2016 15th IEEE International Conference on Machine Learning and Applications

Hedonic Housing Theory – A Machine Learning Investigation

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 hedonic pricing theory suggests that house is a differentiated commodity, whose value depends on its heterogeneous characteristics. Application of the theory has been well implemented using OLS Regression. Our study investigates this econometric concept using machine learning algorithms. An improved pricing will benefit buyers, sellers, investors, banks and real estate professionals. Normality test for the experiment was done using Chi-Square Quantile-Quantile plot and Henze-Zirkler's Multivariate Normality Test. Statistical relationship was based on correlation matrix, Kaiser-Meyer-Olkin and Bartlett tests. Support Vector Regression (SVR), K-Nearest Neighbor (K-NN) and Principal Component Regression (PCR) were used as learning algorithms. Performance comparison of the learning algorithms was done using spearman's rho correlation coefficient. The performance of the model showed that PCR has a slight edge over SVR and K-NN. Also, the study validated the suitability and substitutability of PCR, SVR and K-NN in the implementation of the hedonic pricing theory.

Keywords
linear regression, PCR, SVM, KNN, hedonic pricing model, housing prices prediction, comparative market analysis, Kaiser-Meyer-Olkin, Bartlett tests, Henze-Zirkler's Multivariate Normality Test

1. INTRODUCTION

The last decade witnessed the sudden bubble of the housing market and its subsequent burst. Just like the law of gravity - values of real estate properties went up and came down. The 'American dream' was turned into a nightmare, foreclosure and short sale became the order of the day. On almost every city in the United States, fore-sale signs became ubiquitous like flowers that beautify a city. Homeowners who depended on the equity of their properties could not belief that they were sitting on empty wealth that had evaporated. To reduce losses, investors withdrew their capitals from the failing system. The catastrophic systemic failure sent shocking and panicking waves into the stock market. The resultant effect reverberated into other sectors of the economy; creating economic chaos, insecurity and instability. To stop the bleeding, the government reluctantly bailed-out banks and offered financial reliefs to some struggling homeowners. The system created feigned, contrived and artificial housing values and the public saw it disappeared into oblivion. Reacting to the trend, the government crafted improved policies, legislations and regulations. The effort was to stop the predatory lending practices of the subprime mortgage industry. Blaming unqualified buyers, the mortgagees adopted automated mortgage underwriting and increased credit requirements of would be mortgagors. The policy of the government and the scrutiny of the banks reduced predatory lending. However, the problem of housing price manipulation by major players in the real estate world is still a concern. All stakeholders should have an answer to the question; "how much does this property worth?". The present manual system is prone to circumvention and manipulation.

Sales comparable method has been widely used in estimating the values of real estate properties. Using this approach, the appraised value of a property is based on the average price of at least three (occasionally banks requests for more) recently sold properties. Comparable properties must have the same features and in the same submarket. If there are no recent sales in a submarket, the market is extended to the nearest one. The major setback with the approach is that, it is open to price manipulation. Using the present manual appraisal mechanism, stakeholders can 'creatively' inflate the price of one property. Once the value of one out of three comparable properties goes up, then based on sales comparison methodology, the average price of the submarket goes up. This is one of the challenges during the housing crisis of the last decade. We hereby propose a PCR, K-NN and SVR algorithmic implementation of the hedonic theory using comparable sales approach. Datasets were sold properties between January and December 2006 (period of housing crisis) obtained from the multiple listing services (MLS) of the realtors.

We will discuss related works in the next section. Section three will be about the experiment, while result and analysis will be in section four. Future works will be discussed in section five and conclusion in section six.

2. LITERATURE REVIEW

2.1 Background

From the basic economic theory of laws of supply and demand, the equilibrium price (intersection of demand and supply curves) determines the market price of a commodity. However, the housing market violates this assumption.

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.103

Therefore, there have been different proposals and recommendations about what actually determines the market value of a real estate property. The hedonic pricing approach has been well accepted as a viable, feasible and practicable model. The model suggests that house is a differentiated commodity, whose value is a function of its heterogeneous characteristics. According to a study, the hedonic pricing model was introduced in 1966, and applied into real estate pricing in 1968 [1]. Researchers argued that the value of a real estate property is a function of its utility. According to the study, the price of a property does not depend on its demand like other commodities, rather, it is mostly influenced by its features. Wen *et al*., in another study pointed out that a real estate property is just like any other heterogeneous goods where features determine utility. The combination of the individual features according to the study, is the total utility to the consumer [2].

Applying the hedonic quintessential tool of pricing, the price of a real estate property at the market equilibrium, is a function of its structural and locational attributes.

Mathematically;

$$V(H) = f(S, L) + \in. \quad (1)$$

Where; $V(H)$, $\in$, S and L are the values of a property, reducible error, structural and locational attributes respectively. Thus, the price of a property (response variable) could be computed from its structural and locational features (explanatory variables).

2.2 Related Work

In the recent times, there have been continued interest in the study of real estate pricing. Using 2075 datasets of properties in Shijiazhuang, Zhao and Liu implemented the hedonic model with a OLS. The model explained the relationship between the price of a property and its features in Shijiazhuang housing market [3]. Wang *et al*., developed a multiple linear hedonic model with datasets from more than 400 properties in Harbin City [4]. Kecheng and Wei proposed a multilevel hedonic model to account for spatial dependencies and non-stationarity of properties [5]. Danlin *et al*., used a geographically weighted hedonic regression (GWHR) to develop a new model to identify submarkets. Datasets from the city of Milwaukee were used for the experiment. The performance of the algorithm was compared with cluster analysis. The outcome of the experiment showed that submarkets based on hedonic regression performed better than the uniform market [6]. Wen *et al*., developed a hedonic real estate price model for the city of Hangzhou. The proposed method produced 18 features that determined the value of properties in the city [7]. Zhu et *al*., proposed a hedonic model for residential properties in Nanjing. Residential price was used as the response variable while fourteen other variables were used as the explanatory variables [8]. Another study implemented hedonic price model using rough set (RS) and support vector machine (SVM). The proposed algorithm developed a new mathematical model for the prediction of real estate prices. The rough set reduced the independent variables which improved the convergence speed and prediction accuracy [9]. In our previous work, we developed a real estate predictive model using MVC architecture and linear regression. As demonstrated, the software application produced an interactive portal where stakeholders in the real estate transactions can sell, buy and verify prices of properties. Datasets of Howard county, Maryland used for the experiment was drawn from the Multiple Listing Services of realtors in Washington DC [10].

3. EXPERIMENT

Four thousand raw datasets of eight counties were extracted from the Multiple Listing Services (MLS) of real estate agents in the Washington DC, metropolitan area of the United States [11]. Datasets from the MLS were considered reliable and complete because it is the official repository of real estate transactions in the metropolitan area. Furthermore, values of sold real estate properties on the MLS are government regulated. The study used 2006 dataset because, 2006 was the pick of mortgage crisis in the United States. Features such as garden, household size, neighborhood satisfaction and schools were considered statistically insignificant in a hedonic model as demonstrated by Berna and Craig [12]. Therefore, we excluded this category of features from our experiment.

3.1 Data Exploratory

3.1.1 Box Plot

The distribution of the datasets based on counties were visualized using a box plot graphical representation. A box plot uses whiskey boxes to show the summary of the distribution of datasets based on minimum, first quartile, median, third quartile, and maximum values. Figure 1shows the box plot of the prices of properties in each of the eight counties. As shown in the diagram, the black line in each box shows the media price, while the whiskers above and below show the maximum and minimum prices respectively. The rectangle itself shows the range between the first and third quartile. outliers are shown with small black circles. The figure shows that Baltimore and Prince Georges counties have the closest range, while Anne Arundel county has the highest range. From the box plot, it is obvious that price ranges are not the same in the eight counties.

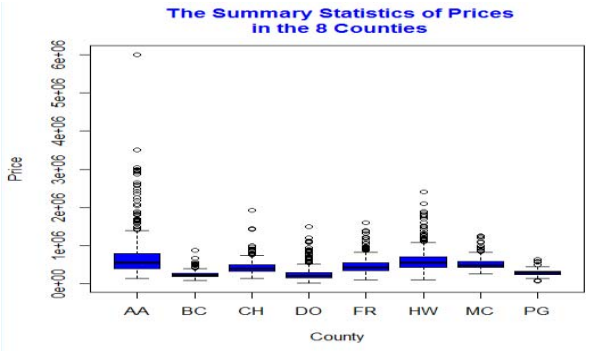

Figure 1. Box Plot of Housing Price in Anne Arundel, Baltimore, Charles, Dorchester, Fredrick, Howard, Montgomery and Prince Georges counties.

Keeping all other factors constant, the figure suggests that home buyers who want low prices may have better options in Baltimore or Prince Georges counties. On the other hand, sellers with expensive properties in Anne Arundel and Howard County counties are in a suitable market.

3.1.2 Normality of Distribution

Considering the fact that our dataset was extracted from different submarket (eight counties), therefore, a normality test was necessary to know if they were normally distributed. The normality of the distribution was investigated using Chi-Square Quantile-Quantile Plots and Henze-Zirkler's Multivariate Normality Test.

3.1.2.1 Chi-Square Quantile-Quantile Plots

We tested the normality of the distribution using a Chi-Square Q-Q Plot. Given that a normal distribution should follow a bell-shaped appearance, the Chi-Square Q-Q plot of a multivariate normal distribution is expected to produce points on a perfect straight line [13]. Mahalanobis squared distance was computed between parameters and plotted against the estimated quantiles in a chi-squared distribution with *p* degrees of freedom and a sample size n. Figure 2 shows Chi-Square Q-Q Plot of the distribution. The number of the parameters is the degree of freedom. Outliers are shown as dots on the graph.

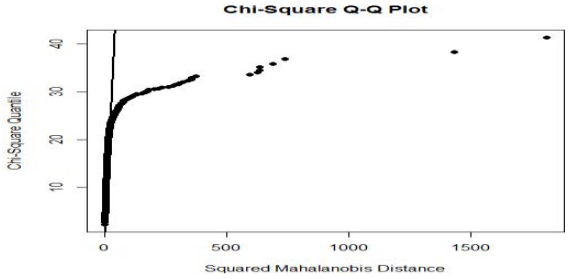

Figure 2. Chi-Square Q-Q Plot of the Distribution

As shown in the figure, the graph does not sit perfectly on the straight line. The Chi-Square Q-Q plot suggests that the datasets come short of a normal distribution.

3.1.2.2 Henze-Zirkler's Multivariate Normality Test

We confirmed the Chi-Square Q-Q plot result using Henze-Zirkler's Multivariate Normality Test. The test was based on Euclidean distance between sample variables in a dataset. Unlike Royston test which is limited to 2000 observations [14], Henze-Zirkler's Multivariate Normality Test has the capability of testing datasets that are more than 2000. We hypothesized that datasets obtained from eight submarkets on the MLS followed a normal distribution. Henze-Zirkler's multivariate test was computed, 9.632816 and 0 .00 were obtained for the HZ and p-values respectively. The result showed a p value less than the significant value of 0.05. Therefore, there is evidence that samples were not drawn from a normal distribution.

3.1.2.3 Normalization

Our tests confirmed that the datasets fell short of a normal distribution. It was therefore necessary to normalize the datasets. We used min-max normalization for rescaling the datasets into a normalized form [15]. All normalized parameters are expected to fall between the range 0 and 1. Normalization was computed using the following relationship; $X_{new} = \frac{X-\min(X)}{\max(X)-\min(X)}$, Where X_{new} is the new value of X, min(x) minimum value of X and max(X) the maximum value of X

3.1.3 Statistical Relationship

We further investigated the normalized datasets and its suitability for a principal component analysis using, correlation matrix, Kaiser-Meyer-Olkin and Bartlett tests.

3.1.3.3 Correlation Matrix

A correlation matrix shows the statistical relationship between two variables of a dataset in a matrix form. Relationship can be positive, negative or none. Figure 2 simplifies the correlation between the variables in the form of circles and ellipses.

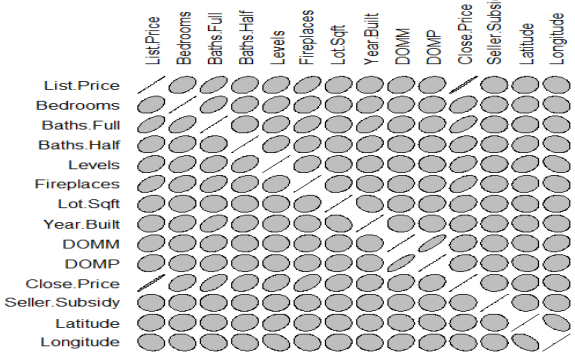

Figure 2. Correlation Matrix

A positive correlation suggests that the two parameters increase and decrease together. Negative correlation on the other hand implies that two variables increase and decrease in opposite directions. From the figure, a perfect circle suggests that two variables are not related (zero correlation). An example of zero correlation is Level vs DOMM (number of levels of a property vs Number of days on the MLS market). The relationship suggests that the number of levels of a property may not necessarily determine how long it will stay in the market, provided other factors are constant. A small ellipse shows a strong relationship. A good example is Close Price vs Full Bath, the shape suggests positive relationship. This implies that provided other factors are held constant, the

number of baths in a property has a considerable impact on its price.

The smaller the eclipse, the stronger the relationship. An eclipse that tilts right is a positive correlation, while the negative tilts left. In numerical form, correlation takes value between -1 and 1. A negative 1 suggests that the two variables are perfectly moving in opposite trajectories, while a positive 1 suggests a perfect trajectory in the same direction. The correlation matrix shows that some components are related, while others are not related. A further test is necessary to investigate the statistical relationship between the variables.

3.1.3.4 Kaiser-Meyer-Olkin Test

We further investigated the statistical relationship between the variables using Kaiser-Meyer-Olkin Test. The MSA (Measure of Sampling Adequacy) takes value between 0 and 1. Overall MSA is the mean value of the statistical relationship between variables [16]. Overall MSA that is more than 0.5 suggests that there is enough correlation between the variables for a dimensioning reduction. The result of our test shows that only two attributes (Year Built and Seller Subsidy) have MSA values that are below the threshold of 0.5. A quick fix was to remove the two parameters, however, the overall MSA value was 0.69. Therefore, a further test was necessary to confirm the relationship.

3.1.3.5 Bartlett tests

Bartlett test confirmed the assumed statistical relationship between the variables. The algorithm measures the homoscedasticity or homogeneity of variances of the distribution. It tests whether the correlation matrix is closer or farther from an identity matrix (in identity matrix, all elements are zero except for the diagonal elements). If the result shows that the correlation matrix follows the identity matrix pattern, it suggests that there is no relationship between the variables. This implies that they are independent. The null hypothesis for a Bartlett test is that variances are the same while the alternative hypothesis is that they are different. Table 1 shows the result of the test

Table 1. Bartlett Tests Result

Chi square	P value	DF
752.6801	9.449e-105	91

As shown in the table, the test generated a chi square value of 752.6801 with a very low p-value. Thus, there exist a strong evidence that the correlation matrix is far from an identity matrix. We therefore reject the Bartlett null hypothesis and accepted the alternative.

3.1.4 Principal Component Analysis

Correlation matrix, Kaiser-Meyer-Olkin Test and Bartlett tests provided enough evidence that there exists statistical relationship between the variables. Therefore, our dataset was suitable for decomposition into its principal components. A principal component analysis (PCA) was necessary to find the core components of the datasets, increase convergence speed and eliminate collinearity. The PCA algorithm transformed the data into some smaller and more meaningful components which were the true representation of the attributes. In other words, new variables were derived from the datasets. PC1 is the first principal component; it is the linear combination of variables with most possible variance in the datasets. The second one, PC2 covers the next variance. A new principal component is assumed to be uncorrelated with all previous components.

Mathematically;

$$nPC_i = (a_{i1}\, V_1) + (a_{i2}\, V_2) + \cdots + (a_{in}\, V_n). \quad (2)$$

Where nP is the number of components, $a_{i1}\ldots..a_{in}$ are the component weights and $V_1\ldots..V_n$ the variables.

3.2 Learning Algorithm

As demonstrated above, dataset for the study were preprocessed, normalized and decomposed. Statistical relationship between variables were also established. According to the hedonic theory, the price of a property is a function of its features. Therefore, the response variable for our experiment was the price of the property, while the features were the explanatory variables. Hedonic housing pricing model was implemented using principal component regression (PCR), support vector regression (SVR) and K nearest neighbors (K-NN) learning algorithms. Datasets were divided into *k* equal sizes. Subset *k* was used for training while the remaining *k*-1 was used for the testing. In all, validation was done *k* times; ensuring *k* subsets were used only once for the training. The final estimate is the average of k results from *k* iterations. The value of *k* for the experiment was 10. Performance comparison of the learning algorithms was done using spearman's rho correlation coefficient.

4. RESULT AND ANALYSIS

4.1 Principal Component Analysis Result

The PCA algorithm as expected derived new variables that were representative of the original variables. The first component captured more variance, followed by the second and so on. Figure 4 shows the scree plot of the principal components. The horizontal line parallel to the component number is the cut off point for principal components that are more than one eigenvalue. These are the five principal components that explained at least 85% of the variance. The question here is "what is the probability that the presumed relationships between the principal components and the price of a property based on hedonic regression is not random chance?" The question was answered by conducting a t-test for the study. Test of significance table of the five principal

components is shown in table 2. A p-value of 0.05 was selected as evidence to either reject or fail to reject the hypothesis. A p-value that is less than 0.05 is considered a significant evidence against the null hypothesis. The null hypothesis here is that a given principal component does not have effect on the hedonic price of a property. If a p-value is less than 0.05 on a "one-tailed" t-test, it is considered a significant evidence against the null hypothesis. Thus suggesting that changes in the value of the principal component are strongly related to changes in the value of a property. Conversely, a p-value greater than 0.05 implies that changes in the value of the principal component do not have a significant effect on the hedonic price of a property.

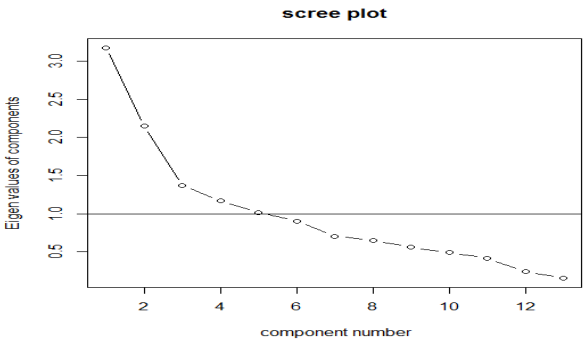

Figure 4. The Scree Plot of principal components.

In the table, the code column shows the level of significance of each component (four is the maximum, one is the minimum). The coef column is the coefficient value of each component, while std.Er and std.c are the standard error and standard coefficient respectively. T-Stat is the test statistics and pVal the p-value.

Table 2 Principal Component Analysis Test of Significance Table

PC	Coef	Std.Er	Std.C	t-Stat	pVal	Code
Pc1	-59032	1545.6	-0.42	-38.2	0	****
Pc2	-77574	1947.1	-0.44	-39.8	0	****
Pc3	30603	2524.2	0.134	12.12	0	****
Pc4	18005	2600.00	0.076	6.925	0	****
Pc5	11851	2787.61	0.047	4.252	0	****
Intercept	440949	2401	N/A	183	0	****

4.2 Performance of Learning Algorithms

We compared the performance of the three learning algorithms using spearman's rho correlation coefficient. Spearman's rho is a measure of the monolithic functional description of the statistical relationship or strength of association between the Close.Price (actual price) of the property and our predicted price. A positive spearman's rho suggests that both variables increase and decrease together, while a negative value suggests that they move in opposite trajectories. Spearman's rho value ranges between positive one and negative one. Exact Positive one and negative one implies perfect trajectories in the same and opposite trajectories respectively; this is rare in most experiments. Figure 6 shows the performance comparison bar chart of the learning algorithms. The figure shows that PCR performed slightly better than SVM and KNN. A 0.886 spearman's rho performance value of the PCR means that the predicted price is 11.4% close to become a perfect monotonic function of the actual value of a property. On the other hand, the K-NN and SVM are 16.6% and 13% respectively away from perfect monotonic functionally related to the actual price of a property. Thus, the strength of association between the PCR and the actual price of the property is the strongest of the three learning algorithms.

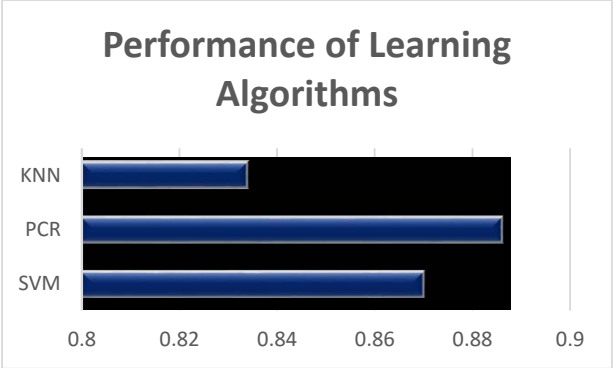

Figure 6. Performance Comparison Bar Chart

5. FUTURE WORK

a. Our study was based only on 2006 datasets of properties listed and sold on the MLS. The time frame may create a bias on the algorithm. Our future work includes building a spatial temporal hedonic model spanning several years.
b. Our model is suitable for estimating the value of a property and not a replacement for an appraisal report. Building models to fully automate the appraisal process is our ultimate goal.

6. CONCLUSION

We have implemented the hedonic theory using three different learning algorithms (PCR, SVR and K-NN). PCA was used for component analysis and decomposition. Normality test was done using Chi-Square Quantile-Quantile plot and Henze-Zirkler's Multivariate Normality Test. Statistical relationship was confirmed by correlation matrix, Kaiser-Meyer-Olkin and Bartlett tests. The outcome of the study shows that;

a. The price of a property is predictable using hedonic theory.
b. There exists an evidence of statistical relationship between the price of a property and explanatory parameters on the MLS datasets.
c. Hedonic pricing theory is implementable using PCR, K-NN or SVR.
d. PCR performs slightly better than K-NN and SVR when implementing the hedonic housing theory.

7. ACKNOWLEDGEMENTS

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

REFRENCES

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

[2] J. F. Lu and L. Lin H. Z. Wen, "An Improved Method of Real Estate Evaluation Based on Hedonic Price Model," in *IEEE*, 2004.

[3] X. j. Liu Y. Zhao, "Hedonic Price Study on Urban Housing: The Case of Shijiazhuang City," in *International Conference on Management and Service Science (MASS)*, Wuhan, 2010.

[4] T. Guang-ji and Z. Hong-rui W. Wei, "Empirical analysis on the housing price in Harbin City based on hedonic model," in *International Conference on Management Science and Engineering (ICMSE)*, Melbourne, 2010.

[5] Wei Shen Kecheng Zhao, "Spatial characteristic with individual house properties and multilevel approach to hedonic models," in *International Conference on Computer Science and Service System (CSSS)*, Nanjing, 2011.

[6] Jingyuan Yin and Feiyue Ye D. Yu, "Novel methods to demarcate urban house submarket - Cluster analysis with spatially varying relationships between house value and attributes," in *IET International Conference on Smart and Sustainable City* , Shanghai, 2011.

[7] 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.

[8] L. Qi-ming, Z. Jian-jun, "An Empirical Study of Residential Hedonic Prices in Nanjing," in *International Conference on E-Business and E-Government (ICEE)*, Guangzhou, 2010.

[9] Y. Q. Li and S. F. Zhao T. Wang, "Application of SVM Based on Rough Set in Real Estate Prices Prediction," in *4th International Conference on Wireless Communications, Networking and Mobile Computing*, Dalian, 2008.

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

[11] MRIS. (2016) [Online]. http://www.mris.com/

[12] Craig Watkins Berna Keskin, "Defining spatial housing submarkets: Exploring the case for expert delinated boundaries," *Urban Studies Journal Limited*, pp. 1-17, 2016.

[13] Mark Gardener, *Begining R. The Statistical Programming Language*. Indianapolis: John Wiley & Sons, Inc, 2012.

[14] Matias Salibian-Barrera and Katarzyna Naczk Patrick J. Farrell, "On tests for multivariate normality and associated simulation," *Journal of Statistical Computation & Simulation*, pp. 1-14, 2006.

[15] Brett Lantz, *Machine Learning with R*. BIRMINGHAM - MUMBAI: Packt, 2015.

[16] Gergely Daróczi, *Mastering Data Analysis with R*. Birmingham: Packt Publishing, 2015.

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