Prediction of House Unit Price in Taipei City Using Support Vector Regression

Tse-Wei Lee
Department of Industrial Engineering and Management
Ming Chi University of Technology, New Taipei City, Taiwan

Kuentai Chen*
Department of Industrial Engineering and Management
Ming Chi University of Technology, New Taipei City, Taiwan
Tel: (+886) 2-29089899-3111, Email: kuentai@mail.mcut.edu.tw

Abstract. Buyers of houses often determine the prices basing on incomplete information, and might be misled by subjective opinions from owners and real estate brokers. In order to provide a rational price for the buyers' reference, we tried to predict the prices of housing units by including both crisp and fuzzy variables for model building. The data are collected from Taipei City's official webpage, which is just available online, and the targets are limited to pre-owned units in buildings of seven floors or more. Two neural network methods, namely Support Vector Regression (SVR) and Adaptive Network-Based Fuzzy Inference System (ANFIS), are applied to build the model and the results are compared. The results show that instead of using all data in Taipei City for the models, SVRs with data of each individual district are recommended.

Keywords: SVR, Neural Networks, ANFIS, Real Estate, House unit price

1. Introduction

For various purposes, valuation of real estate is essential to the businesses and Government, as well as to families who desire to own residences. Financial bureaus such as banks valuate lands and properties for loans or mortgages; the Government appraises those to levy taxes; and individuals ask around to determine the housing prices to trade. Nevertheless, the focal point of real estate appraisal for individuals is the trading prices but not the worth of the property, which is different from the value of the property. In this study we concentrate on the trading prices of those housing units.

Purchasing a house is literately an important event for most families. Buyers of houses often determine the prices basing on partial information, and might be misled by subjective opinions from owners and real estate brokers. In order to provide a rational price for the buyers' reference, we tried to provide rational prices of housing units by including both subjective and objective variables for model building. The objective variables included are measured data collected from official database of the Taiwan Government, and these are viewed as crisp variables. The subjective variables included are collected by interviewing senior real estate brokers, and these are considered as fuzzy variables in this study.

2. Literature review and research methods

There are usually two kinds of real estate pricing, mainly on mass appraisal and valuation of individual real estate property. For the Government, mass appraisal is commonly used to calculate how much to tax, it is the systematic appraisal of groups of properties as of a given date using standardized procedures and statistical testing (Kontrimas & Verikas, 2011). These models are often based on sales comparison approaches, which traditionally use regression analysis.

Since the construction of Hedonic price indices, which uses objectively measured characteristics for product prices by Rosen (1974), many researchers implemented this method in real estate domain and sometimes compared it with other methods, such as Butler (1982), Clapp & Giaccotto (1998), and Cohen & Coughlin (2008). There are some works done using the Hedonic Equation in the Real Estate Market in Taiwan, such as Lin & Ma (2007).

Some later works focus on soft computing algorithms, which can usually provide satisfactory results in a short time and sometimes optimum. Some of these works focused on building a model that can precisely predict the price of a single property such as a house or a unit in an apartment. As the traditional predicting methods for the housing unit price, building statistical regression models is the most common way with acceptable results. Nevertheless, the regression models are sometimes confined by certain function and thus do not always provide desired solutions (Tay & Ho, 1992; Do & Gary, 1992). On the other hand, soft computing methods such as Back Propagation Neural Networks (BPNN) and Support Vector Regression (SVR) have been demonstrated to be advanced when dealing with nonlinear problems (Tsoukalas & Uhrig, 1997; Kuan & White, 1994; Lin & Lee, 1996), where strict pre-assumptions are not required.

In previous research, we can find out that various NNs have become popular and useful for housing price problems with enough historical data. For instance, Do and Gary (1992) compared NN and multi-regression analysis by predicting real estate prices in United States and find out that NN outperformed Regression model with smaller mean absolute error (MAE). For their application, there are 105 samples and MAE is 6.9% for NN while MAE is 11.3% for Regression. Tay and Ho (1992) also used the same two methods for comparison and concluded that NN performed better. They had 822 training data sets and 222 testing data sets for the models predicting prices of Singapore Departments, resulted MAE 3.9% for NN and 7.5% for regression. Furthermore, McCluskey et al. (1997) suggested NN to be a superior method as they took 416 housing prices in Northern Ireland to train NN and resulted in MAE of 7.75%. More researches with similar conclusions can be found in McGreal et al. (1998) and Wong et al. (2001), with targets of Belfast in U.K. and Hong Kong, respectively. All these papers agreed that NNs are valid methods in predicting housing prices. Detailed description for neural networks can be found in Hornik (1989) and a successful application of BPNN in finance can be found in Chen et al. (2009).

Somehow, many financial applications used BPNN, and some others implemented various types of NNs to improve the results. In fact, Support Vector Regression (SVR) and Adaptive Neuro-Fuzzy Inference System (ANFIS) can provide desired outcome in many occasions and can be classified as NNs. Both of these methods are multi-layer neural networks that consist of three typical NN elements: architecture with connected neurons, learning or training algorithms, and activation functions. In this study, ANFIS and SVR are implemented to perform the training and predictions as these two neural networks have been widely implemented in many fields including finance. Since these methods are now widely applied in many domains and the length of conference paper is limited, only brief review and references are provided below.

2.1 Support Vector Regression

Support Vector Regression is proposed by Vapnik et al. (1997), it is constructed based on Support Vector Machine (SVM), which was proposed by Vapnik (1995). Based on statistical learning theory, SVM is developed to solve the classification problem. SVM gained popularity rapidly due to many striking features and promising performance in real-world applications. SVM performs classification by constructing an N-dimensional hyper-plane that separates the data optimally into two categories. SVM is also close to neural networks when using a sigmoid kernel function, which makes SVM equivalent to a two-layer, perceptron neural network. SVM was first applied to pattern recognition (Schmidt, 1996), and then modified for regression estimation of signal processing (Vapnik et al., 1997). A tutorial for SVM in pattern recognition can be found in Burges (1998).

By introducing ε -insensitive loss function including a distance measure, Support Vector Machine is able to deal with nonlinear regression problems and such algorithm is thus named as Support Vector Regression Machine, usually called SVR. The structure of SVR is basically a three-layer neural network, the SVR model produced by support vector classification depends on a subset of the training data. SVR model can also cope with nonlinear data and has only a unique optimal solution with each set of kernel parameter and soft margin parameter. Therefore, other than the training process the problem becomes to search the best combination of parameters, in our study, SVR is combined with Grid algorithm for the best combination of parameters. For detailed description of SVR, the readers are referred to a tutorial in Drucker et al. (1997), and a review in Basak et al. (2007).

SVR minimized both empirical risk and structural risk and thus guarantees the existence of one unique optimal solution while avoiding over-fitting problems. This advantage entitles SVR the success in many applications such as identification, prediction, clustering, and so on. Putting the focus on financial fields, we can find a plenty of successful examples, such as stock prediction in U.S. (Trafalis & Ince, 2000), and financial time series forecasting for Futures contract (Tay & Cao, 2001). Readers are also referred to Holimchayachotikul et al. (2011) for a successful SVR application.

2.2 Adaptive Neuro-Fuzzy Inference System

Another popular algorithm, namely Adaptive Neuro-Fuzzy Inference System (ANFIS), was proposed by Jang (1993), it combines structure of neural networks and the fuzzy inference system in order to acquire strengths from Neural networks and fuzzy logic with learning ability and linguistic interpretation, respectively. ANFIS maps inputs through input membership functions and associated parameters, and then through output membership functions and associated parameters to outputs. Hybrid Learning Rule and Least Squares Estimate are used in this five-layer feed-forward network. Numerous successful ANFIS applications are reported since it was integrated in MATLAB as a toolbox. ANFIS has been cited by more than 7000 papers or reports as of 2016, therefore readers are referred to the book published by Jang et al. (1997).

2.3 Error measure

Two popular error measures, namely Mean Absolute Percentage Error (MAPE) and Coefficient of Determination (R-square), are employed as the indices of efficiency to compare these two methods. R-square was used mainly in pre-test using BPNN, and MAPE was applied for method comparison. The equation of MAPE is as shown in the following:

$$MAPE = \frac{\sum_{k=1}^{N} \left| \frac{d_k - y_k}{d_k} \right|}{N} \qquad \text{--EQ(1)}$$

where d_k is the k th target value; y_k is the k th output; and N is the total number of data. And the equation to calculate R^2 is as shown in EQ(2).

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{N} (t_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \qquad \text{--EQ(2)}$$

where y_i is the i^{th} target value, t_i is the i^{th} output, and N is the total number of data sets.

3. Data, Variables and Analysis

3.1 Data and Crisp variables

The Taiwan government established an open webpage for real estate traders to register the transaction prices of the trading property (Dept of Land Administration, M. O. I., 2016), and the information can be accessed by anyone. The responsible party such as real estate agent or property owner must register the location of the property, price or rent value, and information on the property itself. The information on the property consist of address, trading date, total price, price per ping, total floor area in ping, housing pattern, and so on. For residence, there are various styles of houses such as studio, condominium, flat, and mansion. As different housing patterns may lead to different prices, we concentrate on pre-owned house units in buildings that are higher than eleven floors with elevator. The collected data is from January to December 2013, with total of 3336 trading records. Luxury houses with high prices are excluded using k-means clustering in this study and resulted with 3049 records. Table 1 is the descriptive Statistics for collected data without outliers. The numbers of collected data for each district are listed in Table 2. As shown in Table 1, the resulted total prices of housing units are below 33 million New Taiwan dollars and total area are less than 85 Pings.

Table 1. Descriptive Statistics for collected data

# of data: 3049	Min	Max	Range	Mean	STD
Total Floor Area (Ping)	1.1	84.73	83.63	24.41	11.96
RSLDI*(Ping)	0.05	25.46	25.41	3.59	2.96
House Age(Year)	1	44	43	19.07	11.45
Total Number of Floors	11	27	16	13.55	2.55
Floor Number	1	21	20	7.42	3.85
Total Price (10k)	39.5	3230.0	3190.5	1471.2	745.1
Price Per Ping (10k)	1.9	184.95	1278.1	62.59	21.43

* Rights of shared land divide individually

In the collected data sets, we select numerical factors as crisp variables, including *Total Floor Area*, *Rights of Shared Land Divide Individually, House Age, Total Number of Floors, and Floor Number*. A stepwise regression is performed to observe if there exists linear and cross-product relationship between variables and *Total Price*. As the result, only *Total Floor Area* is selected with adjusted R-square of 0.525, setting 0.1 R-square improvement as the criterion to the addition of each variable. The adjusted R-square with all five variables is 0.548, which is not satisfactory for a prediction model. We suspect that possible causes may be 1) not adequate to pool all data since there are twelve district and/or 2) some important variables may not be included in the model. Therefore, we introduced fuzzy variables and tested our models with data from each district.

3.2 Fuzzy variables

As stated above, there may be some other factors affecting the housing prices, such as *distance to metro station*, *distance to school*, *security*, *distance to park*, *neighborhood*, and so on. The common feature of these factors is they are more or less imprecise. Taking *the distance to park* as example, a difference of 5, 10, even 50 meters would not make the price higher or lower. In other words, it can be represented by linguistic terms such as far, near, or a little bit far. This is the domain where fuzzy logic can usually be useful.

Among these factors, attributes of houses and location of houses are considered important variables to determine the trading prices (Butler, 1982; Huh & Kwak, 1997; Meese & Wallace, 1997; Kiel & Zabel, 2008). Lin and Lin (1993) pointed out that other than the features of the house itself, the environment quality and nearby public facilities also influent the trading prices. Some research such as Can (1990), Ioannides (2002), and Ioannides & Zabel (2003) support that other than structural characteristics of house units, the neighborhood characteristics also affect the living quality. Furthermore, Hardman & Ioannides (1998) show that the socioeconomic characteristics of neighbors in the cluster are highly correlated. In summary, we can conclude that both "crisp" variables and "fuzzy" variables should be considered in the model.

According to Jhuang (2013), we compound the variables other than crisp variables introduced in section 3.1 into three fuzzy variables, namely *Environmental condition*, *Living function*, and *Economic potential*. The first fuzzy variable *Environmental condition*, as the name states, contains environmental conditions such as neighborhood, air quality, noise, safety, public facilities, and "not in my backyard phenomenon". The second fuzzy variable *Living function* is related to life convenience like distances to public facilities (Oh & Jeong, 2002; Andriantiatsaholiniaina et al., 2004). The third fuzzy variable *Economic potential* is about the financial considerations such as the trend of trading prices in surrounding area, city planning, and urban renewal. In this study, fuzzy variable is equipped with three to five membership functions.

3.3 Collection of Fuzzy data

A pre-assumption of our model is that in 2013 there is no drastic change at housing prices in Taipei City.

Before the collection of fuzzy data, we first consulted with several real estate brokers with more than three years experiences on how to divide each district into blocks. We assume that there will be similar fuzzy attributes in each block, that is, if the *living function* of some house unit is Good, the nearby house units are also Good in the same block. Dividing of district differs due to the different city development history. As shown in Figure 1 and Figure 2, some of the divided blocks are rectangle or close to rectangle in shape, which may contain some street blocks. In Figure 1 and Figure 2, Da-An District and Datong District are divided into 18 and 15 blocks respectively. Many dividers of other districts may be curve along with rivers, MRT route, or mountains, such as Shilin District and Beitou District. In Figure 3 and Figure 4, Shilin District and Beitou District are divided into 7 and 9 blocks respectively. The detailed dividing process and the dividing of the other 8 districts will not be displayed here due to the page limit.

The divided map was provided to five experts for each district, with total of 60 experts. The experts rate each block with three overall scores according to their experiment and judgement, in corresponding to three fuzzy variables. Trading records were then combined with fuzzy scores and resulted with five crisp variables and three fuzzy variables. As the number of data sets in each district is relatively small for training NNs, we randomly reserved only 10 data sets in each district for testing and use the rest for training.

4. Results and Conclusion

The results of SVR and ANFIS predictions are shown in Table 2. From Table 2 we can observe that the best testing outcome is generated by ANFIS with MAPE of 6.01% for Nangang District, where there are only 101 samples. Figure 5 and Figure 6 show the SVR and ANFIS predictions for Nangang District using testing data, respectively. On the other hand, the worse predictions are made by also ANFIS with 17.71% for Zhongzheng District, where there are 153 samples. Figure 7 and Figure 8 show the SVR and ANFIS predictions for Zhongzheng District using testing data, respectively. It can be observed that even with the worse MAPE among all districts, the predictions still follow the targets well. This suggests that the sample sizes are not the major cause of predictions.

Note that the training error of ANFIS for the best testing result is 6.01% with training MAPE of 7.29%, as the maximum training MAPE is 16.88% and the minimum training MAPE is 4.78% for all districts. That is, better training does not guarantee better testing in these cases.

The predictions of housing prices in Nangang District are obviously better than in other districts for both SVR and ANFIS. A possible reason may be that most of the trading records of house units in Nangang District have smaller age, that is, they are relatively new and thus are more consistent in price.

From above we can conclude in the following:

1) Both objective data such as structure characteristics of house itself and subjective data such as neighborhood and transportation convenience should be considered in the model. Fuzzy logic does help when converting human opinions to scores.
2) Using data from each district for prediction generates better result than pooling all data in Taipei City. Outliers should be deleted first.
3) SVR is recommended to predict pre-owned housing prices in Taipei City rather than ANFIS.

We suggest to use SVR for prediction of

housing prices in Taipei City based on the following reasons: 1) ANFIS generated better predictions than SVR in three districts, while SVR performed better than ANFIS in 9 districts. 2) The average MAPE for SVR (12.05%) is smaller than average MAPE for ANFIS (14.09%).

Figure 1. Dividing Da-An District into 18 blocks

Figure 2. Dividing Datong District into 15 blocks

Figure 3. Dividing Shilin District into 7 blocks

Figure 4. Dividing Beitou District into 9 blocks

Table 2. The results of SVR and ANFIS predictions

District	Method	SVR		ANFIS	
		Training	Testing	Training	Testing
Shilin	98 Samples	7.48%	**11.12%**	14.15%	14.86%
Datong	160 Samples	8.29%	**13.15%**	7.93%	14.87%
Da-an	285 Samples	8.63%	**11.54%**	9.57%	17.40%
Zhongshan	910 Samples	12.19%	**12.40%**	16.70%	14.38%
Zhongzheng	153 Samples	16.21%	**13.34%**	11.73%	17.71%
Neihu	231 Samples	11.36%	**10.64%**	10.75%	14.62%
Wenshan	187 Samples	7.20%	12.94%	4.78%	**12.56%**
Beitou	164 Samples	8.98%	**12.64%**	7.43%	13.45%
SongShan	316 Samples	10.72%	**13.67%**	11.68%	15.57%
Xinyi	216 Samples	10.46%	**9.21%**	9.36%	12.24%
Nangang	101 Samples	6.45%	7.29%	7.29%	**6.01%**
Wanhua	228 Samples	14.10%	16.69%	16.88%	**15.43%**
Average	Total of 3049 Samples	10.17%	***12.05%***	10.69%	14.09%

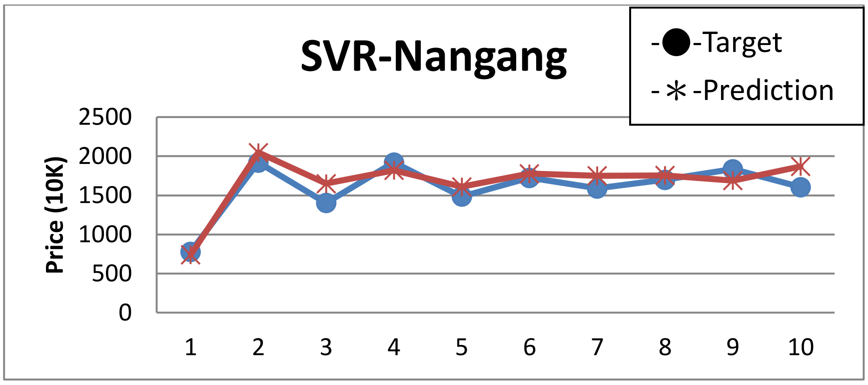

Figure 5. SVR predictions for Nangang District (testing data)

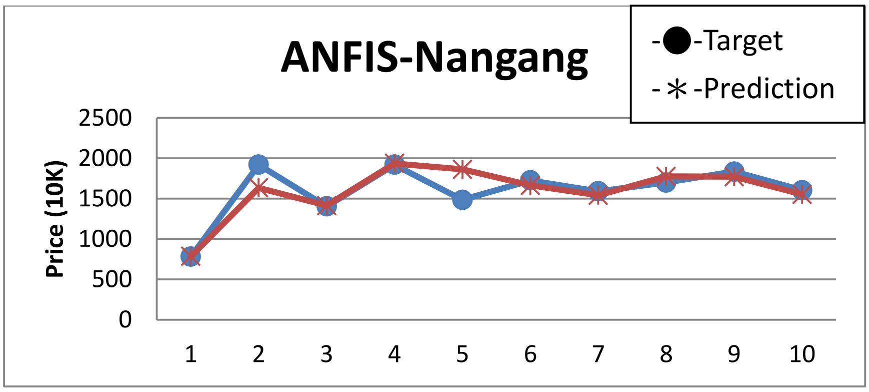

Figure 6. ANFIS predictions for Nangang District (testing data)

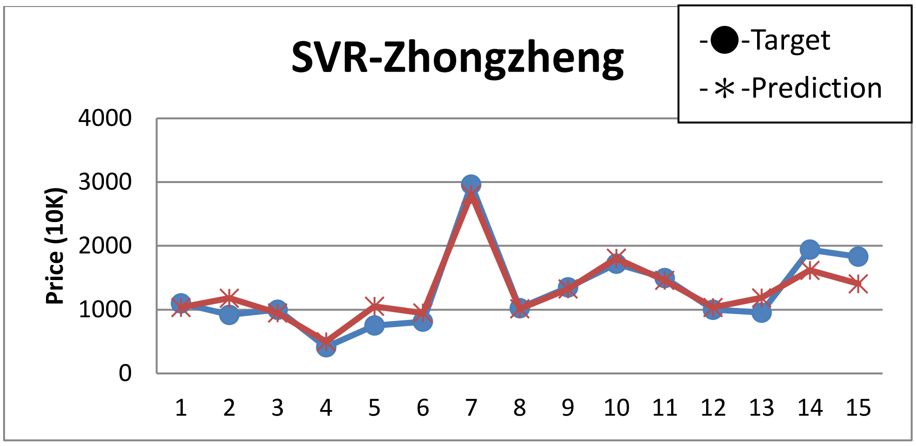

Figure 7. SVR predictions for Zhongzheng District

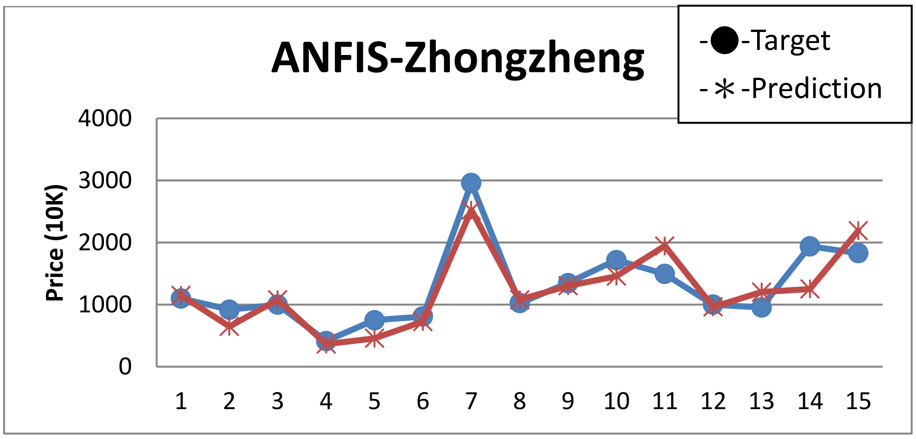

Figure 8. ANFIS predictions for Zhongzheng District

Reference

Andriantiatsaholiniaina, A., Vassilis, S. and Yannis, A. (2004) Evaluating Strategies for Sustainable Development: Fuzzy Logic Reasoning and Sensitivity Analysis, *Ecological Economics.* 48(2): 149-172.

Burges, C.J.C. (1998) A tutorial on support vector machines for pattern recognition, *Knowledge Discovery and Data Mining* 2(2):121-167

Butler, R.V. (1982) The Specification of Hedonic Indexes for Urban Housing, *Journal of Land Economics.* 58(1): 96-108.

Can, A. (1990) The Measurement of Neighbourhood Dynamics in Urban House Prices, *Journal of Economic Geography.* 66(3): 254-272.

Chen K.T., Lin, H.-Y. & Huang, T.-C. (2009) The Prediction of Taiwan 10-Year Government Bond Yield, *WSEAS Transactions on Systems*, 9(8): 1051-1060.

Clapp, J.M. & Giaccotto, C. (1998) Residential Hedonic Models: A Rational Expectations Approach to Age Effect, *Journal of Urban Economics.* 44(3): 415-437.

Cohen, J. P. & Coughlin, C. C. (2008) Spatial Hedonic Models of Airport Noise, Proximity, and Housing Prices, *Journal of Regional Science.* 48(5): 859-878.

Department of Land Administration, M.O.I. (2016) Actual Price Registration http://lvr.land.moi.gov.tw/N11/homePage.action

Do, Q. & Gary, G. (1992) A Neural Network Approach to Residential Property Appraisal, *Real Estate Appraiser.* 58(3):38-45.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A. & Vapnik, V. (1997) Support vector regression machines, *Advances in Neural Information Processing Systems* 9: 155-161. Cambridge, MA, MIT Press.

Hardman, A.M. & Ioannides, Y.M. (1999) Residential Mobility and the Housing Market in a Two-sector Neoclassical Growth Model, *The Scandinavian Journal of Economics*. 101: 315–335.

Holimchayachotikul, P., Leksakul, K., & Guizzi, G. (2011) Robust Design for Etching Process Parameters of Hard Disk Drive Slider Fabrication Using Data Mining and Multi Response Optimization, *WSEAS Transactions on Systems and Control.*

1(6):15-24.
Hornik, K., Stinchcombe, M. & White, H. (1989) Multilayer feedforward networks are universal approximators, *Neural Networks*. 2(5):359-366
Huh, S. & Kwak, S.J. (1997) The Choice of Functional Form and Variables in the Hedonic Price in Seoul, *Urban Studies*. 34(7): 989-998.
Ioannides, Y.M. (2002) Residential Neighborhood Effects, *Regional Science and Urban Economics.* 32(2): 145-165.
Ioannides, Y.M. & Zabel, J.E. (2003) Neighborhood Effects and Housing Demand, *Journal of Applied Econometrics*. 18(5): 563-584.
Jang, J. S. (1993). ANFIS: Adaptive- network-based fuzzy inference system, *IEEE Transactions on Systems, Man and Cybernetics*. 23(3):665-685.
Jang, J.S. R., Sun, C.T. & Mizutani, E. (1997) Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence, Ch 12, Prentice Hall.
Jhuang, J.-S. (2013) ANFIS Prediction of Pre-owned Housing Prices in Taipei City, Master Thesis of Ming Chi University of Technology, Department of Industrial Engineering and Management
Kiel, K.A. & Zabel, J.E. (2008) Location, Location, Location: The 3L Approach to House Price Determination, *Journal of Housing Economics*. 17(2):175-190.
Kontrimas, V. & Verikas, A. (2011) The mass appraisal of the real estate by computational intelligence, *Applied Soft Computing*. 11(1): 443-448.
Kuan, C.M. & White, H. (1994) Artificial Neural Networks: An Econometric Perspective. *Econometric Reviews*. 13(1):1-91.
Lin, C. T. & Lee, C.S. G. (1996) Neural Fuzzy Systems: a neuro-fuzzy synergism to intelligent systems, New Jersey: Prentice Hall.
Lin, C. C. & Lin, S. J. (1993) An Analysis of the Effect of Environment Quality and Public Facilities on Housing Prices and Rents in Taiwan, *Journal of Housing Studies*. 1: 21-45.
Lin, C.C. & Ma, Y.C. (2007) An Application of Mass Appraisal and the Hedonic Equation in the Real Estate Market in Taiwan, *Journal of Housing Studies*. 16(2): 1-22.
Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy*. 82(1): 34-55.
Meese, R.A. & Wallace, N.E. (1997) The Construction of Residential Housing Price Indices: A Comparison of Repeat Sales, Hedonic-regression, and Hybrid Approaches, *Journal of Real Estate Finance and Economics*. 14(1-2): 51-73.
McCluskey, W.J., Dyson, K., McFall, D. & Anand, S. (1997) The Mass Appraisal of Residential Property in Northern Ireland. Computer Assisted Mass Appraisal: an international review 3, ed. By McCluskey, W.J. & Adair, A.S. 59-77. Aldershot: Ashgate.
Oh, K. & Jeong, Y. (2002) The Usefulness of the GIS-fuzzy Set Approach in Evaluating the Urban Residential Environment, *Environment and Planning B: Planning and Design*. 29(4): 589-606.
Schmidt, M. & Gish, H. (1996) Speaker identification via support vector classifiers, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP-96. 1:105-108.
Tay, F.E. & Cao, H.L. (2001) Application of support vector machines in financial time series forecasting, *Omega*. 29(4):309-317.
Tay, D. P.H. & Ho, D. K.H. (1992) Artificial Intelligence and the Mass Appraisal of Residential Apartments, *Journal of Property Valuation and Investment*. 10(2):520-540.
Trafalis, T.B. & Ince, H. (2000) Support Vector Machine for Regression and Applications to Financial Forecasting, Proceedings of the IEEE- INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00), 6:6348-6353.
Tsoukalas, L.H. & Uhrig, R.E. (1997) Fuzzy and Neural Approaches in Engineering, New York: Wiley and Sons.
Vapnik, V.N. (1995) The nature of statistical learning theory. New York: Springer.
Vapnik, V.N., Golowich, S. & Smola, A. (1997) Support vector method for function approximation, regression estimation, and signal processing, *Advances in Neural Information Processing Systems* 9, ed. by Mozer, M., M. Jordan and T. Petsche, pp. 281–287, Cambridge, MA: MIT Press.
Wong, K.C., So, A. T.P., & Hung, Y.C. (2002) Neural Network vs. Hedonic Price Model: Appraisal of High-density Condominiums, *Real Estate Valuation Theory*, ed. by Wang, Ko, and Wolverton, M.L. pp181-198, Boston, Mass.: Kluwer Academic.