Comparative Study On Estimate House Price Using Statistical And Neural Network Model

Azme Bin Khamis, Nur Khalidah Khalilah Binti Kamarudin

Abstract: This study was conducted to compare the performance between Multiple Linear Regression (MLR) model and Neural Network model on estimate house prices in New York. A sample of 1047 houses is randomly selected and retrieved from the Math10 website. The factors in prediction house prices including living area, number of bedrooms, number of bathrooms, lot size and age of house. The methods used in this study are MLR and Artificial Neural Network. It was found that, the value of R^2 in Neural Network model is higher than MLR model by 26.475%. The value of Mean Squared Error (MSE) in Neural Network model also lower compared to MLR model. Therefore, Neural Network model is prefered to be used as alternative model in estimating house price compared to MLR model.

Index Terms: multiple linear regression, artificial neural networks, estimate, house price, mean square error, R^2, model performance

1. Introduction

Housing market is of great important for the economy activities. Housing construction and renovation boost the economy through an increase in the aggregate expenditures, employment and volume of house sales. They also simulate the demand for relevant industries such as household durables. The oscillation of house prices affects the value of asset portfolio for most households for whom a house is the largest single asset [16]. An accurate prediction on the house price is important to prospective homeowners, developers, investors, appraisers, tax assessors and other real estate market participants, such as, mortgage lenders and insurers [6]. Traditional house price prediction is based on cost and sale price comparison lacking of an accepted standard and a certification process. Therefore, the availability of a house price prediction model helps fill up an important information gap and improve the efficiency of the real estate market [2]. According to [17], the results show that the houses with more bedrooms and bathrooms are priced higher. A relatively new house is more expensive than an old house and a house with a garden is priced higher than one without a garden. Recent studies further justify the necessity of housing price analysis with a conclusion that housing sector plays a significant role in acting as a leading indicator of the real sector of the economy and assets prices help forecast both inflation and output [5][[27][4]. Many previous studies find empirical evidence supporting the significant interrelations between house price and various economic variables, such as income, interest rates, construction costs and labor market variables [18][24][28].

- *Azme Khamis is Associate Professor in statistics at University Tun Hussein Onn Malaysia.*
  E-mail: azme@uthm.edu.my
- *Nur Khalidah Khalilah Binti Kamarudin is currently pursuing masters degree program in statistics in University Tun Hussein Onn Malaysia*
  E-mail: hw130037@siswa.uthm.edu.my

Housing market is illiquid and heterogeneous in both physical and geographical perspectives, which makes forecasting house price a difficult task. Moreover, the subtle interactions between house price and other macroeconomic fundamentals make the prediction further complicated. The change in house prices can either reflect a national phenomenon, such as the effect of monetary policy, or be attributed to local factors—circumstances that specific to each geographic market [8]. It can either indicate the changes in the real sector variables, such as labor input and production of goods, or be affected by the activities in the nominal sector, i.e., financial market liberalization [10]. Apart from the above close link to housing investment, house prices have a strong link with both income and interest rates – both via a standard housing demand function and a housing supply function. On the demand side, [19] propose a theoretical model of house price determination that is driven by changes in income and interest rates. [9] studied indicate that in advanced economies real house prices have fluctuated around an upward trend at least since 1970 at euro market, generally attributed in the literature to rising demand for housing space – linked to increasing per capita income as well as a growing population on the demand side. [12] applied artificial neural network to evaluate the current market situation during the world economic crisis in 2008 and predicted the future performance of property in order to help investors and other market players in making important decisions. Although artificial neural network has been limitedly used for valuation or forecasting property price, studies carried out to compare the accuracy of linear regression and artificial neural network discovered that the latter has superiority compared to the former. [23] compared linear regression and artificial neural network in predicting housing value. [22] used artificial neural network and compared its accuracy with that of linear regression in predicting of housing price. [13] in their research also compared linear regression and artificial neural network in the mass appraisal context. [7] simulated a hypothesis in relation to valuing real estate value in Madrid. Forecasting has some degree of uncertainty. However, a high degree of sophistication has been developed recently, with a range of advanced quantitative and qualitative procedure used by institutional investors in property forecasting, including judgemental procedures, causal or econometric procedures, and time series and trend analysis procedures [20]. This study aimed to compare between MLR model and

Neural Network model to predict the house prices in New York. Secondary data from 1047 houses in New York is used in artificial neural network to predict the house price and determine whether the prediction is good or not. The secondary data was collected in year 2012. The data consist of house price, living area, number of bedrooms, number of bathrooms, lot size and age of house [3][21]. The living area, number of bedrooms, number of bathrooms, lot size and age of house will be in input layer while house price will be in output layer. There is a total of 1047 data points in which 70% was used for training, 15% for validation and another 15% for testing. All 1047 experimental data sets are divided for training, validation and testing. Using Neural Network Toolbox (nntool) in MATLAB, different network configuration with different number of hidden neurons is trained and their performance is checked. There are 733 data sets are used for training, 157 data sets for validation and 157 data sets for testing.

2. Research Methodology

2.1 Multiple Linear Regression

Regression is a fundamental operation in statistics and includes techniques for modelling and analyzing several variables at a time. Regression analysis is used for explaining the relationship between a dependent variable, usually denoted by *Y*, and a number of independent variables, $X_1, X_2, \ldots, X_p$. The independent variables are also known as predictor or explanatory variables. In most regression analyses, the variables are assumed to be continuous. In simple regression, there is only one independent variable. However, most real world applications involve more than one variable which influence the outcome variable. The model for Multiple Linear Regression can be represented as:

$$E(Y/X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where β_0 is called intercept and β_j are called slopes or regression coefficients. The difference between the predicted and the actual value of *Y* is called the error (ε) or can be written as $\varepsilon = \hat{Y} - Y$. Then, regression equation can be express as:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$

where Y_i is the actual value and ε_i is the error for the i^{th} observation. We write $X_{i,j}$ for the j^{th} predictor variable measured for the i^{th} observation. The main assumptions for the errors ε_i is that $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$. Also the ε_i are randomly distributed. The predicted value is also denoted by $\hat{Y}$. The various errors are given as:

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 ; \; SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SSR = \sum_{i=1}^{n} \left(\bar{Y} - \hat{Y}_i\right)^2;$$

SSE is the Sum of Squares of Error, *SSR* is the Sum of Squares of Regression, and *SST* is the Sum of Squares Total. *R*-square is the square of the correlation between the response values and the predicted response values. It is also called the square of the multiple correlation coefficients and the coefficient of multiple determinations. The coefficient of determination is the overall measure of the usefulness of a regression. It is given as;

$$R^2 = 1 - \frac{SSE}{SST}$$

The value of R^2 can range between 0 and 1, a higher value indicates a better model. In terms of the sample, the estimate of the population total variance (*SST*) is denoted by Mean Sum of Squares Total (*MST*). *MST* is obtained as *SST*/(*n*-1) where *n* is the sample size. Similarly, the estimated residual or error is called Mean Squared Error (*MSE*) and is calculated as, *MSE* = *SSE* / (*n-p-1*) where *n* is the sample size, and *p* is the number of exploratory variables. A better estimate of the coefficient of determination is made by the Adjusted-R squared statistic:

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{MSE}{MST}$$

The *F-test* in one way Analysis of Variance (ANOVA) is also used as a statistic to find the goodness of fit of the model. It is calculated as:

$$F_{test} = \frac{explained\ variance}{unexplained\ variance} = \frac{SSR/p}{SSE/(n-p-1)}$$

2.2 Artificial Neural Network

Artificial neural network (ANN) is an artificial intelligence model originally designed to replicate the human brain's learning process. ANN is distributed through a dense web of interconnections. A neural network is formed by a series of neurons or nodes that are organized in layers. Neural networks consist of processing units (artificial neurons) and connections (weights) between those units. The processing units transport incoming information on their outgoing connections to other units. The input information is simulated with specific values stored in those weights that give these networks the capacity to learn, memorize, and create relationships between data. Each neuron in a layer is connected with each neuron in the next layer through a weighted connection. The value of the weight w_{ij} indicates the strength of the connection between the i^{th} neuron in a layer and the j^{th} neuron in the next one. The structure of a neural network is formed by an input layer, one or more hidden layers, and the output layer or can be summarized as (input, hidden node, output). The number of neurons in a layer and the number of layers depends strongly on the complexity of the problem studied. Therefore, the optimal network architecture must be determined. The general scheme of a typical three-layered ANN architecture is given in Fig. 1. The w_{ij} is the weight of the connection between the i^{th} and the j^{th} node. The neurons in the input layer receive the data and transfer them to neurons in the first hidden layer through the weighted links. Here, the data are mathematically processed and the result is transferred to the neurons in the next layer. Ultimately, the neurons in the

last layer provide the network's output. The j^{th} neuron in a hidden layer processes the input data, x_i by:

Step 1. Calculating the weighted sum and adding a bias term, δ_j according to equation 1:

$$f(net)_j = \sum_{i=1}^{n} x_i w_{ij} + \delta_j \quad \text{for } j = 1, 2, \ldots, n \quad (1)$$

Step 2. Transforming the $f(net)_j$ through a suitable mathematical transfer function or activation function, and

Step 3. Transferring the result to neurons in the next layer. Various transfer functions are available [15][25]. Some of the most commonly used activation functions are; (i) is linear function, $g(x) = x$. It is obvious that the input units use the identity function. Sometimes a constant is multiplied by the net input to form a linear function; (ii) sigmoid function, $g(x) = \frac{1}{1+e^{-x}}$. This function is especially advantageous for use in neural networks trained by back-propagation, because it is easy to differentiate, and thus can dramatically reduce the computation burden for training. It applies to applications whose desired output values are between 0 and 1, and Figure 2 (iii) hyperbolic tangent function, $g(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$. This function has similar properties to the sigmoid function. It works well for applications that yield output values in the range of -1 and 1.

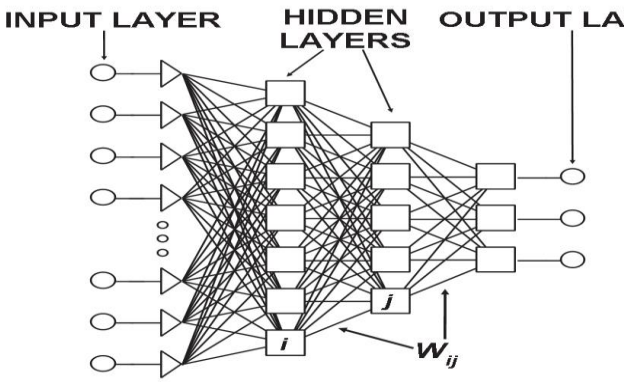

Figure 1. *General structure of a neural network with input layer, two hidden layers and output layer*

The mathematical process through which the network achieves learning can be principally ignored by the final user. In this way, the network can be viewed as a black box that receives a vector with *m* inputs and provides a vector with *n* outputs. Here we will give only a brief description of the learning process; more details are provided for example in the review by [26]. The network learns from a series of examples that form the training data set. Training is formed by a vector $X(inp)_{im} = (x_{i1}, x_{i2}, \ldots, x_{im})$ of inputs and a vector $Y(out)_{in} = (y_{i1}, y_{i2}, \ldots, y_{in})$ of outputs. The objective of the training process is to approximate the function *f* between the vectors $X(inp)_{im}$ and the $Y(out)_{in}$, $Y(out)_{in} = f(X(inp)_{im})$. This is achieved by changing iteratively the values of the connection weights, w_{ij} according to a suitable mathematical rule called the *training algorithm*. The values of the weights are changed by using the steepest descent method to minimize a suitable function used as the training stopping criteria. One of the functions most commonly used is the sum-of squared residuals given by equation (2),

$$E = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{ij} - y_{ij}^{*}\right)^2 \quad (2)$$

where y_{ij} and y_{ij}^{*} are the actual and network's j^{th} output corresponding to the i^{th} input vector, respectively. The current weight change on a given layer is given by equation (3):

$$\Delta w_{ij} = -\eta \frac{dE}{dw_{ij}} \quad (3)$$

where η is a positive constant called the *learning rate*. To achieve faster learning and avoid local minima, an additional term is used and equation (3) becomes:

$$\Delta w_{ij}^{k} = -\eta \frac{dE}{dw_{ij}} + \mu \Delta w_{ij}^{k-1} \quad (4)$$

where μ is the "momentum" term and Δw_{ij}^{k-1} is the change of the weight w_{ij} from the $(k\text{-}1)^{th}$ learning cycle. The learning rate controls the weight update rate according to the new weight change and the momentum acts as a stabilizer, being aware of the previous weight change. The function given by equation (2) is also used as the criterion to optimize the network architecture because it depends on the number of hidden layers and the number of neurons therein. To find the optimal architecture, the most common approach is to plot the value of *E* in equation (2) as a function of the number of nodes in the hidden layer (*q*). Backpropagation is the most common learning algorithm for feed forward network. Backpropagation simply the gradient descent method to minimize the total squared error of output computed by the net. Training a network by backpropagation involves three stages: the feed forward of the input training pattern, the backpropagation of the associated error, and the adjustment of the weights [15]. The number of hidden neurons with lowest MSE will be choose as an optimum number of hidden neurons, and model with higher R^2 and lower MSE was considered to be a good model.

3. Data Analysis and Results

3.1 Regression analysis

The relationship between dependent variable with independent variables was performed using Pearson correlation. It was found that the correlation coefficient indicated that house prices are positive and strongly significance to living area (0.776), bathroom (0.670), bedrooms (0.471) and fireplace (0.460), but negative relationship to age of house (-0.363). Meanwhile, house price is not significance to lot size. The value of *F* statistic is 380.696 and *p*-value is 0.0000, means that the model is suitable and can be fitted to the data. The coefficient of determination R^2 = 0.646 and *Adj-*R^2 = 0.645, it shows that 64.5% variance in house price can be explained by living area, number of bathrooms, number of bedrooms, lot size and age of house. The regression equation for the house price can be written as follow;

House price =

27467.001 + 67.211**Living area* -216.718**Age*+ 16,402.244**Bathrooms* +10,099.817**Fireplace* – 5167.260**Bedrooms.*

3.2 Artificial Neural Network

The number of nodes or neurons in hidden layer are determined by trial and error process. We starts our trial and error with 2 nodes and the process is repeated until 15 nodes. The researcher compares the MSE value and *R* value for all number of nodes. The lowest MSE value with higher *R* value will be selected as optimum number of nodes in hidden layer. Based on Table 8, the lowest MSE value is 1.293E9 with 10 nodes in hidden layer and correlation coefficient is 0.9039. Hence, 10 nodes are selected as optimum number of nodes in hidden layer.

Table 8 : Optimum number of nodes in hidden layer

Number of nodes	MSE	*R*	Number of nodes	MSE	*R*
2	1.574 E9	0.8118	9	1.401 E9	0.8135
3	2.566 E9	0.8215	**10**	**1.293 E9**	**0.9039**
4	1.843 E9	0.8341	11	2.108 E9	0.8845
5	1.443 E9	0.8378	12	1.507 E9	0.8752
6	1.468 E9	0.8563	13	1.433 E9	0.8588
7	1.476 E9	0.8665	14	1.434 E9	0.8231
8	1.476 E9	0.8239	15	1.499 E9	0.8025

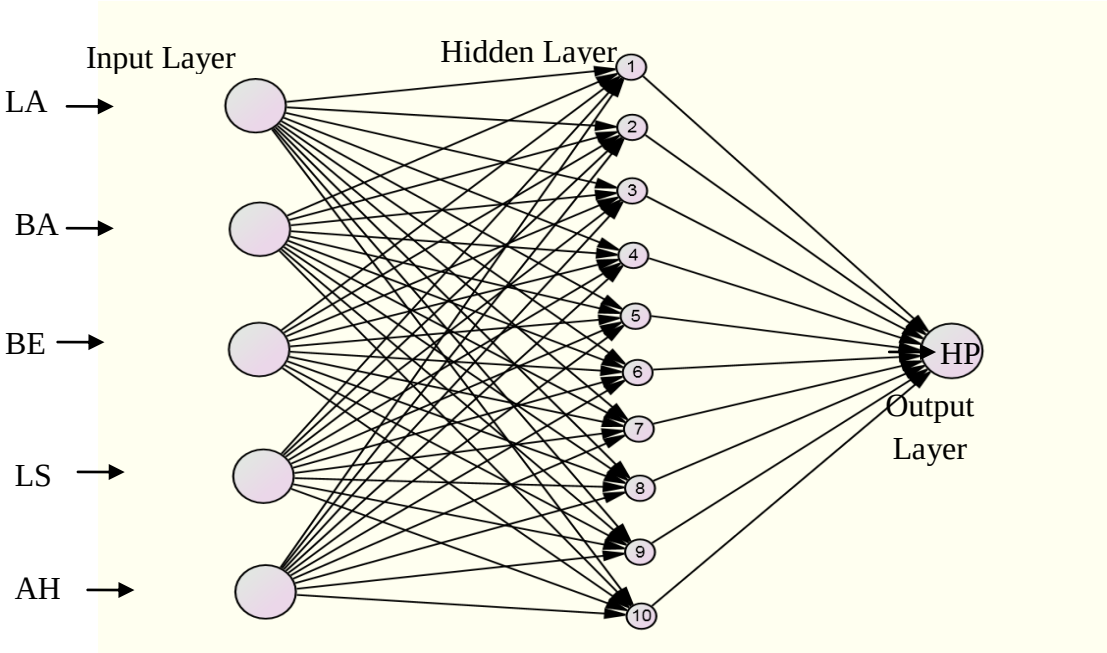

Figure 2 : *Architecture of Artificial Neural Network with three layers*

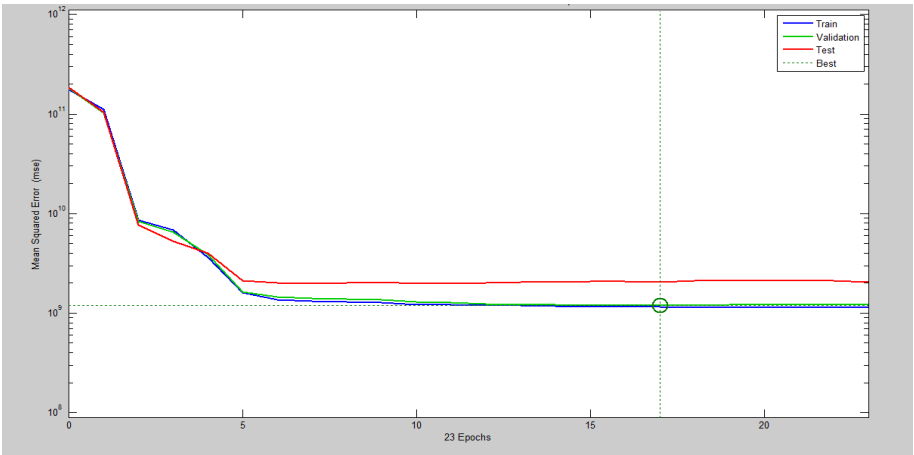

Figure 4 : *Performance plot*

The architecture in Figure 2 shows that there are 5 input layers, 1 hidden layer consists of 10 nodes and 1 output layer. The inputs are living area in square feet (LA), number of bathrooms (BA), number of bedrooms (BE), lot size in acre (LS) and age of house in year (AH). The output is house price (HP).Figure 4 shows retrained performance (MSE) graph of neural network model, created during its training. The training stopped after 23 epochs because the test error increased. It is a useful diagnostic tool to plot the training, validation, and test errors to check the progress of training. The result here is reasonable because the training set error and the validation set error have similar characteristics, and it doesn't appear that any significance over fitting has occurred. After initial training of neural network model, it is retrained for 23 epochs and performance MSE is obtained 1.293E9 in training.

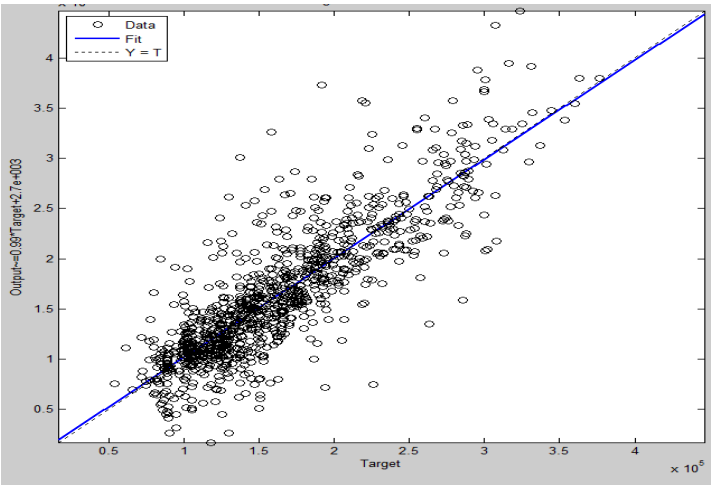

Figure 5 *: Regression plot*

Based on Figure 5, the value for *R* is 0.0.9039 and the value of R^2 is 0.817. This shows that 81.7% of total variation in house price was explained by living area, number of bathrooms, number of bedrooms, lot size and age of house. The value of R^2 and MSE for MLR model is 0.644 and 1.633E9. The value of R^2 and MSE for Neural Network model is 0.817 and 1.293E9. The R^2 value for Neural Network model is higher compared to MLR model. The value of MSE in Neural Network model is lower compared to MLR model. Therefore, Neural Network model is prefered to predict house price.

4. Conclusion

The model's accuracy in predicting house price was measured by a number of criteria. The value of R^2 and MSE were compared to select preferred model. By using ANN, the R^2 value was increase about 26.475% higher than MLR. It can be conclude Neural Network model is prefered to predict house price compared to MLR model and can be used as an alternative way to estimate house price in future.

5. References

[1] M. H. Beale, M. T. Hagan, & H. B. Demuth, "Neural Network Toolbox™ User's Guide". The MathWorks, Inc., 2013.

[2] C. A. Calhoun, "Property valuation models and house price indexes for The Provinces of Thailand: 1992–2000". Housing Finance International, 17: 31 – 41, 2003.

[3] Creative Research Systems. (n.d.). Retrieved December 22, 2013, from Survey System: http://www.surveysystem.com/correlation.htm., 2013.

[4] S. Das, R. Gupta, & A. Kabundi, "Could we have predicted the recent downturn in the South African housing market?" Journal of Housing Economics 4:325-335, 2009.

[5] M. Forni, M. Hallin, M. Lippi, & L. Reichlin, "Do financial variables help forecasting inflation and real activity in the euro area?" Journal of Monetary Economics 6: 1243-1255, 2003.

[6] J. Frew, & G. D. Jud, "Estimating the value of apartment buildings", The J. Real Estate Res., 25: 77 – 86, 2003).

[7] J. Gallego, & Mora-Esperanza (2004). "Artificial intelligence applied to real estate valuation: An example for the appraisal of Madrid". Catastro: 255-265.

[8] L. Gattini, & P. Hiebert, "Forecasting and assessing euro area house prices through the lens of key fundamentals", Working Paper Series, No.124/ October, 2010, 2010

[9] N. Girouard, , M. Kennedy, P. van den Noord, & C. Andr´e, "Recent house price developments: The role of fundamentals", OECD Economics Department Working Paper No. 475, 2006

[10] R. Gupta, S.M. Miller, & D.V. Wyk, "Financial market liberalization, monetary policy, and housing price dynamics". Working paper No. 201009, Dept. of Econ., University of Pretonia. 2010.

[11] H. Hossein, A.Khairil, , H. T. Huam, , K. Naser, , & R. Mohsen, "Artificial neural networks: Applications in management". World Applied Sciences Journal , 14 (7), 1008-1019, 2011.

[12] A. Khalafallah, "Neural network based model for predicting housing market performance". Tsinghua Science and Technology, 13(1): 325-328, 2008.

[13] M. Khashei, & M. Bijari, "An artificial neural network (p, d, q) model for time series forecasting", Expert Systems with Applications, 37: 479-489, 2010.

[14] S. V. Kunwar, & K. B. Ashutosh, "An Analysis of the Performance of Artificial Neural Network Technique for Stock Market Forecasting". International Journal On Computer Science and Engineering , 2 (6), 2104-2109, 2010.

[15] F. Laurene, "Fundamentals of Neural Networks : Architectures, Algorithm, and Applications". United States: Florida Institute of Technology, 1994.

[16] Y. Li, & D. J. Leatham, "Forecasting housing prices: Dynamic factor model versus LBVAR model". Paper for presentation at the Agricultural & Applied Economics Association's 2011 AAEA & NAREA Joint Annual Meeting, Pittsburgh, Pennsylvania, July 24-26, 2011, 2011.

[17] V. Limsombunchai, C. Gan, & M. Lee, "House price prediction : Hedonic price model". American Journal of Applied Sciences , 1 (3), 193-20, 2004.

[18] P. Linneman, "An empirical test of the efficiency of the housing market". Journal of Urban Economics 20(1986): 140-154, 1986.

[19] K. McQuinn, & G. OReilly, "A model of cross-country house prices, Central Bank and Financial Services Authority of Ireland", Research Technical Paper 5 (July), 2007.

[20] M. S. Mohd Radzi, , C. Muthuveerappan, N. Kamarudin, & I. S. Mohammad, "Forecasting house price index using artificial neural network", International Journal of Real Estate Studies, Volume 7, Number 1, 2012

[21] Mymat. "Probabilities and Statistics : Guidance to solve House Prices Data Set Statistical Analysis". Retrieved May 28, 2014, from Math10.com., 2013.

[22] S. T. Ng, , & M. Skitmoreb, "Using genetic algorithms and linear regression analysis for private housing demand forecast". Building and Environment, 43: 1171-1184, 2008.

[23] N. Nguyen & Al Cripps, "Predicting housing value: a comparison of multiple regression analysis and artificial neural networks", Journal of Real Estate Research, Vol . 22, No.3. 2001.

[24] J.M. Quigley, "Real estate prices and economic cycles". International Real Estate Reviews 2: 1-20. 1999.

[25] H. Simon, "Neural Networks : A Comprehensive Foundation", 2nd Edition. Hoboken, New Jersy, Prentice-Hall, 1999.

[26] S. N.Sivanandam, S. Sumathi, , & S. N. Deepa, "Introduction To Neural Networks using MATLAB 6.0". New Delhi: The McGraw-Hill Companies, 2006.

[27] J.H. Stock, & M.W. Watson, "Macroeconomic forecasting using diffusion indexes". Journal of Business and Economic Statistics 2: 147-162, 2002.

[28] K.Tsatasaronis, & H. Zhu, "What drives housing price dynamics: Cross-country evidence?" BIS Quarterly Review March,