JRSSEM 2024, Vol. 04, No. 5, 1687 – 1698
E-ISSN: 2807 - 6311, P-ISSN: 2807 - 6494

Conceptual Cost Estimation In Green Buildings By Using Regression Analysis and Artificial Neural Network Methods To Improve Accuracy

Fauziyah Kamilah, Wisnu Isvara
Universitas Indonesia, Indonesia
Email: fauziyahkamilah@gmail.com, wisnu.isvara@gmail.com
*Correspondence: fauziyahkamilah@gmail.com

ABSTRACT: Conceptual cost estimation is a critical task during the early stages of a construction project, especially for green buildings, which present unique sustainability challenges and design complexities. Traditional methods such as regression analysis are widely used but often rely on experienced estimators and are time-consuming, while Artificial Neural Networks (ANN) offer a modern alternative but are limited by the quality and quantity of available data. This study aims to develop a hybrid model combining regression analysis and ANN to improve the accuracy of conceptual cost estimation for green and conventional high-rise buildings in Indonesia. Using data from 22 high-rise building projects (13 conventional and 9 green buildings), the study employed regression analysis, ANN techniques, and a combination of the two, with eight key variables selected for modeling. The hybrid model demonstrated the highest accuracy, achieving a Mean Absolute Percentage Error (MAPE) of 15.09%, within the acceptable range for conceptual cost estimation (+10–30%) as per AACE standards, outperforming standalone regression and ANN models. These findings highlight that integrating regression analysis and ANN provides a robust tool for early-stage cost estimation, supporting sustainable construction practices and informed decision-making.

Keywords: conceptual cost estimate, green building, artificial neural network, regression

INTRODUCTION

Conceptual cost estimation is carried out in the early stages of a project when information regarding scope, design, and specific requirements is limited. According to Wideman R.M. (2001) Conceptual stage estimation is the first stage of a project where project needs begin to be described. The purpose of cost estimation itself is to provide a rough estimate to guide decision making and budgeting.

The elemental method is the most commonly used method in cost estimation at the conceptual stage. Statistical methods like regression analysis are used to estimate costs but it will depend on estimator experience and need more time. Studies have developed parametric models to estimate pre-construction costs (Cheng, Tsai, & Sudjono, 2020). Regression analysis has been used to estimate parametric costs for construction projects. Statistical methods such as regression analysis are conventionally used in the literature for cost estimation. Several studies have developed parametric models for estimating pre-construction costs. Regression analysis has been used for parametric cost estimation of construction projects (Sonmez & Ontepeli, 2009).

Artificial intelligence is applicable to cost estimation or forecasting problems. Neural networks can identify relationships between parameters and costs. Neural networks can

perform poorly with limited data. Regression and artificial neural network techniques can lead to satisfactory parametric models. Regression combined with artificial neural networks is more accurate than regression and ANN methods. The expected accuracy for conceptual estimation is +10–30% (AACE 56R-20).

In this era green building is commonly used as a requirement According to Christiano Utomo (2022), current urban population growth has long-term impacts on the environment and natural resources. Buildings account for 32% of total global energy use. The concept of green building helps slow the rate of global warming by modifying the microclimate.

As the world becomes more interconnected, the influence of organizational culture on employee behavior and performance has gained significant attention. In the context of businesses and educational institutions, understanding how organizational culture shapes individual efficacy and collective commitment is essential (Kesturi, 2022). Employees' perception of their work environment, along with their belief in their ability to succeed, can directly affect their commitment to the organization. This interconnectedness between culture, self-efficacy, and commitment is particularly critical in sectors where performance and collaboration are vital to success (Yu & Skibniewski, 2020).

Moreover, the relationship between job satisfaction and these psychological factors has been a subject of extensive research. Job satisfaction serves as a key moderating variable, as it enhances the impact of organizational culture and self-efficacy on organizational commitment (Sharma, Najafi, & Qasim, 2013). When employees are satisfied with their work environment, they are more likely to demonstrate stronger commitment and higher performance levels. Understanding how job satisfaction plays this role is crucial for organizations aiming to boost employee engagement and improve overall outcomes (Bates et al., 2015).

The concept of organizational culture itself is multifaceted, encompassing shared values, beliefs, and practices that shape behavior within an organization. A strong organizational culture fosters a sense of belonging and alignment among employees, making them more likely to work toward common goals (Hegazy & Ayed, 2018). It also influences their perception of the organization's mission and their personal role in achieving these objectives. For this reason, it is essential to analyze how various cultural dimensions interact with other factors such as self-efficacy and job satisfaction to drive organizational commitment (Kim, An, & Kang, 2014).

Self-efficacy, or the belief in one's ability to succeed, also plays a pivotal role in employee performance. Research has shown that individuals who possess high self-efficacy are more resilient in the face of challenges and more likely to engage in proactive behaviors (Latief, Wibowo, & Isvara, 2023). In an organizational context, fostering self-efficacy can enhance employees' motivation and productivity, making it an important factor for both personal and organizational growth. Thus, understanding how self-efficacy interacts with organizational culture and job satisfaction is crucial for improving performance and commitment (Setyawati, Creese, & Sahirman, 2023).

The urgency of examining these factors in the context of organizational performance cannot be overstated. In today's competitive business environment, organizations must create an environment where employees feel empowered and motivated to contribute to the organization's success (Haykin, 2018). By investigating the relationships between organizational culture, self-efficacy, job satisfaction, and organizational commitment, this research aims to offer valuable insights that can inform management practices and contribute

to the development of more effective workplace strategies (Liu, Rasdorf, Hummer, Hollar, & Parikh, 2013).

Existing studies primarily address conceptual cost estimation using either regression analysis or artificial neural networks in isolation. While regression models provide interpretable relationships, they often lack predictive accuracy, especially for non-linear data patterns. On the other hand, ANN methods require large datasets to perform effectively, which can be a limitation in specific construction contexts. This research fills the gap by exploring the effectiveness of a combined regression-ANN approach, tailored specifically for green buildings in Indonesia—a sector where data scarcity and sustainability requirements present unique challenges.

The increasing global emphasis on green building practices to mitigate environmental impacts has made accurate cost estimation more critical than ever. Green buildings, with their distinct design and construction complexities, require precise early-stage cost estimations to ensure financial feasibility. In Indonesia, where sustainable development is gaining traction, the lack of accurate and efficient cost estimation methods hinders widespread adoption. This research addresses this urgency by developing a reliable model to support informed decision-making, budget planning, and sustainable project execution in the green construction sector.

This study provides a novel approach by integrating regression analysis and artificial neural networks (ANN) to improve the accuracy of conceptual cost estimation for green and conventional high-rise buildings in Indonesia. Unlike previous research that focuses solely on either statistical models or neural networks, this study demonstrates the synergistic advantages of combining these methodologies, yielding a model with a Mean Absolute Percentage Error (MAPE) of 15.09%, outperforming standalone approaches. This integration bridges the gap between traditional parametric estimation and machine learning techniques, offering a hybrid model tailored for early-stage project budgeting in green building construction.

The primary objective of this study is to develop an accurate and practical conceptual cost estimation model by combining regression analysis and artificial neural networks. The study aims to reduce the prediction error margin and provide a robust tool for estimating costs in both green and conventional building projects. The benefits of this research include providing construction professionals with a reliable estimation tool that enhances financial planning, supports the adoption of green building practices, and streamlines decision-making processes. Additionally, the findings offer valuable insights for future research in leveraging hybrid methodologies for cost estimation, contributing to the advancement of sustainable construction practices.

RESEARCH METHODOLOGY

As with the traditional parametric cost estimation model, the present model also uses regression analysis on collected historical data to define significant building parameters as key cost drivers. This use of stepwise regression is recommended to address multicollinearity issues among input variables (Ji, Li, Xie, Wu, & Zhou, 2013) that are common in statistical data analysis. This outpu will also be use in third model NN2.

A neural network is a system that simulates the human brain's learning process. It consists of interconnected neurons. These neurons interact with each other through weighting. NN are trained with examples and target values. The hidden and output layer neurons process inputs by multiplying them by corresponding weights and using a non-linear transfer function

to produce results. The sigmoid function is commonly used for this. Neurons adjust their weights in response to errors between actual and target output values.

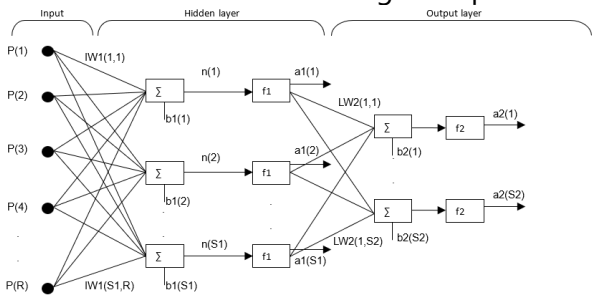


Figure 1. Architecture configuration of neural network

A representative example of an artificial neural network is illustrated in Figure 1. The network illustrated above comprises R inputs, represented by R neurons in the input layer, S1 neurons in the hidden layer, and S2 neurons in the output layer. The number of hidden layers may be varied according to the specific requirements of the application. A network may comprise multiple layers. The outputs of each intermediate layer serve as the inputs to the subsequent layer. Each layer is comprised of a weight matrix W, a bias vector b, and an output vector a. Each element of the input vector p is connected to each neuron input through the weight matrix W. The ith neuron has a sum function that gathers its weighted inputs and bias to form its own scalar output n(i). The various n(i) taken together form an S-element vector n. Finally, the neuron layer outputs form a column vector a.The column vector a1 as the input of layer 2 can be determined by equation (1):

$$a1 = f1(IW1*P+b1) \quad (1)$$

And the vector a2 as the output layer can be determined by formulation given in equation (2):

$$a2 = f2(LW2*a1+b2) = f2(LW2*(f1(IW1*P+b1))+b2) = Y \quad (2)$$

The layers of a multilayer network play different roles. The layer that produces the network output is called an output layer. The layer that gets the inputs is called input layer. All other layers are called hidden layers. It is common for the number of inputs to a layer be different from the number of neurons.

The operation of a neural network can be divided into two principal phases: learning and training. The learning process entails the adaptation of connection weights in response to a set of examples presented at the input layer and, optionally, at the output layer. The objective is to identify a distinct set of weights that can accurately associate all example patterns utilized during the learning process with their corresponding desired output patterns. The multilayer perceptron is the most prevalent neural network model. It is also referred to as a supervised network, as it is presented with input examples and their corresponding desired responses. In this case, the desired outputs are employed to instruct the network in the appropriate responses. The multilayer perceptron employs the backpropagation algorithm for learning. This algorithm incorporates a learning algorithm, the generalized delta rule, which is responsible for training the network.

This rule uses a gradient descent method to determine a unique set of network weights that enable the network to produce outputs that are very close to the desired outputs associated with a number of training examples. The backpropagation algorithm typically employs a non-linear sigmoid transfer function to calculate the output of each neuron, with the exception of those designated as input neurons. In the context of training, the term refers to the process of repeatedly applying input vectors to the network and calculating errors with

respect to the target vectors. This is followed by the identification of new weights and biases through the application of a learning rule. This process is repeated until the sum of the squared errors falls below a specified threshold or the maximum number of cycles/epochs has been reached. Upon completion of the training phase, the neural network will have been developed into a model that is capable of predicting a target value based on an input value.

The operation of a neural network can be divided into two principal phases: learning and training. The learning process entails the adaptation of connection weights in response to a set of examples presented at the input layer and, optionally, at the output layer. The objective is to identify a distinct set of weights that can accurately associate all example patterns utilized during the learning process with their corresponding desired output patterns. The multilayer perceptron is the most prevalent neural network model. It is also referred to as a supervised network, as it is presented with input examples and their corresponding desired responses. In this case, the desired outputs are employed to instruct the network in the appropriate responses. The multilayer perceptron employs the backpropagation algorithm for learning. This algorithm incorporates a learning algorithm, the generalized delta rule, which is responsible for training the network.

This rule uses a gradient descent method to determine a unique set of network weights that enable the network to produce outputs that are very close to the desired outputs associated with a number of training examples. The backpropagation algorithm typically employs a non-linear sigmoid transfer function to calculate the output of each neuron, with the exception of those designated as input neurons. In the context of training, the term refers to the process of repeatedly applying input vectors to the network and calculating errors with respect to the target vectors. This is followed by the identification of new weights and biases through the application of a learning rule. This process is repeated until the sum of the squared errors falls below a specified threshold or the maximum number of cycles/epochs has been reached. Upon completion of the training phase, the neural network will have been developed into a model that is capable of predicting a target value based on an input value

RESULT AND DISCUSSION

Project data consist of 22 Highrise building consist of 13 conventional building and 9 green building. The projects located in Indonesia. The datasets were divided into two parts by random sampling. The first group of data (20 datasets) as data training were used to develop the model and the second group (2 datasets) as data testing that were used to test the model. And a total of 8 input variables were identified whereas the output variable is the contractual construction costs (in IDR), (See Table 1).

Because we have variety of data in terms of location and time of construction, data normalisation is required to ensure that cost data is to ensure that the cost data are on the same basis. In this study, all construction costs in Table 1 have been adjusted to December 2018. Reference using the following formula

formula:

$$C(t)\ f(n)\ f(1).C(r).I(t)/I(r1) + f(2).C(r).I(t)/I(r2) + \ldots f(n).C(r).I(t)/I(rn) \quad (1)$$

where :

f(n) : fraction of time during which construction takes place within year n (if construction of the building takes more than more than 1 year)

C(t) : cost at time of interest, C(r) = cost at reference time

I(t) : cost index at time and place of interest

I(r) : cost index at time and place of reference

As the construction cost index is not yet available in Indonesia, the Consumer Price Index (CPI) was used as a proxy for cost index measures that published by Indonesia Central Bureau of Statistics. Location will use CPI to December 2018 based on each of the area. The same CPI, December 2018 will Be use as numeric for location

Table 1. Variables description

Variables	Description	Range
P1	Location	132,13 = Bekasi ; 135,24 = Surabaya ;135,25 = Jakarta; 139,03= Balikpapan.
P2	Gross floor area	24938,00 - 113020,00 m^2
P3	Number of storey	12– 43
P4	Number of basement	2– 7
P5	Finishing grade	1 = Low, 2 = Middle, 3 = High
P6	Type of foundation	1 = bored pile, 2 = Spun pile
P7	Building Category	1= conventional building , 2 = Green Building
P8	Duration of construction	2-5 years
Y (Output)	construction cost (IDR x 1000)	185433537-2630294208

Regression Analysis

The significance of the primary parameters was assessed in the in the first part of the regression analysis. This paper will use both linear and non-linear regression since In this study, there is no definite theory stating that the relationship between the independent and dependent variables is sinusoidal, logarithmic and so on, therefore, analysis and justification are needed to determine the relationship between variables so that the most dominant variable can be identified.The result as shown in Table 2

Tabel 2. Regression Result

			R	R Square	Adjusted R Square
LINEAR REGRESSION			.912	0,832	0,814
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-9,99E+08	2,04E+08		-4,894	1,37E-04
GFA	1,81E+04	2,85E+03	0,684	6,359	7,14E-06
Building Category	5,42E+08	1,65E+08	0,354	3,293	4,30E-03
			R	R Square	Adjusted R Square
NON- LINEAR REGRESSION			.886	0,785	0,773
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-6,48E+17	3,03E+17		-4,894	1,37E-04
GFA	4,76E+08	5,87E+07	0,886	8,099	2,06E-07

Neural Network

Matlab R2009a software was chosen for developing the neural network model due to its ease of use, speed, and a variety of architectures, including backpropagation. A neural network method was used to develop 22 variants with varying hidden layers and neurons. The experiments set parameters as described in Table 3.

Tabel 3. Neural Network Parameters 1

Parameters	Values	Description
Architecture configuration parameter:		
No of input neurons	8	Number of input variables
No of output neuron	1	Number of output variable
No of hidden layers	1 and 2 hidden layers	
No of neurons in hidden layers	Min 2 and max 12	
Learning algorithm	Backpropagation	
Activation function	Sigmoid bipolar	
Learning function	Gradien Descent Momentum	
Learning rate	0.001	
Maximun epoch	30000	
Goal (min MSE)	10^{-5}	

Performance Of Model

The accuracy performance of all models using data testing is based on Mean Absolute Percent Error (MAPE), given in equation (2):

$$MAPE = \frac{1}{n}\sum \frac{\left|actual\ \cos t - predicted\ \cos t\right|}{actual\ \cos t} x100\% \qquad (2)$$

Performance Of Regression

We use model of Linear regression (R1) because it has better R and the predictors concist of GFA and Building Categoric (Conventional Building or Green Building)

MAPE R1 | 16,60%

Performance Of Neural Network

According to Cerpa and Walczak (2023) When choosing the number of nodes to be contained in a hidden layer, there is a trade-off between training time and the accuracy of training. A greater number of hidden unit nodes results in a longer (slower) training period, while fewer hidden units provide shorter (faster) training, but at the cost of having fewer feature detectors. Too many hidden nodes in an ANN enable it to memorize the training data set, which produces poor generalization performance. Some of the heuristics used for selecting the quantity of hidden nodes for an ANN are using:

- 75 percent of the quantity of input nodes,
- 50 percent of the quantity of input and output nodes, or
- 2*n* + 1 hidden layer nodes where *n* is the number of nodes in the input layer.

This paper also use 25 percent and 100 percent nodes expected a better modelling but won't use 2n + 1 since this paper only use 22 data.The performance of this Neural network with 8 predictors model(NN1) shown in Table 4

Table 4. Summary of Performance NN1

Architecture of configuration	MAPE data training	MAPE data testing
8-2-1	20,58%	17,82%
8-4-1	23,31%	20,36%
8-6-1	27,01%	27,22%
8-8-1.	45,54%	31,13%

8-2-1 is the best model configuration . According to Holm et al (2021) the expected accuracy range of conceptual estimate is ±10-20%, while AACE 56R-20(2020) states that the accuracy range of budget estimate is ±10-30%. Comparing with these references, the proposed model has performed well, although the amount of data training was limited.

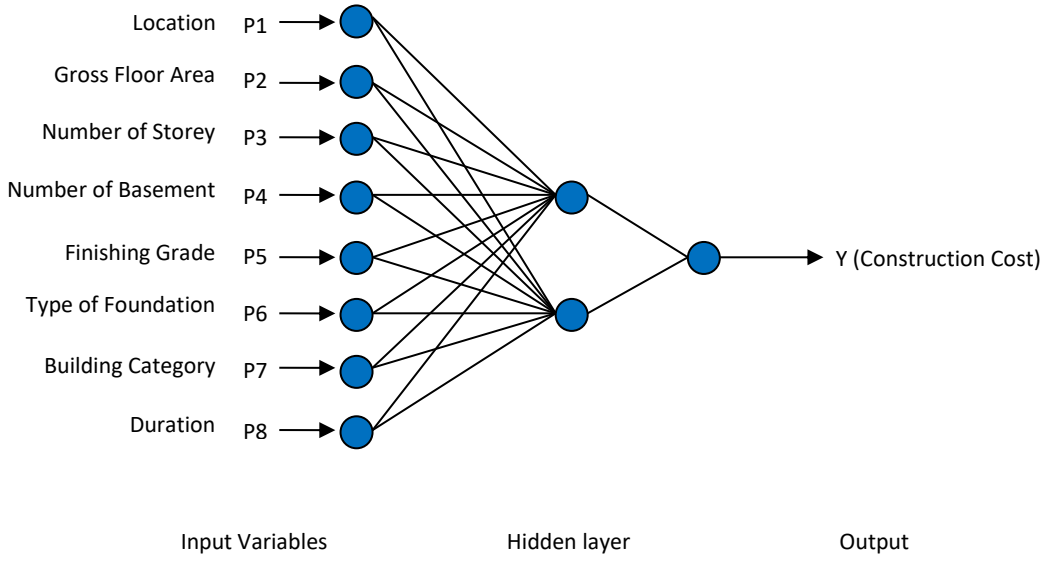
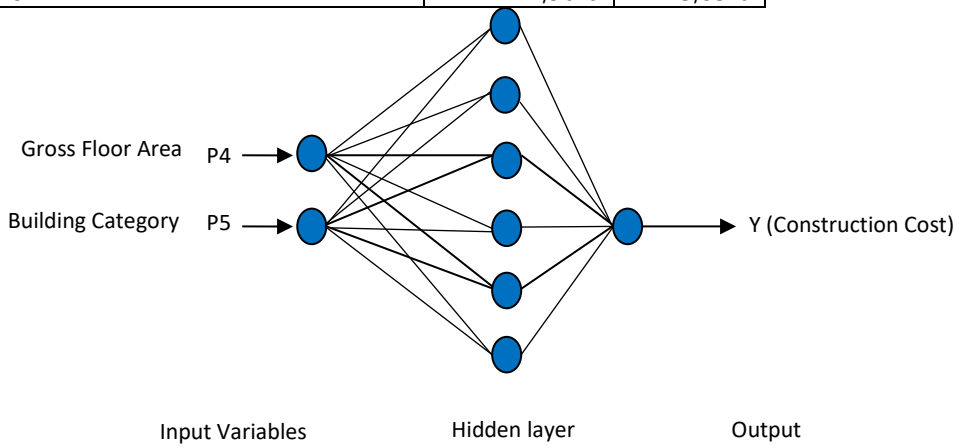


Figure 2. 8-2-1 Neural Network Model

Combination Of Regression And Neural Network

Using the same parameter and nodes with model NN1. This model (NN2) only have two predictors as a result of regression. The result shown in Table

Architecture of configuration	MAPE data training	MAPE data testing
2-2-1	17,27%	29,40%
2-4-1	28,10%	22,37%
2-6-1	21,69%	15,09%
2-8-1	12,90%	15,85%



2-6-1 is the best model configuration . According to Holm et al (2005) the expected accuracy range of conceptual estimate is ±10-20%, while AACE 56R(2020) states that the accuracy range of budget estimate is ±10-30%. Comparing with these references, the model NN2 has also performed well.

CONCLUSION

Three conceptual cost estimation models were developed using data from conventional and green buildings in Indonesia, which could be used to estimate early project costs when construction drawings are not available and detailed cost estimates cannot be made. Regression analysis and neural network techniques were used to develop parametric models. Depending on the project data and the relationships between parameters and costs, each technique has certain advantages.

REFERENCES

Bates, Jennifer, Burton, C. C. E. Dorothy J., Creese, Robert C., Hollmann, John K., Humphreys, Kenneth K., McDonald Jr, Donald F., & Miller, C. Arthur. (2015). Cost estimate classification system–as applied in engineering, procurement, and construction for the process industries. *AACE International Recommended Practice*, *18*.

Cheng, Min Yuan, Tsai, Hsing Chih, & Sudjono, Erick. (2020). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications*, *37*(6), 4224–4231.

Haykin, Simon. (2018). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Hegazy, Tarek, & Ayed, Amr. (2018). Neural network model for parametric cost estimation of highway projects. *Journal of construction engineering and management*, *124*(3), 210–218.

Holm, Len, & Schaufelberger, John E. (2021). *Construction cost estimating*. Routledge.

Ji, Zhonghui, Li, Ning, Xie, Wei, Wu, Jidong, & Zhou, Yang. (2013). Comprehensive assessment of flood risk using the classification and regression tree method. *Stochastic environmental research and risk assessment*, *27*, 1815–1828.

Kesturi, Ludya. (2022). Estimasi Biaya Tahap Konseptual Pada Konstruksi Gedung Perkantoran dengan Metode Artificial Neural Network. *Skripsi Program Sarjana Universitas Indonesia, Jakarta*.

Kim, Gwang Hee, An, Sung Hoon, & Kang, Kyung In. (2014). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and environment*, *39*(10), 1235–1242.

Latief, Yusuf, Wibowo, Andreas, & Isvara, Wisnu. (2023). Preliminary cost estimation using regression analysis incorporated with adaptive neuro fuzzy inference system. *International Journal of Technology*, *1*, 63–72.

Liu, Min, Rasdorf, William, Hummer, Joseph E., Hollar, Donna A., & Parikh, Shalin C. (2013). Preliminary engineering cost-estimation strategy assessment for roadway projects. *Journal of Management in Engineering*, *29*(2), 150–157.

Setyawati, Bina R., Creese, Robert C., & Sahirman, Sidharta. (2023). Neural networks for cost estimation (Part 2). *AACE International Transactions*, ES141.

Sharma, Jwala R., Najafi, Mohammad, & Qasim, Syed R. (2013). Preliminary cost estimation models for construction, operation, and maintenance of water treatment plants. *Journal of Infrastructure Systems*, *19*(4), 451–464.

Sonmez, Rifat, & Ontepeli, Bahadir. (2009). Predesign cost estimation of urban railway projects with parametric modeling. *Journal of Civil Engineering and management*, *15*(4), 405–409.

Walczak, Steven, Pofahl, Walter E., & Scorpio, Ronald J. (2023). A decision support tool for allocating hospital bed resources and determining required acuity of care. *Decision support systems*, *34*(4), 445–456.

Yu, Wen der, & Skibniewski, Mirosław J. (2020). Integrating neurofuzzy system with conceptual cost estimation to discover cost-related knowledge from residential construction projects. *Journal of Computing in Civil Engineering*, *24*(1), 35–44.