International Journal of Technology (2013) 1: 63-72
ISSN 2086-9614

PRELIMINARY COST ESTIMATION USING REGRESSION ANALYSIS INCORPORATED WITH ADAPTIVE NEURO FUZZY INFERENCE SYSTEM

Yusuf Latief[1], Andreas Wibowo[2], Wisnu Isvara[1*]

[1] *Department of Civil Engineering, Universitas Indonesia*
[2] *Agency for Research and Development, Ministry of Public Works*

(Received: December 2011/ Revised: September 2012/ Accepted: January 2013)

ABSTRACT

Preliminary cost estimates play an important role in project decisions at the beginning of design phase of construction project under a limited definition of scope and constraints in available information and time. This study proposes a new approach of preliminary cost estimation model using regression analysis incorporated with adaptive neuro fuzzy inference system (ANFIS). Regression analysis (RA) is used for determination of the significant parameters as input variables in ANFIS model. Datasets of 55 low-cost apartment projects in Indonesia were compiled to demonstrate the advantage of the proposed method. The mean absolute percent error (MAPE) of testing data of the proposed model is 3.98% whereas the MAPE of RA and neural network (NN) models are, respectively, 6.92% and 10.12%, thus indicating better accuracy performance of the proposed model over the latter ones.

Keywords: Preliminary Cost Estimates; Regression Analysis; Neuro Fuzzy

1. INTRODUCTION

Cost estimates evolve with the project life cycle, and can be prepared using a variety of methods appropriate to the level of information and the time available to prepare the estimate (Dell'Isola, 2002). The early cost estimates in construction projects are extremely important to the initial decision-making process by the owner's organization and the project team. At the first phase of design, the schematic design will be prepared and a preliminary estimate can be made as the schematic design develops. The objectives of the preliminary estimate are designing the project within the owner's budget and evaluating alternative design concepts (Pratt, 2011).

The main challenge in early cost estimates often centers on inaccuracy issues. According to Holm et al. (2005), the expected accuracy range of cost estimating at schematic design stage (budget estimate) is ±10-20%, while AACE 18R-97 (2005) states that the accuracy range of budget estimate is ±10-30%. The low accuracy is anticipated because the estimation exercises are based upon limited data and information available at the time of preparing estimates. The most important factor influencing estimate accuracy at the planning stage is the level of information available about the project i.e. the project scope (Ciraci & Polat, 2009). Project information such as the physical and functional characteristics of a project are the important parameters in cost estimation modeling. In the context of the Indonesian construction industry, for instance, Wibowo & Wuryanti (2007) establish a regression-based parametric model using building areas as a single parameter to predict construction cost.

[*] Corresponding author's email: wisnu.isvara@gmail.com, Tel. +62-21-3907401, Fax. +62-21-3102928

Abundant studies to improve small accuracies of early cost estimates have been done and are well documented in extant literature. Statistical methods have traditionally been used to develop cost estimating models, while regression analysis (RA) represents a common alternative. Artificial intelligence approaches are applicable to cost estimating problems related to expert systems, case based reasoning (CBR), neural network (NN), fuzzy logic (FL), genetic algorithms (GA) and derivatives of such (Cheng et al., 2010). Important factors in choosing a construction cost estimating model include a satisfactory degree of accuracy, speed and ease of use, ease in updating, clarity of explanation in construction cost estimations, and consistency in variables stored for long-term use (Kim & Kim, 2010).

This paper presents a preliminary cost estimation model incorporating regression analysis (RA) and the adaptive neuro fuzzy inference system (ANFIS) to improve the accuracy of estimation. RA is used for determination of the significant parameters as input variables in ANFIS model. Given that inclusion of insignificant parameters in the model could lead to a poor prediction outcome, elimination of insignificant parameters may improve the prediction performance of the model (Sonmez & Ontepeli, 2009). ANFIS model is one of the best tradeoffs between neural and fuzzy systems. Fuzzy systems are effective in representing explicit, but ambiguous common sense knowledge, whereas neural networks provide excellent facilities for approximating data, learning knowledge from data, approximating reason, and parallel processing. Regarding the advantages and capabilities of each technique, the proposed model is expected to be more accurate. The approach to hybridize two popular methods featured by the model may contribute to the existing body of knowledge on construction cost estimation. To the best of authors' knowledge, no similar approach has been reported in existing technical articles. Hence, the proposed model may offer an alternative methodology of providing projections for a project's expected cost at early stage of its lifecycle. To demonstrate the model application, it will be tested on case study data and its resulting accuracy level will also be compared with other competing models i.e. traditional RA and NN models.

2. METHODOLOGY

As with traditional parametric estimating cost models, the present model also employs multiple RA on collected historical data to define significant building parameters as key cost drivers. The use of stepwise regression is recommended to address multicollinearity issues among input variables (Ji et al., 2010) that are common in statistical data analysis. The output of this step will be the input for the ANFIS model.

The purpose of the ANFIS model is multifold. It is likely that at the time of preparing cost estimates some accurate data of cost drivers are not yet available. Hence, the fuzzy inference systems implemented in the model allow vague, ambiguous, uncertain, imprecise, noisy, or missing input data. The data are expressed in linguistic terms (e.g., very low, low, medium, high, and very high) that are then translated into the so-called fuzzy numbers (FNs). A FN is defined by a specified membership function (MF), (See Figure 1).

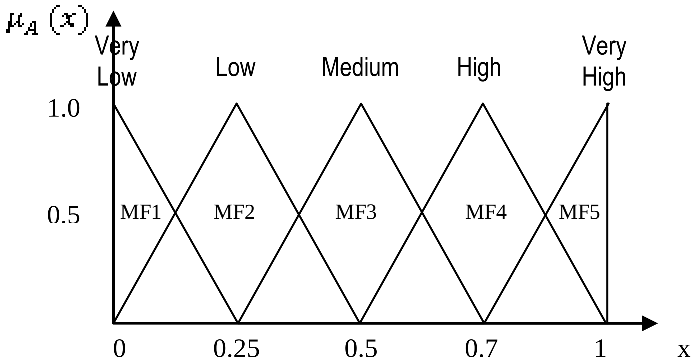

Figure 1 Examples of Membership Function and Definition of Linguistic Terms

A fuzzy inference system consists of three components. Firstly, a rule base contains a selection of fuzzy rules. Rules are some instructions that process input to final output given their membership values and their definitions. Secondly, a database defines the membership functions used in the rules and, finally, a reasoning mechanism to carry out the inference procedure on the rules and given facts.

To this end, the input data will be entered into a neural network (NN)-based process. A neural network is a massively parallel distributed processor made up of simple processing units that have a natural tendency for storing experiential knowledge and making it available for us (Haykin, 1999). It has the ability to model linear and non-linear systems without the need to make assumptions implicitly as in most traditional statistical approaches.

A neural network consists of a large set of interconnected neurons and these neurons are arranged in many layers and interact with each other through weighted connections. The hidden and output layer neurons process their input by multiplying each one of the inputs by the corresponding weights, summing the product, and then processing the sum using a non-linear transfer function to produce a result. Neural networks are regulated by the presentation of a set of examples of associated input and output (target) values. At the end of this training phase, the neural network represents a model, which should be able to predict a target value given the input value.

Jang (1993) develops a combination of a neural network and fuzzy logic called an adaptive neuro fuzzy inference system (ANFIS). It is one of the best tradeoffs between neural and fuzzy systems that provides smoothness due to the fuzzy logic (FL) interpolation and adaptability due to the NN backpropagation. ANFIS model has five levels of layered architecture as shown in Figure 2.

The following represents the detailed calculation procedures. It is assumed that the fuzzy inference system has two inputs (significant parameters) X_1 and X_2 and one output Y and that the rule base contains two fuzzy if-then rules:

Rule 1: If X_1 is A_1 and X_2 is B_1, then $f_1 = p_1x_1 + q_1x_2 + r_1$
Rule 2: If X_1 is A_2 and X_2 is B_2, then $f_2 = p_2x_1 + q_2x_2 + r_2$

The first hidden layer is developed for fuzzification of input variables. If $O_{l,i}$ is the output of the *i*th node of the layer l then every node i in this layer is an adaptive node with a node function:

$$O_{1,i} = \mu_{A_i}(x_1) \qquad for \quad i = 1,2 \quad or \tag{1}$$

$$O_{1,i} = \mu_{B_{i-2}}(x_2) \qquad for \quad i = 3,4 \tag{2}$$

where x is the input node i and $A_i(or\ B_{i-2})$ is a linguistic label associated with this node. Therefore $O_{1,i}$ is the membership grade of a fuzzy set (A_1, A_2, B_1, B_2). A_i or B_i is a lingustic term and typical membership function:

$$\mu_A(x) = \frac{1}{1+\left|\frac{x_i - c_i}{a_i}\right|^{2b_i}} \tag{3}$$

a_i, b_i, c_i is the parameter set and these are referred to as premise parameters.

In the second hidden layer, every node in this layer is a fixed node and represents the fire strength of the rule. The output is the product of all the incoming signals.

$$O_{2,i} = w_i = \mu_{A_i}(x_1).\mu_{B_i}(x_2), \quad i = 1,2 \tag{4}$$

The third hidden layer normalizes the rule's strength. The *i*th node calculates the ratio of the *i*th rule's firing strength to the sum of all the rule's firing strengths:

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2}, \quad i = 1,2 \tag{5}$$

The fourth hidden layer is the layer where the consequent parameters of the rule are determined. Every node *i* in this layer is an adaptive node with a node function:

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i(p_i x_1 + q_i x_2 + r_i) \tag{6}$$

where $\bar{w}_i$ is the output of the layer 3 and (p_i, q_i, r_i) is the parameter set of this node (consequent parameters). The output or the fifth layer computes the overall input as the summation of all incoming signals, and is linear in terms of consequent parameters p, q, and r.

$$O_{5,i} = \sum \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} = \bar{w}_1(p_1 x_1 + q_1 x_2 + r_1) + \bar{w}_2(p_2 x_1 + q_2 x_2 + r_2) \tag{7}$$

The ANFIS is trained by a hybrid learning algorithm, which uses backpropagation learning to determine premise parameters and the least mean square estimation to determine the consequent parameters. The learning procedure consists of two parts. In the first part the input patterns are propagated, and the optimal consequent parameters are estimated by an iterative, least mean square procedure, while the premise parameters are assumed to be fixed for the current cycle through the training set. In the second part the patterns are propagated again, and in this epoch, backpropagation is used to modify the premise parameters, while the consequent parameters remain fixed. This procedure is then iterated to minimize errors.

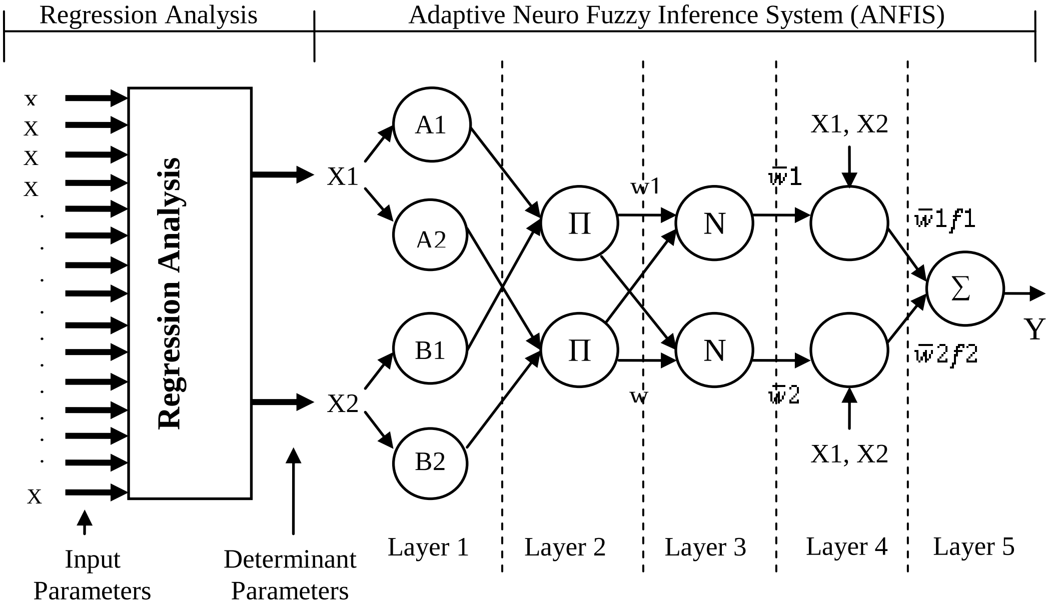

Figure 2 Proposed Model : Regression Analysis and ANFIS Incorporated

3. MODEL APPLICATION

A total of 55 low-cost apartment projects developed by Indonesian Ministry of Public Housing during 2008-2011 in 19 provinces were taken as case studies. So far, the preliminary cost estimates for these projects traditionally used the square meter method with the budget IDR 2.75 million per m^2, and the realization varies from IDR 2.1 – 3.3 million per m^2. This makes the accuracy of the resulting cost estimation highly questionable.

A total of 22 input variables were initially identified as cost driver candidates whereas the output variable is the contractual construction costs (in IDR), (See Table 1). Given the variety of data collected in terms of location and times of construction, data normalization is required to ensure that the cost data are on the same basis. In this study all the construction costs were adjusted to December 2010 and Jakarta as time and location reference, respectively, using the formula:

$$C_{(t)}\ f(n)\ f_{(1)}.C_{(r)}.I_{(t)}/I_{(r1)} + f_{(2)}.C_{(r)}.I_{(t)}/I_{(r2)}\ + ...\ f_{(n)}.C_{(r)}.I_{(t)}/I_{(rn)} \quad (8)$$

where

$f_{(n)}$: fraction of time during which the construction is taking place within year n (if the building construction takes more than 1 year to complete)

$C_{(t)}$: cost at time of interest, $C_{(r)}$ = cost at time of reference

$I_{(t)}$: cost index at time and location of interest,

$I_{(r)}$: cost index at time and location of reference

Since the Construction Cost Index is not yet available in Indonesia, the Consumer Price Index (CPI) was used as a proxy for measures of cost indices.

Table 1 Variables Description

Variables	Description	Range
EZI	Earthquake Zoning Index	0.05 – 1.00
TOF	Type of Foundation	1 = footplate, 2 = shallow bored pile, 3 = driven pile, 4 = bored pile
DOF	Depth of Foundation	2.50 – 30.00 m
NTB	Number of Twin Block	0.50 – 2.00
TOC	Type of Corridor	1 = double loaded, 2 = single loaded
NOU	Number of Units	16 – 196
NOS	Number of Storeys	2 – 5
HOB	Height of Building	9.1 – 15.4 m
BFA	Building Footprint Area	468 – 2,230 m^2
HOS	Height of Storey	2.88 – 3.30 m
LOP	Length of Perimeter	89.4 – 337.8 m
GFA	Gross Floor Area	1,358 – 9,103 m^2
UFA	Usable Floor Area	549 – 4,909 m^2
APU	Area per Unit	18.72 – 39.32 m^2
WFA	Wet Floor Area	94.18 – 663.07 m^2
EWA	Exterior Wall Area	876 – 5,202 m2
UPSR	Number of Units per Number of Storeys Ratio	5.3 – 39.2
UPGR	Usable Floor Area per Gross Floor Area Ratio	0.239 – 0.626
PPGR	Length of Perimeter per Gross Floor Area Ratio	0.099 – 0.033
FPGR	Building Footprint Area per Gross Floor Area Ratio	0.222 – 0.502
TFW	Type of Finishing Wall	1 = brick, 2 = lightweight concrete
DOP	Duration of Project	4.66 – 12.23 months
COST (Output)	Contractual Construction Cost (IDR x 1,000)	4,256,814 – 24,492,799

3.1. Regression Analysis

To develop the regression model, the software SPSS Statistic Release17 was used. The linear and non-linear regressions were applied to determine the model with the results exhibited in Table 2.

3.2. Neural Network

A total of three variants were developed using the NN method based on the variations of input variables and hidden layer neurons, (See Table 3). The NN1 is the model using all identified input variables (22 input variables) while the NN2 used 15 input variables that have correlations with the output variable. The NN3 employed four input variable determinants of non-linear models. The backpropagation learning algorithm and the sigmoid bipolar as function activation records were used for all NN models. Matlab R2009a Software was used to develop the models.

3.3. Regression Analysis – ANFIS (RANFIS) Model

The datasets were divided into two parts by random sampling. The first group of data (50 datasets) were treated as training data were used to develop the model and the second group (5 datasets) were treated as testing data that were used to test the model. Since the values of R^2 and adjusted R^2 of non-linear models are larger than that of the linear model, *GFA*, *APU*, *TOF*, and *UPSR* of the former will be used as input variables to the ANFIS-based method. The software Matlab R2009a was used again to develop the ANFIS model. Table 4 shows the value of each variable.

Table 2 Regression Coefficients

Model	Unstandardized Coefficients		Standardized Coefficient	t	Sig.
	B	Std. Error	Beta		
Linear RA1 Model, with $R^2 = 0.957$ and Adjusted $R^2 = 0.953$					
(Constant)	-1.146E7	3487581.182		-3.286	.002
GFA	2398.948	90.891	.964	26.394	.000
APU	65528.127	17650.164	.118	3.713	.001
HOB	2838828.303	1022280.084	.098	2.777	.008
TOF	608789.626	226849.570	.088	2.684	.010
Non-Linear (Natural Logarithm) RA2 Model, with $R^2 = 0.958$ and Adjusted $R^2 = 0.954$					
(Constant)	8.625	.253		34.150	.000
GFA	.746	.047	.794	15.792	.000
APU	.262	.055	.209	4.811	.000
TOF	.155	.040	.130	3.853	.000
UPSR	.122	.052	.121	2.368	.022

Table 3 Variations in Neural Network Modeling

Specification	Model		
	NN1	NN2	NN3
Input Neuron	22	15	4
Output Neuron	1	1	1
Hidden Layer	1	1	1
Hidden Layer Neuron	11	7	3

Table 4 Description of Selected Cost Drivers

Variables	Range
GFA (m^2)	1,358 – 9,103
APU (m^2)	18.72 – 39.32
TOF	1 = footplate, 2 = shallow bored pile, 3 = driven pile, 4 = bored pile
UPSR	5.3 – 39.2

After data training, the number grid of partitions and the membership functions for each variable were defined. A total of four RANFIS models were developed based on the number of grid partitions, the number of rules, as well as input and output membership function variations as described in Table 5.

Table 5 Variations in ANFIS Modeling

Specification	Model			
	RANFIS1	RANFIS2	RANFIS3	RANFIS4
Input	4	4	4	4
Output Neuron	1	1	1	1
Grid of Partition	3 3 3 3	3 3 3 3	5 5 5 5	5 5 5 5
Rules	81	81	625	625
Input MF	Gaussian	Trapezoidal	Gbell	Trapezoidal
Output MF	Constant	Linear	Constant	Constant

The ANFIS was trained by a hybrid learning algorithm using training data, and the process will be completed after the error tolerance or the maximum number of iterations was achieved, whichever comes first. Once the training process was completed, a fuzzy inference system (FIS) will be subsequently formed and the fuzzy rules will be extracted from training data. The associated fuzzy membership functions of the linguistic terms for input variables of *UPSR* and *GFA* are shown in Figure 3. The knowledge of the model was stored in a fuzzy rule base.

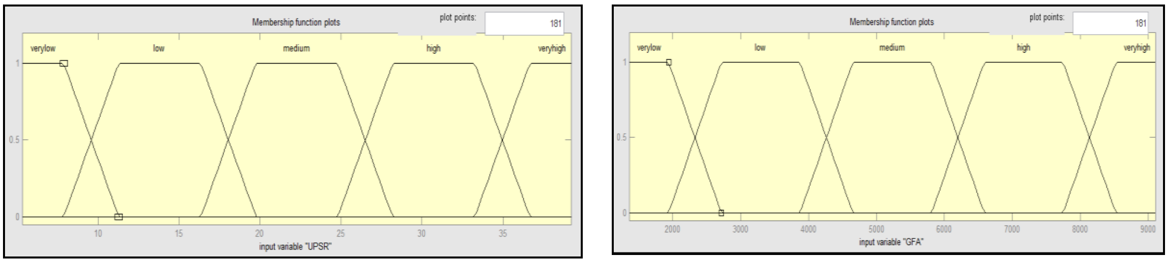

Figure 3 Membership functions for *UPSR* and *GFA*

3.4. Result Comparison

The accuracy performance of all models using 50 data training as well as 5 data testing is based on Mean Absolute Percent Error (MAPE). The results are described in Table 6.

$$MAPE = \frac{1}{n}\sum \frac{|\text{Actual Cost} - \text{Predicted Cost}|}{\text{Actual Cost}} \times 100\% \tag{9}$$

Table 6 Comparisons of Error Rate Result for All Models

Project	Error Rate (%)								
	RA1	RA2	NN1	NN2	NN3	RAN FIS1	RAN FIS2	RAN FIS3	RAN FIS4
MAPE 50 Training Datasets	6.43	5.73	6.35	5.99	10.82	2.46	2.33	2.31	2.54
Project 1	-3.67	-7.32	-45.52	-8.40	4.96	-5.71	-1.31	-1.25	-6.01
Project 2	-8.79	-8.99	-19.32	5.35	4.56	-16.45	-81.31	-12.24	-0.49
Project 3	4.24	13.29	37.53	-5.14	1.97	-0.40	-0.81	-0.40	-0.39
Project 4	7.81	-1.02	-193.19	75.75	-38.62	2.45	2.54	2.45	2.45
Project 5	10.10	8.82	1.43	17.65	-0.50	10.57	10.56	10.57	10.57
MAPE 5 Testing Datasets	6.92	7.89	59.40	22.46	10.12	7.11	19.31	5.38	3.98

4. DISCUSSION

Out of 22, 18 identified variables were removed from RA calculations with the remaining variables able to explain about 96% of cost variations. The very high R^2 and adjusted R^2 coefficients indicate that RA, surprisingly, performs well for both linear and non-linear models to predict cost in the present study. The RA’s MAPEs on 50 training datasets are also deemed favorable if benchmarked to the findings of earlier studies. The cost model developed by Lowe et al. (2006) using gross internal floor area, function, duration, mechanical installations, and piling as key drivers, for instance, has a MAPE of 19.3%.

An almost similar performance is shown by the building construction cost model of Stoy et al. (2008) tested on 70 German residential properties that have a MAPE of 9.6%. Their model was based on compactness of building, number of elevators, project size, expected duration of construction, proportion of openings in external walls, and region as the important cost drivers for residential buildings. Tested on 5 other datasets, the percentage error of RA1 varies between -8.79 and +10.10% and RA2 between -8.99 and 13.29% that are relatively comparable with percentage error of Stoys' that ranges between -12 and 13%.

Neural-network models are underperformance especially for testing datasets. Their MAPEs range between 5.99 and 10.82% for 50 training datasets. Their errors span from as low as -193.19% to as high as 75.75% with MAPEs ranging between 10.12 and 59.40% for 5 testing datasets. Relatively better performance was observed only for the NN3 variant.

The proposed models demonstrate a better improvement in error reductions. On 50 training datasets, the MAPE is only about 2.3-2.5%, which is much lower than other models although one variant, RANFIS2, has a relatively high MAPE when applied on five testing data. This poor performance is attributable to extreme error (-81.31%) in predicting the cost of Project #2. The best prediction was a result of 3.98% as the lowest value of MAPE on five testing data that had resulted from the RANFIS4. This model also has performed very well with a percentage error that ranges between -6.01 and 10.57%. After all, the performance of RANFIS-based model is in general better than the others.

5. CONCLUSION

This paper presents a new approach for preliminary cost estimation models using regression analysis (RA) incorporated with the adaptive neuro fuzzy inference system (ANFIS). This combination allows the merits of each model to be exploited in a positive way. The application of the proposed model has demonstrated that it generally performs better than RA and NN models.

6. ACKNOWLEDGEMENTS

The authors wish to thank the reviewers for their useful comments that improved the quality of this manuscript.

7. REFERENCES

AACE, Association for the Advancement of Cost Engineering, International Recommended Practice No. 18R-97. 2005. Cost Estimate Classification System – As Applied In Engineering, Procurement, And Construction For The Process Industries, TCM Framework: 7.3 – Cost Estimating and Budgeting

Cheng, M.Y., et al. 2010. Conceptual Cost Estimates Using Evolutionary Fuzzy Hybrid Neural Network. *Journal Expert Systems With Applications*, 37, pp. 4224-4231.

Ciraci, M., & Polat, D.A. 2009. Accuracy Levels of Early Cost Estimate, in Light of the Estimate Aims. *Journal of Cost Engineering*, Volume 51, Number 1, pp. 16-24.

Dell'Isola, M.D. 2002. *Architect's Essentials of Cost Management,* AIA. The American Institute of Architects, John Wiley and Sons, New York.

Haykin, S. 1999. Neural Networks - A Comprehensive Foundation, Prentice Hall, Pearson Education.Inc, 2nd Edition.

Holm, L., Schaufelberger J.E., Griffin, D., Cole, T. 2005. Construction Cost Estimating Process and Practices, Pearson Education.Inc, Upper Saddle River, New Jersey, USA.

Jang, J.S.R. 1993. ANFIS : Adaptive-Network-Based Fuzzy Inference System. *IEEE Transaction On System, Man, and Cybernetics*, Volume 23, Number 3, pp. 665-685.

Ji, S.H., et al. 2010. Data Preprocessing-Based Parametric Cost Model for Building Projects: Case Studies of Korean Construction Projects. *Journal of Construction Engineering and Management*, ASCE, pp. 844-853

Kim, K.J., Kim, K., (2010), Preliminary Cost Estimation Model Using Case-based Reasoning and Genetic Algorithms. *Journal of Computing in Civil Engineering*, ASCE, pp. 499-505.

Lowe, D.J., et al. 2006. Predicting Construction Cost Using Multiple Regression Techniques. *Journal of Construction Engineering and Management*, ASCE, pp. 750-758.

Pratt, D. 2011. *Fundamentals of Construction Estimating, Delmar, Cengage Learning, Third Edition*, Clifton Park, New York, USA.

Sonmez, R., Ontepeli, B. 2009. Predesign Cost Estimation of Urban Railway Projects With Parametric Modeling. *Journal of Civil Engineering and Management*, Volume 15, Number 4, pp. 405-409.

Stoy, C., et al. 2008. Drivers for Cost Estimating in Early Design: Case Study of Residential Construction. *Journal of Construction Engineering and Management*, ASCE, pp. 32-39.

Wibowo, A., and Wuryanti, W. 2008. Capacity Factor Based Cost Models for Buildings of Various Functions. *Civil Engineering Dimension*, Volume 9, Number 2, pp. 70-76.