The 25th International Symposium
on Automation and Robotics in Construction

Institute of Internet and Intelligent Technologies
Vilnius Gediminas Technical University
Saulėtekio al. 11, 10223 Vilnius, Lithuania
http://www.isarc2008.vgtu.lt/

June 26–29, 2008

ISARC-2008

A STUDY OF PREPROJECT PLANNING AND PROJECT SUCCESS USING ANN AND REGRESSION MODELS

Yu-Ren Wang
Assistant Professor
Dept. of Civil Engineering
National Kaohsiung Univ. of Applied Sciences
415 Chien-Kung Road,
Kaohsiung, 807 TAIWAN
yrwang@cc.kuas.edu.tw

G. Edward Gibson, Jr.
Professor , Garry Neil Drummond Endowed Chair
Dept. of Civil, Construction and Environmental Engineering, University of Alabama
259 HM Corner, Box 870205
Tuscaloosa, AL 35487-0205 U.S.A.
egibson@eng.ua.edu

ABSTRACT

It is long recognized by the industry practitioners that how well preproject planning is conducted has great impact on project outcome. Through industry project data collection and model analysis, this research intends to investigate the relationship between preproject planning and project success. In early stage of the project life cycle, essential project information is collected and crucial decisions are made. It is also at this stage where risks associated with the project are analyzed and the specific project execution approach is defined. To assist with the early planning process, Construction Industry Institute (CII) has developed a scope definition tool, Project Definition Rating Index (PDRI) for industrial and building industry. Since its introduction, PDRI has been widely used by the industry and researchers have been using the PDRI to collect preproject planning information from the industry. Scope definition information as well as project performance are collected and used for this research analysis. This research summarizes preproject planning data collected from 62 industrial projects and 78 building projects, representing approximately $5 billion in total construction cost. Based on the information obtained, preproject planning was identified as having direct impact on the project success (cost and schedule performance). Two techniques were then used to develop models for predicting cost and schedule growth: statistical analysis, and artificial neural networks (ANN). The research results provide a valuable source of information for the industry practitioners that proves better planning in the early stage of the project life cycle have positive impact on the final project outcome.

KEYWORDS

Preproject Planning, Project Success, Regression Model, ANN Model

1. INTRODUCTION

Preproject planning is "….the process of developing sufficient strategic information with which owners can address risk and decide to commit resources to maximize the chance for a successful project" [1]. It is at this early planning stage that significant decisions are made by the project team. Preproject planning process constitutes a comprehensive framework for detailed project planning and includes scope definition. Project scope definition is the process by which projects are selected defined and prepared for definition. It is a key practice necessary for achieving excellent project performance [2]. and is a key element in the preproject planning process. How well pre-project planning is performed will affect cost and schedule performance, operating characteristics of the facility, as well as the overall financial success of the project [3].

Inadequate or poor scope definition, which negatively correlates to the project performance, is among the most problems affecting a construction project [4]. The result of a poor scope definition is that final project costs can be expected to be higher because of the inevitable changes which interrupt project rhythm, cause rework, increase project time, and lower the productivity as well as the morale of the work force [5]. Success during the detailed design, construction, and start-up phases of a project highly depends on the level of effort expended during the scope definition phase as well as the integrity of project definition package [4]. Therefore, it is important to investigate the relationship between preproject planning and project success with real data from the industry. In order to measure the preproject planning efforts for each construction project, a scope definition tool, Project Definition Rating Index (PDRI) is incorporated in this research to evaluate the completeness of project scope definition.

The Project Definition Rating Index, developed by CII, is a comprehensive, weighted checklist of crucial scope definition elements that have to be addressed in pre-project planning process. It provides the project team a simple and easy-to-use tool to objectively evaluate the current status of a project during pre-project planning. Since its development, researchers at the University of Texas at Austin and Construction Industry Institute (CII) have been collecting preproject planning information using the PDRI. For the uniqueness of the different sectors in the construction industry, two versions of the PDRI have been developed specifically for the Industrial and Building sectors.

In addition to preproject planning information collected using the PDRI, project performance (cost and schedule) information was also collected through the data collection process. Traditional statistical analysis method, Simple Linear Regression, and non-traditional statistical analysis method, Artificial Neural Network, were selected in this research to investigate the relationship between preproject planning and project performance using the sample project data.

2. SURVEY INSTRUMENT AND DATA COLLECTION

The data collection was accomplished through a series of retrospective case studies. A scope definition tool, Project Definition Rating Index (PDRI) is used as a survey instrument in these case studies to measure the preproject planning practices in the industry. Data from 62 industrial projects and 78 building projects, representing approximately $5 billion in total construction cost, were collected and used to conduct an investigation of the early planning practices in the industrial and building industry.

2.1. Project Definition Rating Index

CII constituted a research team in 1994 to produce effective and easy-to-use pre-project planning tools so that owner and contractor companies would be able to better achieve business, operational, and project objectives [6]. This research effort led to the development of the Project Definition Rating Index (PDRI). The PDRI for industrial projects is a weighted matrix with 70 scope definition elements (issues that need to be addressed in pre-project planning) grouped into 15 categories and further grouped into three main sections. In responding to the needs of the building industry, CII developed the PDRI for Building Projects in 1999 [7].

The PDRI provides a means for an individual or team to evaluate the status of a construction project during preproject planning with a score corresponding to the project's overall level of definition. The PDRI helps the stakeholders of a project to quickly analyze the scope definition package and to predict factors that may impact project risk specifically with regard to industrial and building projects [8]. For illustration purposes, Section I – Category A of the Building PDRI is shown in Figure 1. This is one category of 11 in the PDRI for buildings and encompasses eight of 64 scope definition elements. [9]

Each element has a corresponding detailed description. Figure 2 gives an example of an element description. Please refer to CII 1996 and 1999 [6]. [7]. for detailed information on development of the tool, all the element descriptions and application of the PDRI.

SECTION I - BASIS OF PROJECT DECISION							
CATEGORY Element	Definition Level						Score
	0	1	2	3	4	5	
A. BUSINESS STRATEGY (Maximum = 214)							
A1. Building Use	0	1	12	23	33	44	
A2. Business Justification	0	1	8	14	21	27	
A3. Business Plan	0	2	8	14	20	26	
A4. Economic Analysis	0	2	6	11	16	21	
A5. Facility Requirements	0	2	9	16	23	31	
A6. Future Expansion/Alteration Considerations	0	1	7	12	17	22	
A7. Site Selection Considerations	0	1	8	15	21	28	
A8. Project Objectives Statement	0	1	4	8	11	15	

0 = Not Applicable *2 = Minor Deficiencies* *4 = Major Deficiencies*
1 = Complete Definition *3 = Some Deficiencies* *5 = Incomplete or Poor Definition*

Figure 1. Category A, Building PDRI

A1. Building Use

Identify and list building uses or functions. These may include uses such as:

- ☐ Retail ☐ Research ☐ Storage
- ☐ Institutional ☐ Multimedia ☐ Food service
- ☐ Instructional ☐ Office ☐ Recreational
- ☐ Medical ☐ Light manufacturing ☐ Other

A description of other options which could also meet the facility need should be defined. (As an example, did we consider renovating existing space rather than building new space?) A listing of current facilities that will be vacated due to the new project should be produced.

Figure 2. Description of Element A1:Building Use

2.2. Data Collection

In the PDRI survey questionnaires, specific questions were intended to obtain historical and "after the fact" project information. The questionnaires included questions regarding project basics (location, type, budget and schedule), operating information, and evaluation using an unweighted PDRI score sheet. Survey participants were asked to think back at a point just prior to construction document (detailed design) development when they filled out the PDRI evaluation score sheet. The total scores were then calculated based on pre-assigned element weights after the questionnaires were returned. Due to the unique nature of these two different sectors, industrial and building projects were examined separately throughout this research investigation.

The sample projects used in this study were obtained from three different sources: previous PDRI research, CII Benchmarking and Metrics research, and institutional organizational (which prefers remaining anonymous) PDRI benchmarking research. Nevertheless, it is important to note that the collected sample from these three sources is based on organization's volunteering projects and not on a random sample of a known population.

During the development of the PDRI, sample projects of actual construction were collected for PDRI validation purpose. Initially, 23 industrial projects were surveyed to gather information regarding PDRI evaluation and project performance during the PDRI validation process [6]. These projects were selected from companies represented by participants in the CII Front End Planning research project. Following PDRI and project team alignment research efforts obtained 18 more industrial projects from CII owner and contractor member companies for a total of 41 industrial projects. Thirty-three sample projects surveyed during the PDRI-Buildings development effort were used for detailed analysis [7]. These projects were selected from a pool of building projects nominated by the PDRI-Buildings research team member companies. The sample projects represented seven different owner organizations and three contractor organizations.

The CII Benchmarking and Metrics Program was established to provide industry performance norms, quantify the use and value of "best practices," and to help focus CII research and implementation efforts. A committee of industry representatives working with the CII staff has defined critical performance and practice use metrics and developed a strategic approach to CII's collection, analysis, and dissemination of industry data. Questionnaires developed by the committee were sent out to owner and contractor member companies to obtain benchmarking and metrics data. The PDRI was included as part of the questionnaire evaluating the preproject planning practice. Twenty-one projects with PDRI information were selected from the benchmarking database for this research. These projects represented nine contractor and 12 owner companies and they were all industrial projects.

The third PDRI data resource is PDRI-Buildings benchmarking research efforts for an institutional organization. This research is conducted by the researchers at the University of Texas at Austin and the objective is to assist the institutional organization benchmarking their preproject planning practice for their capital facility development program [10]. Questionnaire containing the PDRI were sent out to project managers and information from a total of 45 sample building projects was obtained for the same owner organization.

Although the data from these three sources were collected by different researchers, the data collection methods remain consistent throughout the whole data collection period (1994 ~ 2001). That is, a questionnaire survey was used to collect the data and all the questionnaires are filled out by survey participants and followed up by researchers if necessary. In summary, information from a total of 62 industrial projects representing a total cost of approximately $3.9 billion dollars was obtained for the research. In the meantime, 78 building projects representing approximately $1.1 billion dollars in total budget cost were also collected. A detailed breakdown of the projects is presented in Table 1.

3. DATA ANALYSIS AND MODELING

Specific questions in the PDRI survey questionnaires were intended to obtain historical and "after the fact" project information. The questionnaires included questions regarding project basics (location, type, budget and schedule), operating information, and preproject planning evaluation using a PDRI score sheet. Survey participants were asked to think back at a point of time just prior to construction document (detailed design) development when they filled out the PDRI evaluation score sheet. The total scores were then calculated based on pre-assigned element weights after the questionnaires were returned. Please refer to [6]. and [7]. for detailed development of PDRI element weights.

The PDRI score obtained from the survey is a good indicator of the level of preproject planning for each project. Two project performance aspects are of particular concern for this research: cost and schedule performance. Cost performance and schedule performance are measured by cost and schedule growth. In the survey, respondents were asked to provide estimated costs at the start of construction document development as well as the

Table 1. Industrial and Building PDRI Data Collection

Sector (1)	Resource (2)	No. of Projects (3)	Represented Cost (Billion) (4)
Industrial	PDRI-Industrial Research (1996)	23	$1.6
	Alignment Research (1998)	18	$1.9
	CII BM&M Database (2001)	21	$0.4
	Industrial Projects Total	**62**	**$3.9**
Building	PDRI-Buildings Research (1999)	33	$0.8
	Institutional Organization Benchmarking (2001)	45	$0.3
	Building Projects Total	**78**	**$1.1**

actual costs after construction completion. Total cost growth measures total project cost growth as a percentage of the initial predicted project cost. Cost performance was measured by project *Cost Growth* metric obtained as follow:

$$\frac{Actual\ \ Total\ \ Cost - Initial\ \ \Pr oject\ \ Cost}{Initial\ \ \Pr edicted\ \ \Pr oject\ \ Cost} \quad (1)$$

The total project duration used to calculate project schedule growth was measured from the start date of construction documents development to the date of substantial completion in months. The following equation was used for computing project schedule performance, *Schedule Growth*:

$$\frac{Actual\ \ Total\ \ Schedule - Initial\ \ Schedule}{Initial\ \ \Pr edicted\ \ \Pr oject\ \ Schedule} \quad (2)$$

Two different predictive models, regression model and ANN model, are selected for this research to investigate the relationship between the preproject planning and project performance using PDRI scores and cost/schedule growth.

3.1. Regression Model

The regression methodology models the distribution of a variable (dependent variable) with the help of one or more predictor variables (independent variable). Simple regression analysis involves only one predictor when investigating its relationship with the dependent variable. While in multiple regression, more than one predictor is studied for their relationship with the dependent variable. Though the simple linear regression accounts for only one predictor in modeling a dependent variable, in many situations, a linear function of X, or a suitably transformed X, will be a good first approximation of the true relationship [11].

For simple linear regression, only two variables, independent and dependent variables, exist in the data, $\{ (x_i, y_i): i = 1, \ldots, n \}$. The fitted linear equation is written as $\hat{y} = b_0 + b_1x$, where ŷ is the predicted value obtained by using the equation. The differences between the observed values, y_i, and the predicted values, $\hat{y}_i$, are defined as the residuals, $\{ (y_i - \hat{y}_i) : i = 1, \ldots, n \}$.

The simple linear regression equation is also known as the *least squares* regression equation. The best fitting line is chosen under the criteria that the sum of the *squares* of the residuals should be *least*. That is, the least squares regression equation is the line for which the sum of squared residuals, $\Sigma (y_i - \hat{y}_i)^2$, is a minimum. Under the situation that the sum of squared residuals is minimized, the corresponding b_0 and b_1 in the least squares equation, $\hat{y} = b_0 + b_1x$, are calculated using the following equation [12]:

$$b_1 = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sum (xi - \bar{x})^2}, b_0 = \bar{y} - b_1 \bar{x} \quad (3)$$

The coefficient of determination, R^2, is a statistic that is widely used to determine how well a regression fits the data. R^2 represents the fraction of variability in the dependent variable, y, that can be explained by the variability in the independent variable, x. In other words, R^2 explains how much of the variability in the dependent variable can be explained by the fact that they are related to the independent variable. Eq. (4) shows how the coefficient of determination, R^2, is calculated.

$$R^2 = \frac{S_{yy} - SSE}{S_{yy}} \quad (4)$$

where $S_{yy} = \sum (y_i - \bar{y})^2, SSE = \sum (y_i - b_0 - b_1 x_i)^2$

Using the PDRI score as the independent variable and cost/schedule growth for each project as the dependent variable, a simple linear regression analysis can be performed to examine the relationship between the independent variable (PDRI score) and dependent variable (cost/schedule growth). First, a scatter plot of the PDRI score and Cost Growth was constructed and a best fit line was then calculated and plotted on the scatter plot. The linear regression analysis is performed using SPSS 12.0. For industrial projects, the ANOVA results showed that Significant F equals to 9.5E-05, which is less than 0.05 and indicates the linear relationship is statistically significant. That is, there is linear relationship between the PDRI score and Cost Growth for the surveyed industrial projects. In order to examine how the obtained linear regression equation represents the data, two sets of parameters, R and R^2 are calculated. Firstly, the linear correlation coefficient, R, measures the strength and the direction of a linear relationship between two variables. The value of R ranges between -1 and +1. An R value close to +1 indicates that two variables have strong positive linear correlation. The obtained R value for this particular model is 0.475, which indicate a relationship that as values for PDRI scores increase, the values for cost growth increase as well. Secondly, the coefficient of determination, R^2, records the proportion of variation in the dependent variable (cost growth) explained or accounted for by variation in the independent variable (PDRI score). For this particular model, the obtained R^2 equals 0.23 which indicates that 23% of the variation in project cost growth can be explained by the variation in PDRI scores. Figure 3 shows the scatter plot and regression results for the industrial projects.

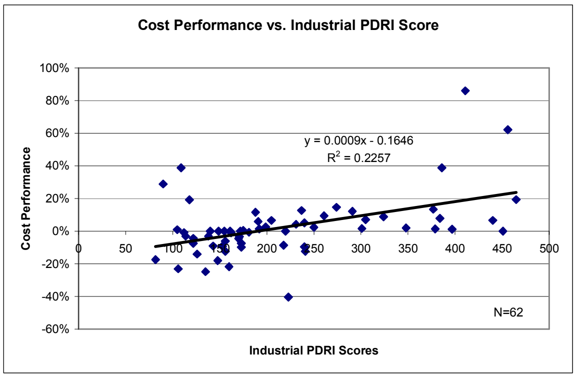

Figure 3. Simple Linear Regression: Cost vs. PDRI

The analysis results show that the linear relation between PDRI score and project cost growth is statistically significant and PDRI score (completeness level of scope definition) can be used to explain a certain proportion of project cost growth. Similar results are observed for project schedule growth vs. PDRI score. However, it should be noted that many factors may influence the project after preproject planning and therefore, can contribute to cost overruns and schedule slippage such as poor contract documents, unforeseen conditions, market conditions, strikes, and Acts of God, and so on. Nevertheless, the PDRI score still serves as a good indicator as to if the project is heading to the right direction in the beginning stage of the project life cycle.

3.2. Artificial Neural Network Model

Studies have shown that ANNs have sevreal advantages over the traditional statistical methods such as multiple regression analysis and multivariate analysis [13]. ANNs does not require that the data must follow a specific statistical distribution and does not require predetermination of the relationships between inputs and outputs. In addition, ANNs have very strong capability of self-learning and self-updating. Considering these advantages, this study has chosen neural networks for modeling project success.

The principle of Neural Networks is based on the assumption that a highly interconnected system of simple processing elements can learn complex interrelationships between independent and dependent variables [14]. A typical neural network consists of an input layer, an output layer, and one or more hidden layers. These layers are connected by neurons to form a parallel distributed processing system. Each neuron is viewed as a processing element (PE) that receives inputs and generates outputs through an activation function. Each of the connections between the process elements has an associated weight. Fig. 4 shows a typical three-layered neural network with an input layer (I), a hidden layer (H), and an output layer (O).

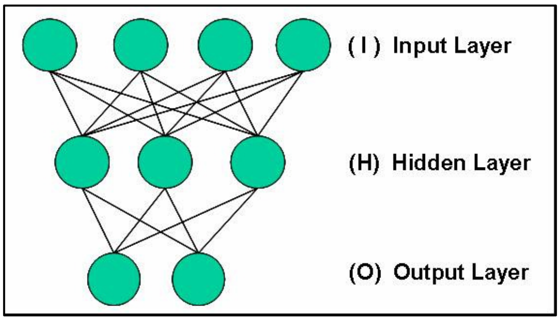

Figure 4. Example of three-layered neural network

In the hidden layer, each neuron receives an activation signal (input), and generates a signal (output) through an activation function. The activation signal is the weighted sum of all the signals entering the neuron, as shown in Eq. (5). In Eq. (5), x_j is the activation signal that the neuron j in the hidden layer receives; I_i is ith input in the input layer; and W_{ij} is the weight of the connection between the neuron j in the hidden layer and the input I_i. The neuron (Process Element) produces an output through an activation function that can be any form. The activation function can be either linear or non-linear, and one of the most commonly used activation function is the *sigmoid function*. The general form of *sigmoid function* is shown in Eq. (6), where h_j equals output of the neuron j in the hidden layer and x_j equals input for the neuron j.

$$x_j = \sum_i I_i W_{ij} \tag{5}$$

$$h_j = f(x_j) = \frac{1}{1+e^{-x_j}} \tag{6}$$

$$y_k = \sum_j h_j W_{jk} \tag{7}$$

As presented by Eq. (7), the neurons in the output layer receive activation signals (weighted sum of inputs to neuron k) from the neurons in the hidden layer. In Eq (7), y_k is the input of the neuron k in the output layer and W_{jk} is the weight of the connection between the neurons j and k in the hidden and output layers, respectively. In the output layer, these activation signals are transformed (through activation function) again to generate the outputs of the neural network. This process is shown in Eq. (8), where o_k is the predicted value of the outputs. Then the outputs are compared with desired or actual values, d_k.. The error (difference between predicted value and desired/actual value) at the output neurons is defined by Eq. (9). The best performance of the neural network is achieved when the error is minimized.

$$o_k = f(y_k) = \frac{1}{1+e^{-y_k}} \tag{8}$$

$$E(W) = \frac{1}{2}\sum_k (d_k - o_k)^2 \tag{9}$$

For supervised neural networks (models with specific actual/desired outputs), one of the most effective and popular technique to minimize the error function $E(W)$ is the back-propagation (BP) algorithm. For back-propagation neural networks, the error at the output layer *propagates backward* to the hidden layer and then to the input layer to update the weights for each of the connections in the neural networks. These forward process (input layer to hidden layer to output layer) and backward process (output layer to hidden layer to input layer) are repeated to minimize the error.

These repeated processes are viewed as learning (training) process. The relationships between inputs and outputs of the system are memorized through the connection weights. It should be noted that before the learning process starts, small random numbers (e.g., between −0.1 and 0.1) are assigned as the initial weights to the connections between the neurons. This ensures that the network is not saturated by large values of the weight, and prevents some training pathologies. Sometimes, the data will be normalized before to obtain convergence within a reasonable number of cycles.

To develop the ANN model, a commercial software package *Neurosolution* was chosen for its ease of use (built-in with EXCEL), speed of training, and host of neural network architectures, including back-propagation with flexible user selection of training parameters. Data from 32 industrial projects are randomly chosen as the training set for the model. One project is used as cross validation to see if over-training has occurred. The model training process includes a total of three runs and each run with 1000 epochs. The mean square error for training of this model is shown in Figure 5.

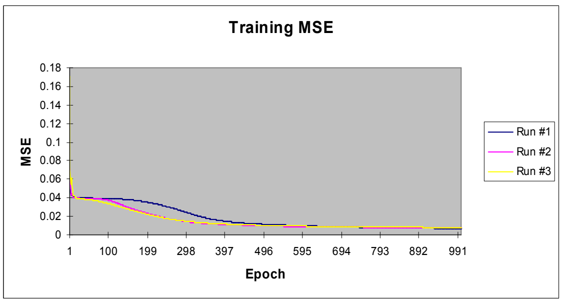

Figure 5. Mean Square Error for Training

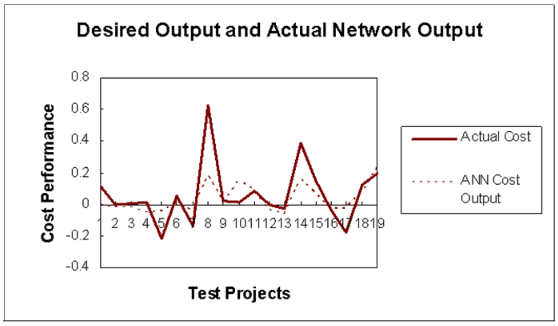

Figure 6. ANN Model Test Output

The best results (model with least mean square error) and the corresponding weights are kept for model testing. Nineteen projects are used to test the model with and Figure 6 shows the test results. The results show that the trend of the ANN outputs matches the trend of the actual cost. In addition, ANN predictions are close to the actual cost data except for some cases where actual cost growths are very high. The calculated coefficient of correlation is about 0.75, which indicates that the ANN model is better than the linear regression model.

4. CONCLUSION

This paper studied the preproject planning of industrial and building construction projects and investigates its relationship with project success (cost and schedule performance). Questionnaire surveys were used to obtain information related to the status of preproject planning and project performance. The score obtained from the Project Definition Rating Index in the questionnaire was used to measure the completeness level of preproject planning for each surveyed project. This PDRI score was considered the independent variable in the model development. Also from the surveyed questionnaires, project Cost Growth and Schedule Growth were calculated as the dependent variable in the model. A scatter plot was first produced to examine the relationship between the independent variable (PDRI Score) and dependent variable (cost and schedule growth respectively). Based on the collected data, this research developed two models to predict the project performance (Cost and Schedule Growth individually) using the PDRI score. The fist model was simple linear regression model and the second was artificial neural network (ANN) model. Data collected from a total of 62 industrial and 78 building projects were used for the model development. Both models show positive relationship between PDRI score and cost/schedule growth for this particular sample of projects. The results indicate that projects with better preproject planning are more likely to have a better project performance at completion.

REFERENCES

[1] The Construction Industry Institute (1995). *Pre-Peoject Planning Handbook*. Special Publication 39-2, Austin, TX

[2] Merrow, E. W., and Yarossi, M. E. (1994). “Managing Capital Projects: Where Have We been – Where are We Going?” *Chemical Engineering*, October: 108-111

[3] Gibson, G. E. and Hamilton, M. R. (1994). *Analysis of Pre-Project Planning Effort and Success Variables for Capital Facility Projects*, A report to the Construction Industry Institute, Source Document 105, Austin, TX

[4] Gibson, G. E. and Dumont, P.R. (1996). *Project Definition Rating Index (PDRI)*, A report to the Construction Industry Institute, Research Report 113-11, Austin, TX

[5] O’Connor, J. T., and Vickroy, C. G. (1986). *Control of Construction Project Scope – Source Document 6*, Austin, Texas: Construction Industry Institute, The University of Texas at Austin

[6] The Construction Industry Institute (1996). *Project Definition Rating Index (PDRI) – Industrial Projects*. Implementation Resource 113-2, Austin, TX

[7] The Construction Industry Institute (1999). *Project Definition Rating Index (PDRI) – Building Projects*. Implementation Resource 155-2, Austin, TX

[8] Cho, C. S. (2000). Development of the Project Definition Rating Index (PDRI) for Building Projects. Ph.D. Thesis, University of Texas at Austin, Austin, TX

[9] Cho, C. S. and Gibson, G. E. (2001). Building Project Scope Definition Using Project Definition Rating Index. *Journal of Architectural Engineering*, Vol. 7, No. 4, pp. 115-125

[10] Wang, Y.R. (2002). *Applying the PDRI in Project Risk Management*. Ph.D. Thesis, University of Texas at Austin, Austin, TX

[11] Montgomery, D.C., and Peck, E.A. (1982). *Introduction to Linear Regression Analysis*. John Wiley & Sons Inc., NY

[12] Milton, J.S., and Arnold J.C. (2003). Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences, 4e. McGraw-Hill, NY

[13] Elhag, T. M. S., and Boussabaine, A. H. (2002). Tender price estimation using artificial neural networks. II: Modelling. *J. Finan. Mange. Property Constr*. 7(1), 49–64.

[14] Elhag, T. M. S. (2004). Cost modeling: Neural networks vs. regression techniques. *Int. Conf. on Construction Information Technology (INCITE)*, Langkawi, Malaysia, Construction Industry Development Board Malaysia (CIDB).