Review On Prediction of Cost Overrun in Construction Projects Using Non-Traditional Methods

Karan Gotlur[1], Gurunath Pujar[2], Shreenidhi Nayak[3] and Jhaswanth Raj BH[4]

[1]*Department of Civil Engineering, R.V College of Engineering, Bengaluru, Karnataka, India*
[2]*Department of Civil Engineering, R.V College of Engineering, Bengaluru, Karnataka, India*
[3]*Department of Civil Engineering, R.V College of Engineering, Bengaluru, Karnataka, India*
[4]*Department of Civil Engineering, R.V College of Engineering, Bengaluru, Karnataka, India*

---***---

Abstract – *The paper talks about different types of non-traditional methods are used to make a prediction model for determining percentage cost overrun in construction projects. Identification of factors that contribute to cost overrun are important for the collection of data. The datasets obtained are focused on certain factors that escalate the cost overrun. These datasets are quantified and calculated with the help of non-traditional methods like regression analysis, fuzzy logic and artificial neural network. An in depth analysis focus on each of their methodology and this can be used in the functioning of a model. The data sets considered in the study are in the form of actual costs and budgeted costs. The analysis of the following methodologies are compared and the nearest best result is considered. The paper has given more stress on the non-traditional methods as they have proven to be equally efficient compared to the traditional methods.*

***Key Words*: Cost Overrun, Non-Traditional Method, Linear regression, Fuzzy Logic, Artificial Intelligence, Artificial Neural Network, Model Efficiency.**

1. INTRODUCTION

The construction sector plays an important role in the economic growth of the country. The government of India aims for the development of construction and services through large budget allocations in this sector. The Indian construction industry is expected to grow at an annual average of 6.6% between 2019 and 2028. The government of India targets to build 5 crore homes over the next five years. Around 8% of Gross Domestic Product (GDP) is covered by the construction industry.

In the current scenario, as per the *Ministry* of *Statistics* and Program Implementation (MSPI) report, the infrastructure in India shows a cost overrun of 3.89 lakh crores from 373 projects. Because of such an issue, construction companies need to adopt better methods to predict the cost overrun. Since we are predicting the extra burden cost, it will not surprise an investor or contractor. It will boost the confidence to invest in this industry.

Prediction models help in analyzing factors. Considering the necessary steps, it reduces cost overrun. Major factors that lead to cost overrun are Design Changes, Escalation of Materials Price, Scheduling, Over Budgeting and Government Policies. An overview of factors that affect the cost overrun was given by Larsen et al. [1]. He ranked major factors based on their effect on the cost overrun [1]. The statement made by Senouci et al.[2] was noticed that the overrun between 2007 and 2013 was less, compared to the projects between 2000 and 2007 [2]. Bhargava et al. [3] suggested the model that helped to enhance the estimation of overrun in final cost and time delay of projects. A light on a few more factors was thrown upon which affect the project and the process of bidding [3]. Hinze et al. [4] linked cost overrun of Washington State Highway Projects and noticed that the cost overrun is expressed as a percentage of the contract amount, which seemed to increase with the increasing size of the project [4]. A reference by Rowland et al. [5] to the escalation of cost and time overrun was due to contract size, complexity, and miscommunication of information in large projects [5]. Another instance of cost and time overrun Chang et al. [6] states when the design changes go beyond the scope of the owner or consultant which is an increase in the scope of work, legislation changes or changes in standards or any archaeological findings [6]. Authors like Akpan and Igwe [7] refer to cost overruns are due to the cost of material and labour charges, insufficient examination, bad costing methods, poor scheduling and shortage of information [7]. [8] Few authors mentioned that risk factors were correlated to the project design, construction and project environment (Akinci and Fischer) [8]. The prediction model was proposed by Karla et al. [9] using fuzzy logic and took design changes as one of the key factors affecting cost overrun. The model used a rating system is given by the user on project characteristics and occurrence of a risk event and a relationship function between these two was used to determine the overall cost overrun percentage [9]. Li et al. [10] proposed a fuzzy logic model to track the project performance by using project duration and also taking into account qualitative and quantitative factors that affect cost overrun in construction projects. Discussion by Mohammed et al [11]. gave an application of a multilayer feed-forward network in construction projects. Required data for the input layer was mentioned. The study revealed that ANN is more accurate than the other methodologies [11]. Regression

Analysis and Neural Networks are used in the prediction model of the cost overrun, which is given by Kang et al. [12]. The prediction model which uses ANN has considered reconstruction projects to determine the cost overrun. Factors affecting the reconstruction projects are considered [12]. A fuzzy logic model that integrates daily site reporting of activity progress and delays, with a schedule updating and forecasting system for construction project monitoring and control by Adriana V. Ordóñez Oliveros and Aminah Robinson Fayek et al. [13]. A fuzzy logic model by Lorterapong and Moselhi et al. [14] is used for estimating the duration of project activities, by accounting for factors such as site conditions, weather, and labour performance, all of which are best expressed linguistically.

2. PREDICTION MODELS

A prediction model can be used to estimate and analyse the cost overrun in a construction project. These models help the contractors to quote the best price for the activity and instil confidence in the clients to allocate the required amount. Models can be classified into Statistical and Non-Statistical. The model under the Statistical method is a Multiple Linear Regression model. Models under the Non-Statistical methods are Fuzzy Logic and Artificial Neural Network.

The following methods are considered in the Prediction Models:

2.1 Regression Analysis
2.2 Fuzzy Logic
2.3 Artificial Neural Network

2.1 Regression Analysis

The Statistical approach has many sub-categories namely Linear Regression, Multiple Linear Regression, Three-Stage Regression Analysis, Multivariate Regression. Multiple Linear Regression (MLR) involves two or more explanatory (independent) variables with a single response (dependent) variable. Three-Stage Least Square (3SLS) combines a system equation (unrelated regression) with a two-stage least square estimation. Assuming that each equation of the system is at least identified.

To attain the objective, proof about the theoretical and empirical, cost overrun and time overrun have to be taken into account simultaneously. For the development of such models, the three-stage least square approach is used.

2.1.1 Investigating the Necessity for Simultaneous Equations

The objective is to use a model and estimate the construction cost overrun and time overrun on the data that is available during the initial phase. The equations mentioned below are used:

$$c = \alpha_c + \beta_c X_c + \lambda_c t_c + \varepsilon_c \qquad (1)$$

$$t = \alpha_t + \beta_t X_t + \lambda_t c + \varepsilon_t \qquad (2)$$

Here, c and t are cost overrun and time overrun respectively, X_c and X_t are the vectors of factors affecting cost overrun and time overrun respectively; β's are vectors of estimate parameters; λ's estimable scalars; and ε's are disturbance terms capturing unseen effects.

The following equations are made such that the time overrun impacts the cost overrun [including the term 't' as an explanatory variable, RHS variable in Eq(1)] and the cost overrun impact the time overruns [including the term 'c' as an explanatory variable, RHS variable in Eq(2)] vice versa.

If equations 1 and 2 are individually considered by Ordinary Least Squares (OLS) an important assumption is breached, the reason being the RHS variables (t and c) are endogenous which implies that any alterations done on the LHS will in-turn affect the variables of the RHS variables. Thus, informing OLS regression is best suited for, when the RHS variables are exogenous (LHS variable do not have an impact on the RHS variables).

2.1.2 Method to Solve Simultaneity Problem

From the literature study several correction techniques have been found to solve the problem of endogenous variable some of these techniques are indirect least squares, two-stage least squares and limited information maximum likelihood also some are system equation techniques like 3 stage least square and full information maximum likelihood. The single equation technique succeeds in dealing with the endogenous problem but is not responsible for the probable relationship between the disturbance terminologies. Not considering such a relationship shall make the model inefficient. Amongst the 2 commonly used system equation methods both the three-stage least square and the full information maximum likelihood possess almost equal variance-covariance matrix, therefore either of them can be used for analysis. Hence the usage of three-stage least square is being made.

The first step of the three-stage least square is to revert (regress) endogenous variables in opposition to all the exogenous variables (all the X apart from the c and t) and use the outcome equation-estimated values of c and t RHS variables are the estimation of equations 1 and 2. In the seconds' step equation estimates are used to calculate disturbances to find the correlation between disturbance terms (ε_c and ε_t). In the third and the final step, we use generalized least squares (GLS) to calculate the final limits of the estimates. In GLS an alternative to estimating limits using standard matrix algebra form of OLS $\hat{\beta} = [X^T X]^{-1} X^T Y$ estimation is done by $\hat{\beta} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1}$ where Ω matrix represents the past correlation among disturbances.

We found from the literature study that project type had a strong influence on cost surge. A test was conducted to find out if separate models were accountable for dissimilar project types. Outcomes suggested that different models need to make to examine different types of projects to know the impact of the independent variables on cost overrun and also time overrun. Adding to the three-stage least square model-estimation results were compared to models by the OLS and single-equation technique and we observed that the three-stage least square model provides a statistically higher fit. The t-statistic is the ratio of the regression coefficient (of a given independent variable) to that of its standard error. It tells us the influence of the independent variable on that of the response variable with a certain level of confidence (a large value of t-statistics will have a high level of confidence which tells that the coefficient capable of estimating to a fair degree of accuracy).

The three-stage least square model outcomes portraited that for all project types, cost overrun and time overrun showed a simultaneous relationship in the cost overrun model, the variable which represented the time overrun was statistically significant (at 99% level of confidence) independent variable regardless of the type of project. Similarly was the case of the time overrun model. Adding on, for all the estimated models so far the three-stage model's result portraited a high correlation between error terms of cost overrun and time overrun model equations, this also shows that simultaneous relationship is present between cost and time overrun.

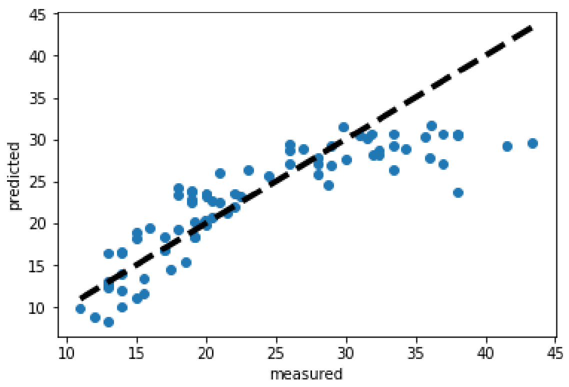


Fig -1: Expected Outcome of Regression Analysis showing the best fit curve

The best fit line for the calculation of regression analysis can be done using the software, python providing previous data information (factors) the model can be tested for and later act as a validator for the other cases.

2.2 Fuzzy Logic

Fuzzy logic is a flexible machine learning technique which mimics the logic of human thought. A logic sometimes has two values representing two possible solutions. Fuzzy logic is a multi-valued logic that allows intermediate values to be defined and provides an interference mechanism that can interpret and execute commands.

Fuzzy systems are suitable for uncertain or approximate reasoning as the construction cost depends on various uncertain factors that need to be quantified. These factors have to be given a rating system to arrive at a result. Due to the imprecise nature of many factors that affect construction projects and a general lack of data for proper quantification of factors, fuzzy logic lends itself well to construction applications.
Example: How can u define if the cost is too much or too less? If the design changes done by a client are many then what does too many indicators in terms of value?

That is where fuzzy logic comes into practice.

Fuzzy logic has 3 components in it as shown in Fig.2

1. Input variables using membership function: predefined membership functions such as decreased, actual and increased.
2. Fuzzification
3. Knowledge-based rules in the developed fuzzy inference
4. Defuzzification of the output variables

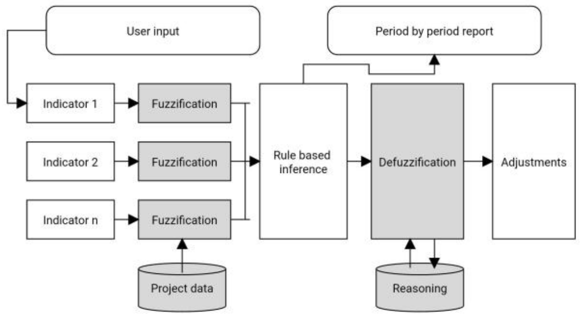


Fig -2: Developed Forecasting Method

The input variables are performance indicators of labour, material prices, and equipment, etc. These indicators provide ratios of actual against budgeted values and in the membership function fuzzified.

The fuzzy rules which are developed, assist in forecasting the cost and duration of the construction project. The rule-based inference has been set to generate the output. The rules set are usually in terms of IF...THEN condition. In defuzzification, the rules and the user inputs are combined to arrive at a result. In the period of executing a construction process, the factors have to be quantified and should be used to calculate the cost overrun percentage.

The factors to be valued need to be given certain parameters such that the inference rule-based system can be applied to it.

For a better understanding, an example is given where the input and output membership functions have been set which is then fuzzified and, the fuzzy values are converted into crisp values by defuzzification.

Input Member Functions

Decreased = {-5,-4,-3,-2,-1}

Actual = {0}

Increased = {1, 2, 3, 4, 5}

Output Member Functions

Cost = {Underrun, Actual, Overrun}

After the parameters are set, as shown above, the rules are set in inference.

RULE 1: (IF the cost of material is decreased THEN the cost is underrun).

RULE 2: (IF the cost of material is actual THEN the cost is actual).

RULE 3: (IF the cost of material is increased THEN the cost is overrun).

The fuzzification is done when the rules check the condition with the input given by the user.

User input: 5

Category: Increased

So the implementation of RULE 3 has taken place.

Defuzzification: Conversion from fuzzy value to crisp value is called defuzzification. Here the output generated is 'overrun'. Finally, in the adjustments, the parameters have to be set to show the percentage of cost overrun that has occurred. Adjustment: For user input '5' assume that overrun is '50%'.The adjustment also has limits. Shows the cost overrun or the percentage increase cost overrun.

2.3 Artificial Neural Network

Artificial Neural Network (ANN) is an information processing model that is inspired by the biological nervous system. ANN is composed of a large number of interconnected processing elements called neurons. Each neuron is connected with other neurons by connection link. Connection links are associated with weights which contain information about input data. Input data is used by the neural network to solve a problem. The processing elements of ANNs are learn, recall and generalize from the input data. ANN can be trained in previous situations. Training of the network is required to change weights that allow the neural network to predict outputs. Predicted outputs are expected to be nearer to actual outputs. Through this, network will generalize the new data.

The following section presents the steps performed to design the ANN model, Shown through a flow chart.

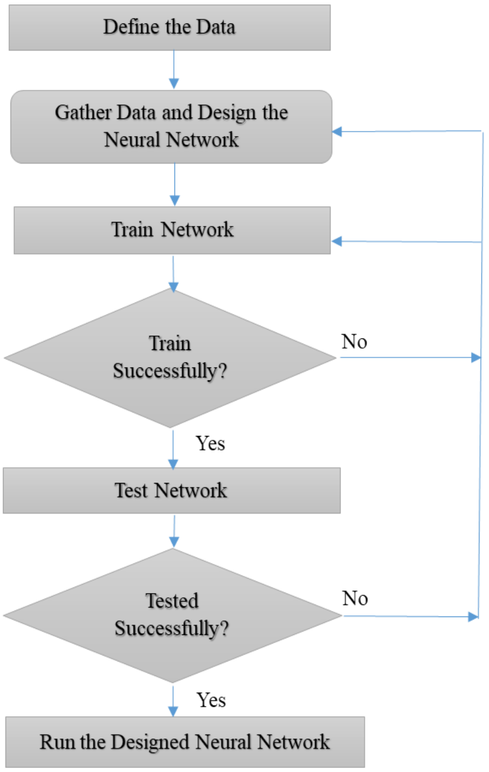
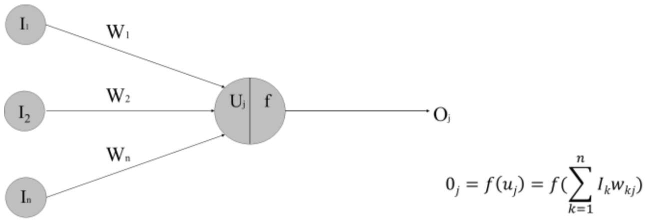


Chart-1: Neural Network Design

Multilayer feedforward network is the most commonly used ANN for developing the prediction models. Multilayer feed-forward network consists of an input layer, output layer, and a hidden layer. The numbers of input neurons and output neurons are not restricted. By conducting the trial and error process number of neurons in the hidden layer are determined. The hidden layer develops arbitrary mapping between input and output layers.

Multilayer feed-forward network is used to estimate the construction cost overrun. The architecture of multilayer feed-forward is shown in fig 4. This network uses log sigmoid transfer function and tan sigmoid transfer function to calculate the output of each neuron in the hidden layer and output layer respectively. But the input layer calculates the output through the linear transfer function.



$$0_j = f(u_j) = f(\sum_{k=1}^{n} I_k w_{kj})$$

Fig -4: Architecture of Multilayer Feed-Forward Network

Multilayer feed-forward network is trained using a back-propagation algorithm that uses a gradient descent approach. Back-propagation algorithm is done to minimize the error between predicted output and actual output. During training, the network undergoes a large number of iterations to reduce the error. After each iteration, errors are propagated back, to update the connecting weights between the neurons of different layers. Training of the network does not stop until the output predicted has converged with an acceptable level. Designing of the neural network includes defining the problem, deciding method to gather the information, defining the network, training the network and testing the trained network.

The following steps help in the development of the prediction model for cost overrun,

1. Conducting the Literature survey to define the problem statement and gathering the information on the topic.
2. Listing the factors that influence the cost overrun. Further factors are categorized into time-dependent and time-independent.
3. Time independent factors such as total floor area, the budget amount, number of floors.
4. The time-dependent factors are actual duration, percentage changes in contract amount and percentage of budget spent.
5. These factors are selected after the questionnaire survey and in-depth discussion with the experts.
6. These factors are used as input variables in a neural network.
7. The database is created to store factors that are affecting a large number of projects.
8. For the accurate results, all the factors are normalized in the range of 0 to 1 using the normalization formula.
9. For the Learning process of the model, certain steps have to be followed. Such as Initialising Parameters. These are set accordingly in a pattern to facilitate the learning process.
10. The Input Data is randomly divided. And the K-fold or 10-fold cross-validation method is used for the learning validation. Each subsample, in turn, is used as testing data and remaining as training subsamples.
11. The model is trained by Neural Network-Long Short Term Memory (NN-LSTM) using time-dependent and time-independent variables. The model's performance is measured to evaluate error between predicted output and actual output as mentioned by Cheng et al. [8].
12. Fitness evaluation is done using the Mean Square Error (MSE) equation. Optimal results are received after the required number of iterations. Few performance measures are used to measure the accuracy of the model.

A case study is taken into consideration:

One reinforced concrete project, selected from a Taiwan construction company, was. The potential factors influencing construction duration framed how case data are analyzed. Key information was derived from the collected data for 22 periods. This study used the best performance of training.

Estimated Schedule at Completion (ESTC) Prediction

NN-LSTM model was used to predict the time remaining to complete the projects at different periods. From the actual and predicted results, the accuracy of the models were determined by calculating the performance indices. MAPE was 2.9%, MAE was 0.69%, and RMSE was 0.031952, while R is 99.47%.

Calculation of Estimated Schedule at Completion (ESAC),

$$ESAC = (AT + ESTC) \times CD$$

Where AT = actual time; and CD = contract duration, as provided in the project contract information. An estimation was done midway through the project, which, in this case, was the 11th period. If schedule progress is accepted, then the project may continue at the same pace. However, the possibility of delay will impact actions to mitigate or avoid this delay. The decision on whether to react to the results will be based on effectiveness of the reaction. Using the results of the model, the ESAC at the 11th period was determined:

Actual value: AT = 0.621583; ESTC = 0.480576; And CD = 695 days

Actual ESAC: (0.621583 + 0.480576) × 695 = 766 days

Predicted value: AT = 0.621583; ESTC = 0.47441; And CD = 695 days

Predicted ESAC: (0.621583 + 0.47441) × 695 = 762 days

Variation from the contract duration is determined by obtaining the difference between the actual and the predicted ESAC and contract duration as follows:

Variation from the actual ESAC: 695 − 766 = −71

Variation from the predicted ESAC: 695 − 762 = −67

Negative variation value implies, the project is behind schedule.

Decision-Making

From the calculated values, this project is behind the schedule within 71 days of commencing, while the prediction results indicate that, the project will be 67 days behind schedule at completion. Project delays are caused due to various factors, which is previously indicated. The delay may lead to losses due to penalties from or conflicts between client and contractor. Project crashing, a schedule compression technique that analyzes cost and schedule trade-offs, is one way of potentially avoiding delays. The time to stop crashing is when it is no longer cost-effective, which is either when the additional cost begins to exceed the consequence cost or when no further schedule compression is possible. The consequence cost can come in the form of penalty fees due to contract duration. For the study, the penalty that assumed was 0.1% of the contract amount for each additional day calculated:

Contract amount : NTD 153,500,000

Penalty per day : 153,500,000 ×0.1/100= NTD153,500.00

Actual penalty : 71 × 153,500 = NTD10,898,500.00

Predicted penalty : 67 × 153,500 = NTD10,284,500.00

The preceding figures shown that the difference between actual penalty and predicted penalty is NTD 614,000, which is in a reasonable margin. These predicted penalty results gave project managers an idea about what to expect at the end of the project if current progress is maintained. It directed them to action, to put proper measures in places in order to mitigate the risk of delays. On the side, failure to take this prediction step will lead the project to continue on the same planned course, resulting in a penalty assessment on the contractors of NTD 10,898,500.

3. CONCLUSION

Many factors influence the Cost Overrun in a project. Time-dependent and Time independent are the two main categories. The essential benefit of the model based on ANN is, it uses a larger number of variables, so the model becomes more diverse. ANN has proven useful and suitable for complex problems. It is a user-friendly predictive model. ANN can rearrange patterns found in the data to provide a larger opportunity for investigating different options and project control techniques. Neural networks can be an alternative modelling technique for problems that may include a higher degree of uncertainty in the data and when statistical analysis may not be practical.

REFERENCES

1. Larsen, J. K., Shen, G. Q., Lindhard, S. M., & Brunoe, T. D. (2016). *Factors Affecting Schedule Delay, Cost Overrun, and Quality Level in Public Construction Projects. Journal of Management in Engineering*, 32(1), 04015032.
2. Senouci, A., Ismail, A., & Eldin, N. (2016). *Time Delay and Cost Overrun in Qatari Public Construction Projects. Procedia Engineering*, 164, 368–375.
3. Bhargava, A., Anastasopoulos, P. C., Labi, S., Sinha, K. C., & Mannering, F. L. (2010). *Three-Stage Least-Squares Analysis of Time and Cost Overruns in Construction Contracts. Journal of Construction Engineering and Management*, 136(11), 1207–1218.
4. Hinze, J., Selstead, G., and Mahoney, J. P. _1992_. "*Cost overruns on the state of Washington construction contracts." Transp. Res. Rec.*, 1351,87–93.
5. Rowland, H. _1981_. "*The causes and effects of change orders on the construction process.*" Ph.D. thesis, Georgia Institute of Technology, Atlanta.
6. Chang, A. S. _2002_. "*Reasons for cost and schedule increase for engineering design projects.*" *J. Manage. Eng.*, 18_1_, 29–36.
7. Akpan, E. O. P., and Igwe, O. _2001_. "*Methodology for determining price variation in project execution.*" *J. Constr. Eng. Manage.*, 127_5_, 367–373.
8. Akinci, B., and Fischer, M. _1998_. "*Factors affecting contractors' risk of cost overburden.*" *J. Manage. Eng.*, 14_1_, 67–76.
9. Karla Knight and Aminah Robinson Fayek, A.M.ASCE. '*Use of Fuzzy Logic for Predicting Design Cost Overruns on Building Projects' Journal of Construction Engineering and Management.* Vol 128, No. 6, December 1, 2002. ©ASCE.
10. J. Li; O. Moselhi; and S. Alkass. forecasting Project Status by *Using Fuzzy Logic'. Journal of Construction Engineering and Management*, Vol.132, No. 11, November 1, 2006. ©ASCE, ISSN 0733-9364/2006/11-1193–1202.
11. Attalla, M., & Hegazy, T. (2003). *Predicting Cost Deviation in Reconstruction Projects: Artificial Neural Networks versus Regression. Journal of Construction Engineering and Management*, 129(4), 405–411.
12. Kim, G.-H., An, S.-H., & Kang, K.-I. (2004*). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Building and Environment*, 39(10), 1235–1242.
13. Adriana V. Ordóñez Oliveros and Aminah Robinson Fayek, A.M.(2005) *Fuzzy Logic Approach for Activity Delay Analysis and Schedule Updating. Journal of Construction Engineering and Management, Vol. 131, No. 1, January 1, 2005. ©ASCE*
14. Lorterapong, P., and Moselhi, O. (1996). *"Project-network analysis using fuzzy set theory analysis." Journal of Construction Engineering and Management.*
15. Cheng, M.-Y., Chang, Y.-H., & Korir, D. (2019). *Novel Approach to Estimating Schedule to Completion in Construction Projects Using Sequence and Nonsequence Learning. Journal of Construction Engineering and Management*, 145(11), 04019072.