospatial Decision Tree Model

Interestingly, both models identified Full Baths as the root, indicating that the number of full baths in each observation is the most important parameter in determining the value of a property. The root assigns properties having full baths less than 4.5 to the left while others were assigned to the right. For properties having baths greater than 4.5, the predicted price is $2028000. Properties with less than 4.5 full baths in the submarket model were considered based on their counties. While properties in the geospatial model with less than 4.5 were further split to properties with less than or greater than 1.5 full baths. Properties with less than 1.5 full baths were predicted with a value of $291400.

5. FUTURE WORK

The experiment was based on single families, our future shall include condominium and townhomes where factors such as condominium and homeowners' association fees may be a factor. Also, factors such as government regulations, frauds, and natural disasters may have an impact on the algorithm.

6. CONCLUSION

The study examined the spatial dependency and substitutability of geospatial and submarket boundary parameters in a real estate hedonic model. Datasets of four thousand properties listed and sold in 2006 in eight counties were processed and analyzed. Our study shows that:

a. The geographical location of a property has a considerable impact on its price.
b. Geospatial model has a better performance than submarket model.
c. There is collinearity between submarkets boundaries and geospatial parameters.
d. A hedonic regression model can use either geospatial or submarket parameters as the geographical information in estimating the value of a property

7. ACKNOWLEDGEMENTS

This work is funded in part by the National Science Foundation grant number HRD-1238784.

REFERENCES

[1] Jean Dubie and Diego Legros Sotirios Thanosa, "Putting time into space: the temporal coherence of spatial applications in the housing market," *Regional Science and Urban Economics*, vol. 58, pp. 78–88, 2016.

[2] Sumit Chopra, John Leahy, Yann LeCun, and Trivikrmaman Thampy Andrew Caplin. (2008, December) Machine Learning and the Spatial Structure of House Prices. [Online]. HYPERLINK http://yann.lecun.com/exdb/publis/pdf/caplin-ssrn-08.pdf

[3] Hu Zhaohui, "Spatial Econometric Analysis of Housing Price of Chinese Provinces," in *2012 Second International Conference on Business Computing and Global Informatization, Shanghai*, Shanghai, 2012.

[4] D. Wang and Y. Wei Y. Wang, "Spatial differentiation patterns of housing price and housing price-to-income ratio in China's cities," in *21st International Conference on Geoinformatics*, Kaifeng, 2013.

[5] L. Feng and X. Tan, "Specification of Housing Bubbles Based on Time-Varying Present Value Model: A Case of China," in *) International Conference on Management and Service Science (MASS)*, Wuhan, 2011

[6] Lang He and Junfeng Jiao Yuan Li, "A spatial analysis of housing prices in Chinese coastal cities, a case study of the city of Xiamen, China," in *6th International Association fo China Planning Conference (IACP)*, Wuhan, 2012.

[7] W. Lijuan and L. Guiwen, "Spatial Variation Analysis of the Housing Price in Multi-center City: A Case Study in Chongqing City, China," in *Fifth International Conference on Computational and Information Sciences (ICCIS)*, Shiyang, 2013, pp. 450-453.

[8] X. Gao and Y. Asami, "Influence of spatial features on land and housing prices," *Tsinghua Science and Technolog*, vol. 10, no. 3, pp. 344-353, 2005.

[9] Tom Engsted and Thomas Q. Pedersen, "Predicting returns and rent grows in the housing markets using rent-price ratio: evidence from the OECD ocuntries," *Journal of International Money and Finance*, vol. 257-275, p. 53, 2015.

[10] E.C.M. Hui Xian Zheng, "Does liquidity affect housing market performance? An empirical study with spatial panel approach," *Land Use Policy*, vol. 56, pp. 189-196, 2016.

[11] Z. Chunguang, H. Lan, W. Yan and Y. Bin Z. Yi, "Support Vector Regression for Prediction of Housing Values," in *International Conference on Computational Intelligence and Security*, Beijing, 2009.

[12] Eva Cantoni and Martin Hoesli, "Predicting House Prices with Spatial Dependence: A Comparison of Alternate Methods," in *15th Conference of the Pacific Rim Real Estate Society (PRRES),* Sydney, 2009.

[13] Eva Cantoni and Martin Hoesli Steven C. Bourassa, "Spatial Dependence, Housing Submarkets, and House Prices," *International Center for Financial Asset Management and Engineering*, no. 151, 2005.

[14] Timothy Oladunni and Sharad Sharma, "Predictive Real Estate Multiple Listing System Using MVC Architecture and Linear Regression," in *ISCA 24th International Conference on Software Engineering and Data Engineering*, San Diego, California, 2015.

[15] LIU Yan and WU Yong-xiang, "Analysis of Residential Product's Value," in *International Conference on Management Science and Engineering*, Moscow, 2009.

[16] J. F. Lu and L. Lin H. Z. Wen, "An improved method of real estate evaluation based on Hedonic price model," in *IEEE International Engineering Management Conference*, 2004.

[17] Daniela Witten, Trevor Hastie and Robert Tibshirani Gareth James, *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 2015

[18] Robert Tibshirani and Jerome Friedman Trevor Hastie, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. California: Springler, 2008.

[19] Metropolitian Regional Information Services. MRIS. [Online]. HYPERLINK "http://www.mris.com/" http://www.mris.com/ .

[20] Andy Kirk, *Data Visualization:a successful design process*. Birmingham: Packt Publishing, 2012.

[21] Roberto Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537-550, 1994.

[22] T. Fischetti, Data Analysis with R, Birmingham: Packt Publishing, 2015.

[23] Scott Shaobing Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, 1998.