2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)

Housing Price Prediction Using Neural Networks

Wan Teng Lim, Lipo Wang, Yaoli Wang, and Qing Chang

***Abstract*—The forecast of Singapore condominium prices is important for potential buyers to make informed decisions. This paper applies two algorithms to predict Singapore housing market and to compares the predictive performance of artificial neural network (ANN) model, i.e., the multilayer perceptron, with autoregressive integrated moving average (ARIMA) model. The more superior model is used to predict the future condominium price index (CPI). The lower mean square error (MSE) of the ANN models showed the superiority of ANN over other predictive tools.**

***Index Terms*—Housing Price Prediction; Time series; Forecast; Neural Networks.**

I. Introduction

Housing can be a shelter to fulfil the fundamental need of an individual, and it can also be a form of investment. Data from Singapore Department of Statistic shows that 81.9% of residents live in public housing (Government of Singapore, 2014) with the rest living in private housing, such as apartments, condominiums or landed properties. Condominiums were introduced in 1972 and are generally more popular among private housing (Wong & Yap, 2003) as there are built-in facilities such as clubhouse, barbecue area, gymnasium, and swimming pool. Most condominiums also have 24-hours security surveillance and they are well located, with easy access to public transportation.

The forecast of Singapore condominium prices is important for potential buyers to make informed decisions. With realistic condominium price estimates, potential buyers could acquire information of the housing price trends before executing any transactions.

The objective of this study is to compare the predictive performance of artificial neural networks (ANN) with autoregressive integrated moving average (ARIMA) and multiple regression analysis (MRA) for Singapore condominium prices. The future condominium prices will be predicted using the model which could achieve the highest accuracy in estimating.

Housing prices are a form of time series. Various techniques, such as ANN, ARIMA and MRA, have been used in predicting many types of time series (e.g., Geva, 1998; Wang et al, 2001), including housing price in other parts of the world (e.g., Limsombunchai et al, 2004; Khalafallah, 2008; Nguyen and Cripps, 2009; Hamzaoui and Perez, 2011) and financial markets (e.g., Aussem et al, 1998; Wang and Gupta, 2013; Dong et al, 2013; Fang et al, 2014).

In particular, time series ARIMA models can be used to model relationship between data such as prices, quantities etc. that are collected over time (Yan, et al., 2007). ARIMA (p,d,q) represents an Autoregressive moving average (ARMA) model with p autoregressive lags, q moving average lags, and difference in the order of d (Katchova, 2013). Autoregressive (AR) is model on how the value of a variable, y in at a given time is related to its historical values. Moving average (MA) models examine the relationship between a variable and the residuals from past periods. When a time series variable is not stationary, the variable will be integrated in order of d.

MRA is a traditional, frequently used prediction tool; it has been used in fields such as financial analysis, market policy decision and many more. Dependent variables can be forecasted via independent variables (Sun, 2012). The adjusted R2 is used to examine the significance of the independent variables in influencing dependent variables. The greater an adjusted R2, the more the independent variables contribute to the model.

II. Forecasting the Condominium Price Index (CPI)

The housing prices are influenced by many different factors. The main variables considered in the design of our ANN models are: consumer spending, monthly average wages, gross domestic product (GDP), consumer price index, prime lending rate, real interest rate, population, the Singapore Housing and Development Board (HDB) resale price index, change of HDB resale price index, Straits Times Index, the number of available condominium, and the condominium price index (CPI).

The variables, except the CPI, are the input (independent) variables. CPI is the target (dependent) variable. The time series data used in this study are obtained from the Real Estate Information System (REALIS) of the Singapore Department of Statistics and Trading Economy website. Most of the data are collected quarterly from 1990 to 2013; however, population and annual inflation rate are only available on a yearly basis and we converted them into quarterly basis by cubic spline interpolation (CSI) (Stoer & Bulirsch, 2002). There are a total of 96 time steps for each of the time-series data.

Determination of an ANN structure is a central issue as it will substantially affect the performance of learning and prediction of the resulting networks. Mean square error (MSE) is the average squared difference between the actual and predicted CPI and is used to measure network performance.

Different network structures were tested in order to select the best ANN model that would achieve a good estimate of CPI. Table 1 illustrates the results obtained from the best 5 examined network structures. Model 5 which is highlighted in bold depicts the best network with the smallest MSE.

Wan Teng Lim and Lipo Wang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (email: elpwang@ntu.edu.sg)

Yaoli Wang (corresponding author) and Qing Chang (corresponding author) with College of Information Engineering, Taiyuan University of Technology, Taiyuan, China.

978-1-5090-4093-3/16/$31.00 ©2016 IEEE

In the design of ARIMA model, the best ARIMA model, i.e., ARIMA (2,1,1) with drift, is selected using the auto.arima() function in the R package forecast based on the AIC criterion.

Table 1: Five Best Examined Network Structures

Model	Data Sample Distribution	No. of Hidden Neurons; No. of delays	Training, Validation, Testing	MSE	R
1	60:20:20	10;4	Training	9.23	0.997
			Validation	12.9	0.996
			Testing	19.2	0.994
2	60:20:20	10;6	Training	7.07	0.998
			Validation	22.2	0.993
			Testing	18.1	0.994
3	60:20:20	9;4	Training	2.53	1.000
			Validation	27.5	0.984
			Testing	25.0	0.995
4	70:15:15	9;4	Training	6.44	0.998
			Validation	20.4	0.993
			Testing	18.2	0.990
5	**70:15:15**	**10;4**	**Training**	**8.81**	**0.998**
			Validation	**8.00**	**0.998**
			Testing	**13.1**	**0.993**

Different performance indicators, as shown in Table 2 are used to evaluate the prediction performance of the ANN (Model 5) and the ARIMA model. The mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) of the ANN model is smaller when compared with the ARIMA model. Thus, the ANN model is concluded as the better predictive tool; the future CPI trend will be predicted using the ANN model.

The ANN model has ten hidden neurons in a single hidden layer and there are four time delays. It has a data sample distribution of 70:15:15.

In total, there are 92 predicted values, starting from 1991 Quarter 1 (Q1) to 2013 Q4. The forecasting only starts from the fifth time step as the network uses a "moving window" size of 4.

Table 2: Performance Comparison of ANN and ARIMA.

	MAE	RMSE	MAPE
ANN	1.9	3.1	1.7%
ARIMA	3.2	5.0	2.5%

Figure 1 shows the target and output time series and the errors between them. It illustrates a high accuracy of the forecast as the predicted and actual values are in close agreement. The errors range from -11.6 to 8.3.

Five-time-steps-ahead predictions (2014Q1 to 2015Q1) are shown in Table 3, the results illustrate that ANN is able to produce accurate forecasts for the first three-time-steps-ahead as the percentage of errors only range from 0.66% to 2.24%, as time series forecasting relies highly upon the historical data. Since the future inputs are predicted, the accuracy of predicting CPI future trend will get poorer as time passes.

Table 3: Difference between Actual and Predicted CPI

Period	Actual CPI	Predicted CPI	Error	Error of Percentage
2014Q1	197.7	195.9	1.8	0.91
2014Q2	196.9	195.6	1.3	0.66
2014Q3	196.5	200.9	4.4	2.24
2014Q4	194.4	203.9	9.5	4.89
2015Q1	-	230.9	-	-

When the target is changed to the CPI of each individual region (East, West, Central, and Northeast), the network that is saved in the MATLAB workspace is then used to estimate the CPI of each individual region. Each individual region is trained 20 times as the weight and bias is different every time and an average is computed. Table 4 shows the performance of each individual region.

Table 4: Predictive Performance Results of Each Region

	RMSE	MAPE (%)
Central	2.1	1.3
East	2.4	1.6
North East	2.3	1.6
West	4.5	2.6

Table 4 shows that central region has the lowest RMSE and MAPE. This could be due to the reason that CPI of the central regions is more dependent on the independent variables. The condominiums that are located in the central region tend to have better facilities and are more expensive; hence the CPI values are more dependent on the independent variables. For instance, the economic factors will determine whether the people have sufficient purchasing power to buy the condominiums in the central region, whereas the CPI values of other regions may not be as highly dependent upon the independent variables.

III. Forecasting the Condominium Asking Price (CAP)

The real estate value is based on attributes and location of a particular condominium unit. In this section, the ANN and MRA models use the data gathered from the STProperty website (online resources). The dataset contains 800 samples and these samples only cover condominium units that are posted between January and March 2015.

The target (dependent) variable is the CAP and the other nine variables are the input (independent) variables, as shown in Table 5.

Similarly, several ANN structures were configured and the best 5 examined network structures are shown in Table 6, which shows that the best performed ANN is architecture 9-15-15-1 which is highlighted in bold.

As shown in Table 2, the ANN model can predict the CAP more accurately as it has a lower RMSE and a higher Regression value (R-value) when compared with the MRA model. R value is the correlation between the predicted and actual values. The future CAP will be predicted using the ANN model.

Table 5: Characteristics of the Variables

Variable Name	Definition (measurement)	Range
CAP	Asking price of condominium unit (millions of Singapore dollars or SGD)	0.6-36.0
Bedroom	Number of bedrooms in a unit (numeric)	1.0-8.0
Bathroom	Number of bathrooms in a unit (numeric)	1.0-9.0
Floor Area	Size of a unit (square feet)	263-12000.0
Tenure	Strata tile - binary variable: “0” if lease for 99 years, else “1”	0.0-1.0
Age	Age of a condominium unit (number of years)	1.0-42.0
MRT	Linear distance to the nearest MRT (km)	0.1-3.4
School	Linear distance to the nearest School (km)	0.1-3.9
Shopping Mall	Linear distance to the nearest Shopping Mall (km)	0.1-3.2
Childcare Centre	Linear distance to the nearest Childcare Centre (km)	0.0-3.5

Table 6: Best Five Neural Network Architecture

Data Sample Distribution Ratio: 60:20:20				
Neural Network Architecture	Training MSE	Validation MSE	Testing MSE	Overall R
9-15-1	0.832	0.899	0.657	0.974
9-12-12-1	0.798	0.897	0.725	0.975
9-12-15-1	0.147	0.337	0.439	0.982
9-14-15-1	0.618	0.550	0.916	0.980
9-15-15-1	**0.532**	**0.759**	**0.322**	**0.983**

Table 2: Comparison of MRA and ANN Model

	RMSE	R-value
ANN	0.732	0.983
MRA	2.134	0.716

The ANN model consists of two hidden layers with fifteen hidden neurons in each layer and has a sample distribution of 60:60:20. Figure 2 plots the training, validation and test MSE against the iteration number. The predicted outputs are of an acceptable range. The condominium units in central region may be at the higher end where there are better and more facilities (i.e. spa and karaoke etc.).

IV. Conclusion

This study aimed to predict Singapore condominium prices using an effective predictive tool, i.e., the ANN. First, the ANN model showed its superiority over the ARIMA model in predicting CPI. The forecasts were based on time series data of variables that are believed to influence the condominium prices in Singapore. These variables and the CPI were the inputs and output to the models, respectively. The predicted and actual CPI were in close agreement as the models have the ability to deduce and generalize the relationship between the input and output variables through learning. However, one of the possible limitations in this section is that the size of data set, which consists only of 96 time steps, may not be sufficiently large. Hence, the performance of the ANN is not optimized due to the lack of training data set and insufficient verification output data. Moreover, some of the data for the independent variables such as the population and annual inflation rate are converted from yearly into quarterly basis; the accuracy of the predicted outputs will be affected.

Second, the ANN has proven to be a better predictive tool than MRA in predicting the CAP. The input variables are the housing characteristics and the output variable is the CAP. The high R-values of the ANN model shows a good fit between the independent and dependent variables. The model is able to map the non-linear relationship between attributes of the condominium units (independent variables) and the CAP (dependent variables). However, the best ANN model is determined based on trial-and-error. Hence, it is not possible to determine whether the best result had been generated by the model.

In conclusion, the results show that ANN model can generate high accuracy. In future studies, we shall use more rigorous techniques to select input features (e.g., Fu and Wang, 2003; Chu et al, 2004; Wang et al, 2008) and other predictive models (e.g., Frayman and Wang, 1998; Zhou and Wei, 2010; Majidpour et al, 2015.)

References

A. Aussem, J. Campbell, and F. Murtagh, “Wavelet-based feature extraction and decomposition strategies for financial forecasting,” J. Comput. Intell. Finance, vol. 6, no. 2, pp. 5–12, 1998.

F. Chu and L. P. Wang, “Applications of support vector machines to cancer classification with microarray data,” *International Journal of Neural Systems*, vol.15, no.6, pp.475-484, 2005.

F. Chu, Wei Xie, and L.P. Wang, "Gene selection and cancer classification using a fuzzy neural network", *Proceedings of the North-American Fuzzy Information Processing Conference (NAFIPS 2004),* vol.2, pp.555-559, 2004.

G. Dong, K. Fataliyev, and L. P. Wang, “One-step and multi-step ahead stock prediction using backpropagation neural networks,” in Communications and Signal Processing (ICICS) 2013 9th International Conference on Information, 2013, pp. 1–5.

Y. Fang, K. Fataliyev, L.P. Wang, X.J. Fu and Yaoli Wang, "Improving the genetic-algorithm-optimized wavelet neural network approach to stock market prediction," 2014 International Joint Conference on Neural Networks (IJCNN 2014), pp. 3038-3042.

Y. Frayman and L. P. Wang, "Data mining using dynamically constructed recurrent fuzzy neural networks," Research and Development in Knowledge Discovery and Data Mining, vol. 1394, PAKDD'98 (Regular Paper), pp. 122-131, 1998.

X.J. Fu and L. P. Wang, "Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance", *IEEE Trans. Systems, Man, Cybern, Part B-Cybernetics,* vol.33, no.3, pp. 399-409, 2003.

A. B. Geva, "ScaleNet-multiscale neural-network architecture for time series prediction," IEEE Trans. Neural Netw., vol. 9, no. 6, pp. 1471–1482, Nov. 1998.

Y.Q. He, K. Fataliyev, and L. P. Wang, "Feature selection for stock market analysis," the 20th International Conference on Neural Information Processing (ICONIP2013), Daegu, Korea, 3-10 November 2013, Invited Paper, Part II, LNCS 8227, pp. 737–744, 2013.

Government of Singapore, Department of Statistics, http://www.singstat.gov.sg/statistics/latest_data.html, 31 Oct, 2014.

S. Gupta and L. P. Wang, "Stock Forecasting with Feedforward Neural Networks and Gradual Data Sub-Sampling," Australian Journal of Intelligent Information Processing Systems, vol.11, pp.14-17, 2010.

Y. Hamzaoui, J. Perez, "Application of Artificial Neural Networks to Predict the Selling Price in the Real Estate Valuation Process," 2011 10th Mexican International Conference on Artificial Intelligence (MICAI), pp.175 – 181, 2011.

A. Katchova, "Time Series ARIMA Models", https://drive.google.com/file/d/0BwogTI8d6EEiaDJCRXd0dmU1ZDA/edit?pli=1, 2013.

A. Khalafallah, "Neural network based model for predicting housing market performance," Tsinghua Science and Technology, vol.13, no.S1, pp.325 – 328, 2008.

V. Limsombunchai, C. Gan, M. Lee, "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network," American Journal of Applied Sciences, vol.1, pp.193-201, 2004.

B. Liu, C.R. Wan, and L. P. Wang, "An efficient semi-unsupervised gene selection method via spectral biclustering", *IEEE Trans. Nano-Bioscience*, vol.5, no.2, pp.110-114, June, 2006.

M. Majidpour, C. Qiu, P. Chu, R. Gadh, H.R. Pota, "Fast Prediction for Sparse Time Series: Demand Forecast of EV Charging Stations for Cell Phone Applications," IEEE Transactions on Industrial Informatics, vol.11, pp.242-250, 2015.

N. Nguyen, A. Cripps, "Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks," Journal of Real Estate Research, vol. 22, no. 3, pp.313-336, 2009.

STProperty Condo Search, http://www.stproperty.sg/condominium-directory/top-from-1970/top-to-2014/sort-top-asc/page1/box-1

J. Stoer, R. Bulirsch, Introduction to Numerical AnalysisNew York: Springer Science, 2002.

B. Sun, "A Study about the Prediction of University Library Lending Based on Multiple Regression Analysis," In B. Sun, Advances in Automation and Robotics, vol.1, pp.525-532, Springer Berlin Heidelberg, 2012.

K. K. Teo, L. P. Wang, and Z. Lin, "Wavelet Packet Multi-layer Perceptron for Chaotic Time Series Prediction: Effects of Weight Initialization," in Computational Science - ICCS 2001, V. N. Alexandrov, J. J. Dongarra, B. A. Juliano, R. S. Renner, and C. J. K. Tan, Eds. Springer Berlin Heidelberg, 2001, pp. 310–317.

L. P. Wang, "Learning and retrieving spatio-temporal sequences with any static associative neural network," *IEEE Trans. Circuit and Systems-II: Analog and Digital Signal Processing,* vol. 45, no.6, pp. 729-738, June, 1998.

L. P. Wang, F. Chu, and W. Xie, "Accurate cancer classification using expressions of very few genes," *IEEE-ACM Trans. Computational Biology and Bioinformatics*, vol.4, no.1, pp. 40-53, Jan.-March, 2007.

L. P. Wang and X. J. Fu, Data Mining with Computational Intelligence. Berlin/Heidelberg: Springer-Verlag, 2005.

L. P. Wang and S. Gupta, "Neural Networks and Wavelet De-Noising for Stock Trading and Prediction," in Time Series Analysis, Modeling and Applications, W. Pedrycz and S.-M. Chen, Eds. Springer Berlin Heidelberg, 2013, pp. 229–247.

L. P. Wang, K.K. Teo, and Z.P. Lin, "Predicting time series with wavelet packet neural networks", 2001 IEEE International Joint Conference on Neural Networks (IJCNN 2001), pp.1593-1597, 2001.

L. P. Wang, Nina Zhou, and Feng Chu, "A general wrapper approach to selection of class-dependent features," *IEEE Trans. Neural Networks*, vol.19, no.7, pp.1267-1278, 2008.

T.-C. Wong, A. Yap, "From universal public housing to meeting the increasing aspiration for private housing in Singapore," Habitat International, vol. 27, pp. 361–380, 2003.

Y. Yan, W. Xu, H. Bu, Y. Song, W. Zhang, H. Yuan, S.-Y. Wang, "Method for Housing Price Forecasting based on TEI@I Methodology," Systems Engineering - Theory & Practice , pp.1-9, 2007.

T. Zheng, K. Fataliyev, and L. P. Wang, "Wavelet neural networks for stock trading and prediction," SPIE Defense, Security, and Sensing, 29 April - 3 May 2013, Baltimore, USA, vol.8750, 0A (8750-9).

H. Zhou and Y. Wei, "Stocks market modeling and forecasting based on HGA and wavelet neural networks," in 2010 Sixth International Conference on Natural Computation (ICNC), 2010, vol. 2, pp. 620–625.

N. Zhou and L. P. Wang, "Effective selection of informative SNPs and classification on the HapMap genotype data," *BMC Bioinformatics,* 8:484, 2007.

M. Zhu and L. P. Wang, “Intelligent trading using support vector regression and multilayer perceptrons optimized with genetic algorithms,” in The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1–5.

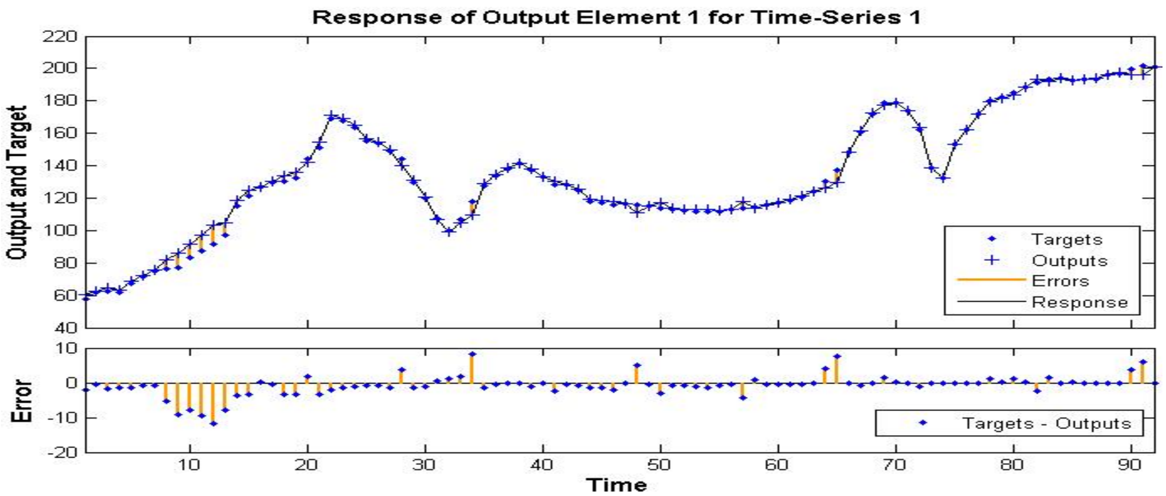

Figure 1: Time Series Response

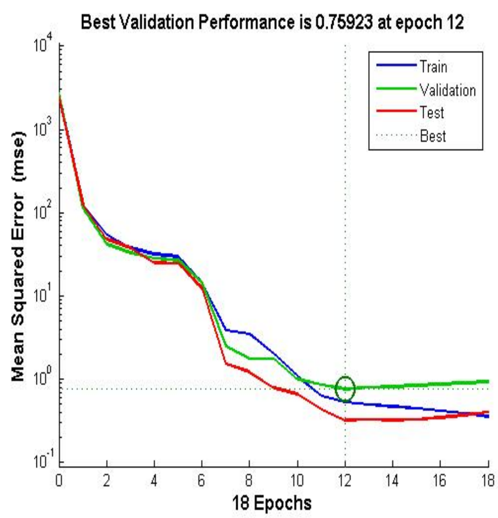

Figure 2: Training, validation, and test errors of the best neural network at different epochs.