RESEARCH ARTICLE OPEN ACCESS

Comparison of Cost Estimation Methods using Hybrid Artificial Intelligence on Schematic Design Stage: RANFIS and CBR-GA

WisnuIsvara*, Yusuf Latief*, Andreas Wibowo**, MurthadaAskari*
*(Department of Civil Engineering, Faculty of Engineering, University of Indonesia, Indonesia)
**(Agencyfor Research and Development, Ministry of Public Works and Public Housing, Indonesia)

ABSTRACT
Cost estimating at schematic design stage as the basis of project evaluation, engineering design, and cost management, plays an important role in project decision under a limited definition of scope and constraints in available information and time, and the presence of uncertainties. The purpose of this study is to compare the performance of cost estimation models of two different hybrid artificial intelligence approaches: regression analysis-adaptive neuro fuzzy inference system (RANFIS) and case based reasoning-genetic algorithm (CBR-GA) techniques. The models were developed based on the same 50 low-cost apartment project datasets in Indonesia. Tested on another five testing data, the models were proven to perform very well in term of accuracy. A CBR-GA model was found to be the best performer but suffered from disadvantage of needing 15 cost drivers if compared to only 4 cost drivers required by RANFIS for on-par performance.
***Keywords*-**Case basedreasoning, costestimation, geneticalgorithm, neuro fuzzy, regressionanalysis

I. INTRODUCTION

The accuracy of estimation of construction costs from the beginning to the end of projectis a critical factor to the success of the project. At the first phase of design, the schematic design will be prepared and a preliminary estimate can be made when the schematic design develops. The objectives of the preliminary estimate are to design the project within the owner's budget and to evaluate alternative design concepts [1]. However, due to lack of detailed design information during the planning phase, accurate cost estimation is hard to obtain even for professional cost estimators [2].

It has been widely acknowledged that inaccuracies are the main problem in early cost estimates. This is because estimation exercises are often based upon limited data and information available at the time of preparing estimates. According to Holm et al [3], the expected accuracy range of cost estimating at schematic design stage is $\pm$10–20%. Oberlander and Trost [4] explained that what was known about the project (scope) and how was the estimate preparedhave relative influence of 38.6% and 23.5% to the accuracy of early cost estimates, respectively.

A large number of cost estimation models have been developed by previous research. Cost estimation models were traditionally based on statistical methods, including the widely used regression analysis (RA). This analysis method is a very powerful statistical tool that can be used as both an analytical and predictive technique in examining the contribution of potential new items to the overall estimate reliability [5]. A major disadvantage of regression-based techniques is that the mathematical form has to be defined before any analysis can be performed [6]. However, the rapid development of information and computational technologies now allows the use of more sophisticated techniques, such as neural networks (NN), fuzzy logic (FL), case based reasoning (CBR), genetic algorithms (GA)to overcome this constraint andmany research studies have been done in this area. For examples, NN affords a capacity to learn from past data and generalize solutions for future applications; FL allows for tolerance of real world imprecision and uncertainties; CBR solves the new problem by adapting previously determined solutions of the similar previous cases and storing the new successful solution for future use; and GA facilities global optimization of parameters.

At present, more recent research efforts have been directed toward artificial-intelligence (AI) hybrid models that combine AI techniques from one another to obtain better algorithms and, thus, better accuracies. These AI methods include as neurofuzzy system (NFS), fuzzy neural network (FNN), evolutionary fuzzy hybrid neural network (EFHNN), adaptive neuro fuzzy inference systems (ANFIS), case based reasoning-genetic algorithm (CBR-GA), neural network-genetic algorithm (NN-GA) that gain growing popularity for cost estimation models.

Latief et al [7] developed a preliminary cost estimation model incorporating RA and ANFIS, named as RANFIS model. They found that their RANFIS model performed much better than RA and NN models. RANFIS as a hybrid model that integrated three powerful techniques (RA, NN, and

FL) has demonstrated better accuracy performance than other models that only used RA or NN alone. The next question is how about the performance of RANFIS model when compared with another hybrid model that combines several different techniques. This paper further tests the performance of RANFIS against another type of hybrid AI method, a CBR-GA technique.

II. METHODOLOGY

2.1 RANFIS Model [7]

The RANFIS model is an early-costestimation model incorporating RA and ANFIS, as described in Fig. 1. This model employs RA on collected historical data to determine significant building parameters as key cost drivers. The stepwise regression method wasused to address multicollinearity issues among input variables that are common in statistical data analysis [8]. Given that inclusion of insignificant parameters into a model could lead to a poor prediction outcome, elimination of insignificant parameters may improve the prediction performance of the model [9]. The output of this stage will be the input for the ANFIS model.

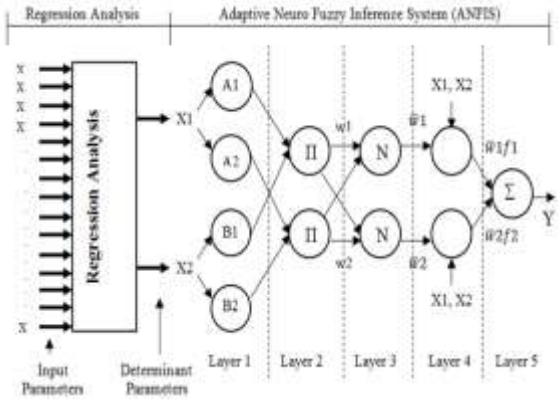

Figure 1. RANFIS Model : Regression Analysis and ANFIS Incorporated

ANFIS was developed by Jang [10] and is one of the best tradeoffs between neural and fuzzy systems. Fuzzy systems are effective in representing explicit, but ambiguous common sense knowledge while NNs provide excellent facilities for approximating data, learning knowledge from data, and parallel processing. ANFIS model has five levels of layered architecture. The nodes in the first and fourth layers are adaptive nodes and the nodes in the second, third and fifth layers are fixed nodes.

It is assumed that the fuzzy inference system has two inputs (significant parameters) X1 and X2 and one output Y and that the rule base contains two fuzzy if-then rules:

Rule 1: If X1 is A1 and X2 is B1, then f1 = p1x1 + q1x2 + r1
Rule 2: If X1 is A2 and X2 is B2, then f2 = p2x1 + q2x2 + r2

In the first layer, input values are converted to their respective membership values by corresponding membership functions as in equations (1) and (2). The membership function can be any appropriate parameterized membership function such as generalized bell function like in equation (3), where a_i, b_i, c_i are the parameter sets and referred to as premise parameters.

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

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

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

In the second layer, every node in this layer is a fixed node and represents the fire strength of the rule. The output is then the product of all 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 layer normalizes the rule strengths. The *i*-th node calculates the ratio of the *i*-th rule's firing strength to the sum of all rule's firing strengths:

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

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

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

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

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

ANFIS is trained by a hybrid learning algorithm since it combines the gradient descent method and the least squares method. It employs gradient descent to fine-tune the premise parameters that define membership functions and uses the least square method to identify consequent parameters

that define the coefficients of each output equations. ANFIS modeling process starts by obtaining a data set (input-output data pairs) and dividing it into training and testing datasets. The training dataset is used to find the initial premise parameters for the membership functions by equally spacing each of the membership functions. A threshold value for errors between the actual and desired output is determined. The consequent parameters areobtained using the least square method, so error for each data pair can befound. If anerror is larger than the threshold, the premise parameters will be updated using the gradient descent method. This process is then iterated to minimize errors and will be terminated when the final error is less than the threshold.

2.2CBR-GA Model

The CBR-GA model is a hybrid approach that combines CBR with GA. The CBR is the process of retrieving previous cases similar to a new problem, solving the new problem by adapting previously determined solutions of the similar previous cases, and storing the new successful solution for future use [11]. A CBR requires four steps: retrieve, reuse, revise and retain [12], as shown in Fig. 2. Cases are represented by attributes describing the circumstance of the problem and its solution. Similar previous cases best matching the new problem are retrieved. The solutions of the retrieved cases are then adapted to fit the new problem. Finally, new solutions are retained for future use once it has been approved.

The GA is a method of intelligently searching for an optional solution based on the genetics and natural selections [13]. GA is an iterative procedure that maintains a population of candidate solutions to optimize the similarity function. The role of GA in the CBR process is used to optimize the outcome in the retrieval process. In this process GA isused to find the value of the variable weight.

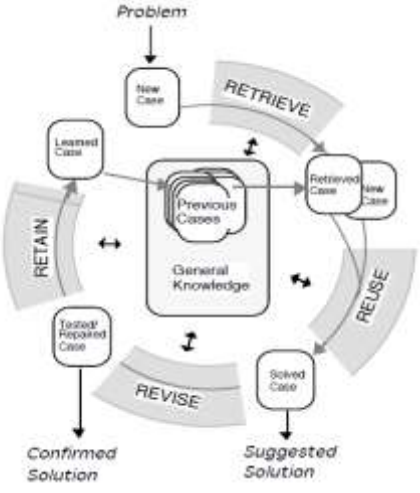

Figure 2. Case Based Reasoning Cycle

To make the similarity function, it is assumed that cost of a project for a particular case can be formulated with appropriate weighting variables:

$$C_k = w_1.X_{1k} + w_2.X_{2k} + w_3.X_{3k} + ...+ w_i.X_{ik} \quad (8)$$

Where$C_k$is cost of project *k*, $w_i$is variable weight *i*and$X_{ik}$is the value of variable *i* in case of project *k*.

However, the concept of equation 8 is technically difficult to apply because the values of w_i should be the inverse of the values of input variables X_{ik}. While the space of input variable is not necessarily an $n \times n$ symmetric matrix, the order is not possible for inversion. To address this problem, it needs optimizations using a GA to obtain the weight of each variable (w_i). The first step is to find the distribution of the distance between case-based variables with a test-case variable using the formula:

$$D_{it} = \sqrt{(x_{i1} - x_{t1})^2 + (x_{i2} - x_{t2})^2 + .. (x_{ik} - x_{tk})^2} \quad (9)$$

Where$D_{it}$is the distribution of the distance of base case to test case variable, x_{ik} is the value of variable *i* in case *k*, and X_{tk} is the value of variable *i* in the test case. The variable weights w_i, can be obtained using the equation:

$$w_i = \frac{1}{J-1}\left[1 - \frac{D_{it}}{D_{ittotal}}\right] \quad (10)$$

Where$w_i$is variable weight *i*, *J* is number of variables and $D_{ittotal}$ is total D_{it}. It is essential to find the value of the similarity between the value of variable *i* in case *k* with the value of variable *i* in test case.

The value of attribute similarities can then be obtained using the following equation:

$$VS_i = \frac{Min\ (Xik,\ Xit)}{Max\ (Xik,\ Xit)} w_i \quad (11)$$

where VS_i is value of similarity for variable *i*, Min (x_{ik},x_{it}) is minimum between variable *i* in case *k* (i.e. base case) and variable *i*in test case, Max (x_{ik},x_{it}) is maximum between variable *i* in case *k* and variable *i*in test case.

Case similarity values can be obtained by summing up *VS* values of each variable and those with the highest total similarity will be the selected case to be reused. The contract value of the reuse process cannot be simply used here. It should be revised based on the difference value of variables that exist in the reuse case and test case, using the equation:

$$RC_i = (X_{it}-x_{iu}) \times UC_i \quad (12)$$

WhereRC_i is Revision Cost for variable *i*, X_{it} is the value of variable *i* in test case, x_{iu} is the value of variable *i* in reuse case and UC_i is unstandardized coefficient of variable *i* from RA.

III. MODEL APPLICATION

3.1 Data and Variables [7]

The historical data of 55 low-cost apartment projects in 19 provinces built by Indonesia's ex-Ministry of Public Housing from 2008 to 2011 were taken as case studies. A total of 22 inputs were initially identified as cost-driver candidates and the output variable was contractual construction costs (see Table 1).

Table 1. Variables Description

Variables Description	**Range**
Earthquake Zoning Index (EZI)	0.05 – 1.00
Type of Foundation (TOF)	1 = footplate, 2 = shallow bored pile, 3 = driven pile, 4 = bored pile
Depth of Foundation (DOF)	2.50 – 30.00 m
Number of Twin Block (NTB)	0.50 – 2.00
Type of Corridor (TOC)	1 = double loaded, 2 = single loaded
Number of Units (NOU)	16 – 196
Number of Storeys (NOS)	2 – 5
Height of Building (HOB)	9.1 – 15.4 m
Building Footprint Area (BFA)	468 – 2230 m^2
Height of Storey (HOS)	2.88 – 3.30 m
Length of Perimeter (LOP)	89.4 – 337.8 m
Gross Floor Area (GFA)	1,358 – 9103 m^2
Usable Floor Area (UFA)	549 – 4909 m^2
Area per Unit (APU)	18.72 – 39.32 m^2
Wet Floor Area (WFA)	94.18 – 663.07 m^2
Exterior Wall Area (EWA)	876 – 5202 m2
Number of Units per Number of Storeys Ratio (UPSR)	5.3 – 39.2
Usable Floor Area per Gross Floor Area Ratio (UPGR)	0.239 – 0.626
Length of Perimeterper Gross Floor Area Ratio (PPGR)	0.099 – 0.033
Building Footprint Areaper Gross Floor Area Ratio (FPGR)	0.222 – 0.502
Type of Finishing Wall (TFW)	1 = brick, 2 = lightweight concrete
Duration of Project (DOP)	4.66 – 12.23 months
Construction COST (IDRx1000)	4256814 – 24492799

Data normalization to adjust costs for location and time is required to ensure that the cost data are on the same basis. In this study, December 2010 and Jakarta were selected as the base year and the base location, respectively. The datasets were divided into two parts by random sampling. The first group containing 50 datasets were used as training data to develop the model and the second group of 5 datasets were used as testing data to test the model.

3.2RANFIS Model

To develop the regression model, software package SPSS Statistic Release17 was used. Running of the same project data, linear and non-linear regressions were applied to determine the best fitmodel. Because the computed coefficient of determination (R^2) as well as the adjusted R^2 of non-linear model were higher than those of the linear model, the former was employed for subsequent analysis. This model suggested that Gross Floor Area (GFA), Area per Unit (APU), Type of Foundation (TOF), and Number of Units per Number of Storey Ratio (UPSR) were found to be statistically significant and therefore used as input variables for ANFIS-based method. The Matlab R2009a was used to develop the ANFIS model. This research developed two RANFIS models based on the variation of the number grid of partitions, rules and the type of membership functions for each variable, as described in Table 2. Considerable time must be spent in determining the grid partition numbers and types of membership functions, which also required a few trial and error processes.

The ANFIS was trained by a hybrid learning algorithm. Once the training process was completed, a fuzzy inference system (FIS) will be subsequently formed. The associated fuzzy membership functions of the linguistic terms for input variables are shown in Fig. 3. The knowledge of the model was stored in a fuzzy rule base, as shown in Fig. 4. All these findings have been reported in Latief et al [7] and interested readers may wish to consult with this reference for more detailed discussions.

Table 2. Variations on RANFIS Model

Parameters	**RANFIS1**	**RANFIS2**
Input/ 4Var	GFA, APU, TOF, UPSR	GFA, APU, TOF, UPSR
Output/ 1 Var	Construction Cost (IDR×1000)	Construction Cost (IDR×1000)
Grid of Partition	5 5 5 4	5 5 5 5
Rules	500	625
MF of Var. 1	Generalized bell	Trapezoidal
MF of Var. 2	Generalized bell	Trapezoidal
MF of Var. 3	Trapezoidal	Trapezoidal
MF of Var. 4	Triangular	Trapezoidal
Output MF	Constant	Constant

The fuzzy decision rules can be visualized and evaluated manually by domain experts. For example, the first fuzzy IF-THEN rule of the fuzzy rule base, shown in Fig. 4, can be interpreted as follows: IF number of units per number of storeys ratio (UPSR) is 17 AND area per unit (APU) is 32.17 m^2 AND total gross floor area (GFA) is 8,043 m^2 AND type of foundation (TOF) is driven pile (3), THEN the total construction cost is around IDR(×1000) 2.2E+007 or IDR. 22 billion.

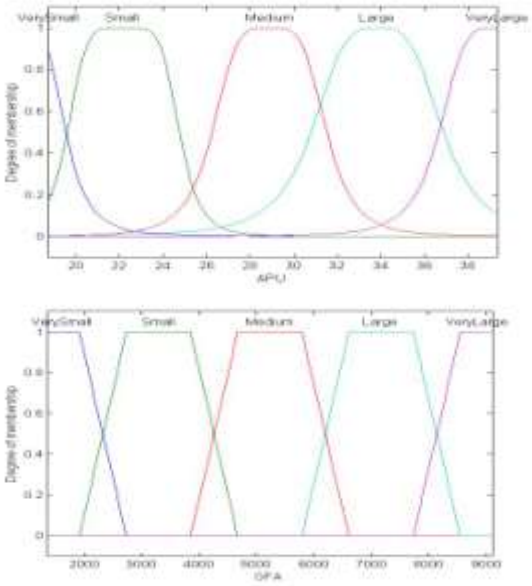

Figure 3. Membership Functions ofAPU and GFA of RANFIS1

Figure 4. Fuzzy Rule Base of RANFIS1

3.3CBR-GA Model

Likewise, two CBR-GA models were developed based on the determinant variables from the correlation analysis and RA, as presented in Table 3. These statistical methods were used to select independent variables that have significant influence to construction cost. As shown, one model using correlation analysis (significant at p =.05) to obtain significant cost drivers (CBR-GA1) employed a much larger number of variables (i.e. 15 variables) for cost estimation. Another model used the same input variables of that of RANFIS based models (CBR-GA2). The case based on each model was then developed using the same 50 datasets.

Table 3. Variables of CBR-GA Modeling

Model	**Variables**
CBR-GA1 (Correlation)	NTB, NOU, UPSR, APU, LOP, GFA, UFA, PPGR, HOB, BFA, NOS, HOS, EWA, TOF, FPGR
CBR-GA2 (Non linier Regression)	GFA, APU, TOF, UPSR

IV. RESULT AND DISCUSSION

The RANFIS and CBR-GA models were tested on the same 5 testing data. The accuracy performance of the four developed models were evaluated based on their Error rate and Mean Absolute Percent Error (MAPE)calculated as follows (see Table 4 for the results):

$$Error\ rate = \frac{Actual\ Cost - Predicted\ Cost}{Actual\ Cost} x\ 100\% \quad (13)$$

$$MAPE = \frac{1}{n}\sum |Error\ rate| \quad (14)$$

Table 4. Performance Comparison between RANFIS and CBR-GA Models

5 Project Testing dataset	Error rate (%)			
	RAN FIS1	RAN FIS2	CBR-GA1	CBR-GA2
Project no 1	-1.22	-6.01	-1.21	-1.22
Project no 2	-2.98	-0.49	-0.20	-5.54
Project no 3	-0.40	-0.39	-0.39	-0.40
Project no 4	2.45	2.45	2.46	2.46
Project no 5	10.57	10.57	10.57	10.67
MAPE (%)	3.52	3.98	2.97	4.04

All the proposed models demonstrate satisfactory accuracies for an early cost estimate. Their errors span from as low as -6.01% to as high as 10.67% with MAPEs ranging between 2.97% and 4.04% for 5 testing datasets. On average, each model yields almost the same error for each data testing. However, RANFIS2 had the worst performance if compared to other models in predicting the cost of Project #1. The largest error (i.e. 10.67%) was resulted from CBR-GA2 model for Project #5.

On another front, RANFIS1 that used4 variables and 3 types of membership functions (generalized bell, trapezoidal and triangular) performed very well with error range of between -2.98% and+10.57% and MAPE of 3.52%. RANFIS2 with only used a single type of membership function (trapezoidal) had a rather poorer performance than RANFIS1, although its number of rules is larger than that of RANFIS1. These findingsevince that the selection of membership function type of each variables becomes important and a large number of grid partition leading to increasing number of rules does not

guarantee that the model performance will be better.

The CBR-GA1 outperformed other models in terms of MAPE and error ranges. Nevertheless, the primary disadvantage of this model is that it required 15 variables to generate a low MAPE.On the other hand, CBR-GA2 that used much smaller number of variables (i.e. 4 variables) performed slightly poorer than RANFIS1 and RANFIS2. The large number of variables will require more resources such as the information, time and cost, whereas these resources are limited at the beginning of the design process.

V. CONCLUSION

In this study, RANFIS and CBR-GA models were developed using datasets of 55 low-cost apartment projects in Indonesia. All the models demonstrate excellent accuracies owing to the fact that they are early-cost estimation models which were based on limited data and information. It has been shown that a CBR-GA model has the best accuracy level but it requires much more input variables. On the other hand, RANFIS models with only four variables were proven to have on-par performance.

RANFIS model can be superior for estimating construction costs at schematic design stage with only very limited information, time and cost while CBR-GA model which requires much more input variables can be more appropriate in value engineering for reviewing the schematic design alternatives by focusing several variables that influence the construction cost to achieve the best value of the project.

Some future directions for additional research can be pursued, such as the application the models to other type of projects to validate the models and generalize the effects of the suggested methods, and compare the models against other hybrid AI methods.

REFERENCES

[1] D. Pratt, *Fundamentals of Construction Estimating* (Delmar, Cengage Learning, 3rd Edition, Clifton Park, New York, 2011).

[2] W. Yu, M.J.Skibnewski, Integrating Neurofuzzy System with Conceptual Cost Estimation to Discover Cost-Related Knowledge from Residential Construction Projects, *Journal of Computing in Civil Engineering, ASCE, January/February,* 2010, 35-44.

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

[4] G.D. Oberlender, S.M. Trost, Predicting Accuracy of Early Cost Estimates Based on Estimate Quality, *Journal of Construction Engineering and Management, ASCE,127(3),* 2001, 173–182.

[5] R.M. Skitmore, B.R.T. Patchell, *Developments in Contract Price Forecasting and Bidding Techniques. Quantity Surveying Techniques: New directions* (P.S. Brandon,ed., BSP Professional Books, Oxford, U.K., 1990).

[6] R.C. Creese, andL. Li, Cost estimation of timber bridges using neural networks, *Cost Engineering, Vol.37, No. 5,* 1995, 17-22.

[7] Y. Latief, A. Wibowo, and W. Isvara, Preliminary Cost Estimation using Regression Analysis Incorporated with Adaptive Neuro Fuzzy Inference System, *International Journal of Technology, Vol 4, Issue 1,* 2013, 63-72.

[8] S.H. Ji, M. Park, and H.S. Lee, Data Preprocessing-Based Parametric Cost Model for Building Projects: Case Studies of Korean Construction Projects, *Journal of Construction Engineering and Management, ASCE, August,* 2010, 844-853.

[9] R. Sonmez, B. Ontepeli, Predesign Cost Estimation of Urban Railway Projects With Parametric Modeling, *Journal of Civil Engineering and Management, 15(4),* 2009, 405-409.

[10] J.S.R. Jang, ANFIS : Adaptive-Network-Based Fuzzy Inference System, *IEEE Transaction On System, Man, and Cybernetics, Vol. 23, No. 3, May/June,* 1993, 665-685.

[11] S.K. Pal, S.C.K. Shiu, *Foundations of Soft Case-Based Reasoning* (Wiley, Hoboken, N.J., 2004).

[12] A. Aamodt, E. Plaza, Case-based Reasoning: Foundational Issues, Methodological Variation and System Approaches, *AI Communication, 7(1),* 1994, 39-59.

[13] K.J. Kim, K. Kim, Preliminary Cost Estimation Model Using Case-based Reasoning and Genetic Algorithms, *Journal of Computing in Civil Engineering, ASCE, November/December,* 2010, 499-505.