Cost Estimation Model for Highway Projects using ANN

Lily Gurung*[1], Manoj Chhetri[2]

[1]Civil Engineering Department, College of Science and Technology, Royal University of Bhutan
[2]Information Technology Department, College of Science and Technology, Royal University of Bhutan
Email: lilygurung.cst@rub.edu.bt*[1], manoj_chhetri.cst@rub.edu.bt[2]

Abstract
Construction and maintenance works are carried out by contractors for a period of few months to few years. Those contracts show cost variations between the actual and estimated costs. For example, cost estimation of 3 out of 5 projects from 2015-19 in Trashigang Dzongkhag have not only been underestimated but the actual costs are on average 27% higher than the estimated costs. Moreover, researches reveal that there are issues in cost estimation at the conceptual stage affecting the project progress and cost (Flyvberg, 2002). In this study neural network has been used to predict the actual cost based on the initial estimation and duration. During the study, it was discovered that the neural network demonstrated its usefulness as a tool for cost estimation and could be effectively employed by decision makers. However, it is important to note that the neural network is not designed to replace the entire cost estimation method. Instead, it should be regarded as one of the tools to be utilized alongside other methods when estimating costs.
Key words: *ANN, Highway projects, Cost estimation, RMSE*

1. INTRODUCTION

Until the beginning of 1960s, Bhutan had been isolated from the rest of the world in terms of basic infrastructure of transportation. Currently, there are 18,264.60 kms of roads of various categories which are constructed and maintained by various agencies. A massive share of national budget is kept aside for construction and maintenance of roads. According to Business Bhutan, the road maintenance works amounted to Nu. 2666.54 million, from 2013-14 to 2017-18.

Construction and maintenance works are carried out by contractors for a period of few months to few years. Those contracts show cost variations between the actual and estimated costs. For example, cost estimation of 3 out of 5 projects from 2015-19 in Trashigang Dzongkhag have not only been underestimated but the actual costs are on average 27% higher than the estimated costs. Moreover, researches reveal that there are issues in cost estimation at the conceptual stage affecting the project progress and cost (Flyvberg, 2002).

Cost estimation is an essential component in any infrastructure project and development of a more accurate estimation technique for road projects at the initial phase itself is inevitable. This research proposes to develop a model to accurately estimate the cost of road projects using Artificial Neural Network (ANN) as a tool. Road is the lifeline of a country's economy; therefore, it is essential to evaluate the cost variations and formulate a benchmark to successfully evaluate new road projects and help the government from cost overruns.

2. LITERATURE REVIEW

Accurate cost estimation of road projects at the initial stage is very vital for precise planning and feasibility check. Several researchers have used artificial neural network and estimated construction costs for various projects. Tijanic et. al. (2019) elaborated the road construction cost overruns in Republic of Croatia. They used the database of roads constructed on the territory of Croatia to establish a model to generate road construction costs using neural networks namely multilayer perceptron (MLP), radial basis function neural network (RBFNN) and generalized regression neural network (GRNN).

Mahalakshmi and Rajasekaran (2019) developed ANN model to predict the construction cost of highway with 52 projects' data from National Highway Authority of India. It was concluded that MLP with backpropagation algorithm was capable of predicting the cost with an accuracy of 95.2%.

Faiq and Ibraheem (2017) worked on forecasting the cost of structure of infrastructure projects using ANN model during the feasibility study and taking highway projects in particular as a case study. They developed a model using the data from Stat Commission for roads and bridges in republic

of Iraq, which predicted the project cost to an accuracy of 93.19%.

Hashemi et. al. (2020) reviewed the articles published within years from 1985 to 2020, which used the machine learning techniques to estimate cost and predict construction projects. They concluded that most of the prediction techniques met the expectations and could help decision makers. Among the various methods, ANN and RA were found to be the most popular that were used in the reviewed papers.

Shemi and Ashok (2020) examined the cost overrun in highway projects by developing regression and artificial neural network models. The research considered importance of the causes influencing the accuracy of the cost estimation. It was concluded that the best results were obtained by using 10 inputs, 13 hidden layers and one output with the help of MATLAB software. Sandhya and Judit (2020) mentioned the importance the of cost estimation of civil engineering projects at the beginning of the project. They used ANN to estimate the cost of various buildings using excel solver and MATLAB software. Input parameters used were cost of the buildings, number of storeys, building area, storey height, dimensions, quantity, rate, and plan. Bachav and Saharkar (2020) states the huge investment required in the construction of roads and also the repair and maintenance of old roads. They tried to estimate the life cycle cost of bituminous and concrete roads using ANN and made a conclusion that the stable pavement has almost two times longer life cycle that of the comfortable pavement.

Therefore, Cost estimation is an integral part of any construction project. Developing a model to estimate the cost of a project at conceptual or initial stage will not only help decision making at early stages but also enable the projects to progress smoothly without any cost overrun.

3. METHODOLOGY

Highway projects in Bhutan often express cost overruns and developing a Model using artificial neural network (ANN) to estimate the cost of such projects will be a new beneficial venture for a developing country like Bhutan. ANNs as a state of the art, are a part of artificial intelligence which helps to predict or estimate results. Fig. 1 summarizes the research approach that was closely followed.

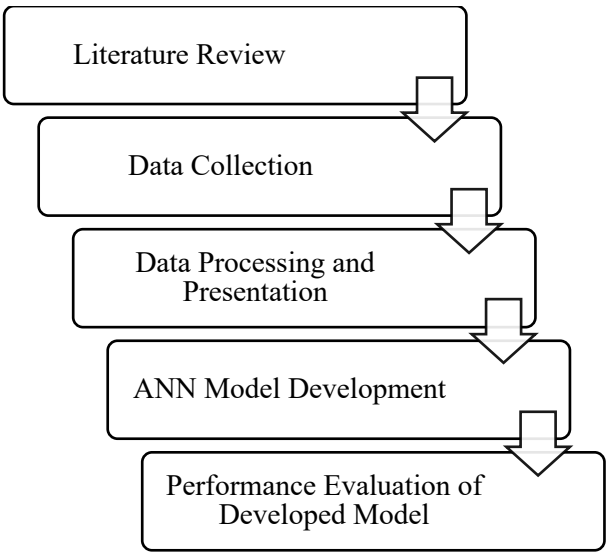


***Fig. 1**: Methodology*

There are several research projects which have adopted ANN for making decisions in highway or any construction project. A thorough literature review was carried out related to estimating or forecasting costs for highway projects. Documents published by government or any agencies like Department of Roads (DoR) and Thromdes were reviewed as well, to gather as much information as possible. From the literature review, the factors influencing the cost estimation of such projects, data processing method, ANN architecture and performance matrices was identified.

The preliminary data was collected from various agencies of Bhutan responsible for construction and maintenance of highway projects; agencies like Department of Roads (DoR), Thromdes and Gewog centers. The data was collected by surveying, interviewing engineers, contractors and collecting details of construction and maintenance of highway projects from Ministry of Works and Human Settlement.

After identifying the variables influencing the cost estimation from literature and data collection, the variables were processed and are presented as descriptive statistics. The data to be used for the research are presented in a tabular form.

One type of neural network was identified from literature review to be used to estimate the highway project cost. The selection was based on wide usage of the ANN and good accuracy with least errors. Number of input layers, output layers and number of neurons were accordingly identified while developing the model. Then the

type of training method was identified, followed by training and testing of the dataset.

4. COST ESTIMATION USING ANN

4.1. Data collection

Collection of data is typically difficult from government organization. Data for the research was collected after approaching multiple organizations with majority of organizations refusing to share data citing various reasons ranging from policy to non-compilation. Following agencies look after respective road networks and were the initial prospects for data collection:

- National Highways by Department of Roads (DoR)
- Dzongkhag Roads by Dzongkhag/ Dungkhag/ Gewog Administration
- Thromde Roads by Thromdes
- Farm Roads by Ministry of Agriculture and Forest (MoAF) / DoR
- Access Roads by concerned agencies/ communities

Dataset process varied including questionnaire surveys, phone surveys and interviews wherever questionnaire surveys fail. Since many engineers working in those organizations are previous students of College of Science and Technology, it was easier to collect survey data and some missing value data through phone calls. They were also of immense benefit while clarifying doubts. But nonetheless many of the engineers did not have the authority to share data and managements of some of those organizations did not share the data that they had. Parameters influencing cost estimation, duration of project, type of project, etc. were collected. Additional data for training the model was also be collected from published government documents and literature.

Although multiple private and government organizations like mentioned above were potential source of data, majority of the data that was used in this study was provided by Department of Roads, Ministry of Works and Human Settlement. The dataset we have are from 2017 to 2020.

The data collected was segregated into various files based on the year of allocation for construction work. These individual files were merged to form a unified Excel file. The combined data consisted of a total of 1030 records and encompassed 18 distinct parameters, including the serial number.

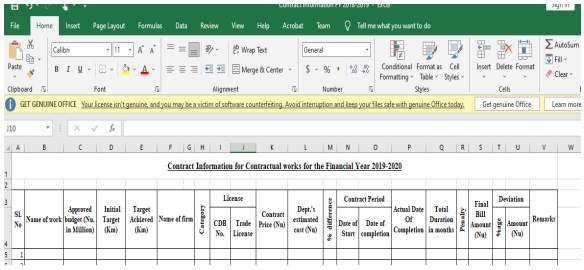


***Fig. 2**: Dataset collection in Excel Format*

4.2. Data preprocessing

a. Dataset Description and Data Cleaning

The initial dataset contained a total of 18 parameters and 1030 records. The specific parameters can be found in the figure below.

Sl.No	Name of work	Approved budget (Nu.)	Initial Target	Target Achieved	Name of firm	Category	License		Contract Price (Nu.)	Dept.'s estimated cost (Nu.)	% difference	Contract Period		Actual Date Of Completion	Total Duration in months	Penalty	Final Bill Amount (Nu.)	Deviation		Remarks
							CDB No.	Trade License				Date of Start	Date of completion					%age	Amount(Nu in million)	
1																				
2																				

***Fig. 3:** Format of data collected*

Not all parameters collected in the dataset were relevant to our study. Features such as "Actual data of completion" and "penalty", etc. are examples of parameters that are only known once the project is completed. Since our objective was to estimate the cost before the project is tendered out, we excluded these parameters from our study. “Contract price” was retained as it shows the lowest bid value considered for tendering.

Certain parameters, including "trade license number," "name of work," and "CDB no," lacked statistical significance in our study. As a result, we made the decision to exclude these parameters from our analysis.

In light of their significance, we introduced two additional parameters in our study: "dzongkhag" and the "month” in which the tender was allocated." Following the removal and inclusion of the above 2 parameters, we ultimately concluded our study with a set of 8 parameters, which encompassed the target variable "final cost."

The ordinal parameter "size of firm" had varying references, such as the use of notations like "L" or "Large." To ensure consistency, the values were cleaned and unified to have a

common reference throughout the column. Similarly, the initial duration of the project was initially mentioned using different time periods, such as months or years. To maintain uniformity, the time period was standardized to be expressed exclusively in months.

Table 1: *Final parameters and their datatype*

Name of Parameter	Data type
Dzongkhag	String
Size of the firm	Character [L, M, S]
Contract Price	float
Departments Estimated cost	float
%Difference	float
Total duration of project in months	int
Final bill	Float

b. Missing Values

The data collected had a lot of missing values. Many of these missing values were of paramount importance and hence the appropriate individuals were contacted to ask for the values. The authors did not use mean imputation or other imputation methods as it just creates bias in the dataset, rather the authors have manually filled the missing values by reaching out to the concerned person or through desktop research. Some missing values were missing completely at random (MCAR) and some had large portion of missing values hence imputation did not make much sense in our case. Some of the records with multiple missing values were dropped and cases and columns with large missing values were not considered for this study.

c. Outliers

Outliers are basically values which are so different from the other values such that, to an observer it feels like the values were actually recorded using a completely different method. While an outlier may indeed be valid and accurate, incorporating such values can significantly impede the generalization process of the neural network. The primary objective of this research is to enhance the neural network's ability to accurately assess the majority of cases, rather than placing emphasis on exceptional cases. In order to identify the outliers for continuous and integer variables, box plot analysis was used. Box plot basically gives 5 descriptions of the data as listed below:

- Minimum
- First quartile
- Median
- Third quartile, and
- Maximum.

Additionally, 2 whiskers linking the lowest and highest values were also used. While using outlier analysis using box plot and whisker plots, any values lying away from the two whiskers were considered as outliers. The outliers were removed from the dataset.

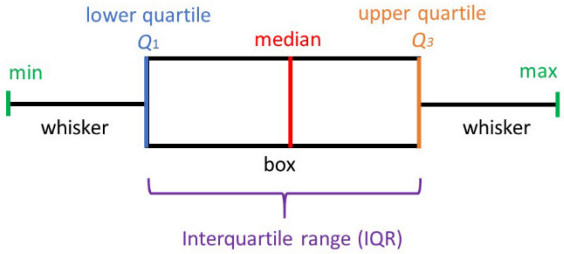


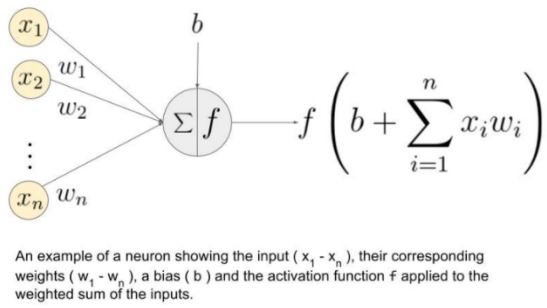
Fig. 4: *Bos plot*

d. Data Normalization

Data normalization is an important step as it improves convergence and gives equal importance to features. The categorical variables were normalized using one hot encoding and the other variables were normalized using min max scaling (Bahri,1990), which scales the values between 0 to 1.

4.3. Neural Network (NN)

The working of neural network is based on the human brain. The neural network consists of a number of neurons which simulate the neurons in a human brain. How an individual neuron works is, it takes all the inputs, these inputs are connected to a neuron by weighted edges. Here in this study the parameters or individual values in the excel column are the input values represented in the image below by 'x'. The inputs(x) are multiplied with a weight value of the edge. Initially the weight values are randomly initialized. After multiplication, the values are summed together and then it goes through an activation function which converts it to a value and only the values which are above a range are allowed to pass through to the next neuron thus mimicking an actual human neuron.



An example of a neuron showing the input (x_1 - x_n), their corresponding weights (w_1 - w_n), a bias (b) and the activation function f applied to the weighted sum of the inputs.

Fig. 5: *Working of a neuron*

The output is then compared to the original value. If there is a difference between the actual and predicted value, then the difference is calculated and the error is minimized by adjusting the weights. But the disadvantage of a single neuron is, it will not be able to predict or represent nonlinear data. This is overcome by including multiple neurons thus forming a neural network. In our study we used multi layered perceptron which is a universal approximator.

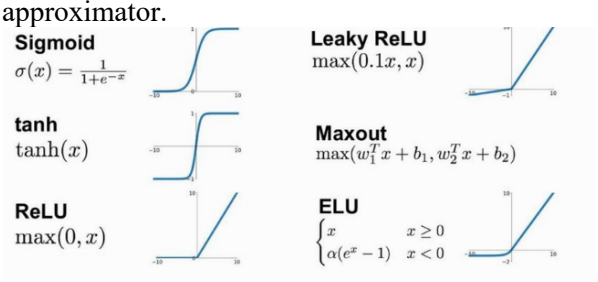


Fig. 6: *Activation functions*

In this study multiple configurations of neural network with multiple different sets of layers, learning rate and activation functions were tested. In our study, hyperparameters were optimized using grid search. The models were trained for 1000 epochs and early stopping was implemented with a patience value of 10. The selection of the final hyperparameter values was based on the evaluation method of RMSE, which was utilized in our study. The configuration of the final model looks as below:

Table 2: *Final parameters and their datatype*

Hyperparameter	Values
No of hidden Layers	2
Activation function [hidden layer]	Tanh
Activation function [output layer]	Linear
No of neurons	7-64-32-1
Epoch	1000[with early stopping]
Optimizer	Adam

5. RESULTS AND CONCLUSION

5.1. Training

The first step while using any supervised learning method is to train the algorithm. In this study, the neural network is trained by using 80% of the entire dataset. During training, both the input parameters and the output is shown to the network. After each row of data, the network will calculate the difference between the output value and the actual ground truth value that is provided to the network. This difference is corrected by back propagating the error by changing the values of the edges until the correct value is predicted. The same process repeats for every row of values. Once all the values are sent through the NN, it is considered as one epoch. The network was trained for 1000 epoch using early stopping with a patience value of 5. The network used 2 layers of hidden network with size of 64 and 32 with tanh activation function. Basically, the activation function checks whether the values are greater than 0 or not. If it is less than zero it returns 0 else the value it got. In our case we got the training RMSE (Root mean squared error) using equation (1) as 0.041. RMSE was used as the evaluation metrics.

5.2. Testing and Results

The remaining 20% of the dataset was used for testing purpose and the result obtained was considered as the accuracy of our model. The model was able to get the resultant RMSE value of 0.052.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - \hat{y}_i)^2}{n}} \quad \ldots\ldots (1)$$

Since there was no standard prediction rate to compare against, basic linear regression was used and found the proposed network to outperform linear regression by a large margin. The linear regression value was 0.081.

When we compared the RMSE of the "Departments Estimated cost" with the actual "final bill" the RMSE score was 0.09. Based on the results obtained from the study, NNs proved to be an effective tool for cost estimation and can be used by the highway projects beforehand to find the actual cost of the project based on the initial estimation and the duration of the project. Currently the neural network relies on the "estimated cost "carried out by the department and the MLP suggested

in this work is meant to be used as an additional tool for cost estimation and not a replacement. The authors intend to enhance the existing work by gathering supplementary data for the years 2021 and 2022. Additionally, we plan to investigate the reasons behind project delays, which were not within the scope of the current paper.

6. ACKNOWLEDGEMENT

The authors would like thank the Royal University of Bhutan for providing AURG funding to support this research. The authors are also grateful to the Department of Roads, Ministry of Works and Human Settlement for providing the data and other agencies as well for success of the research work.

REFERENCES

Al-Zwainy, F. M., & Aidan, I. A. A. (2017). Forecasting the cost of structure of infrastructure projects utilizing artificial neural network model (highway projects as case study). *Indian J. Sci. Technol, 10(20), 1-12.*

Bahri, A., & Li, Y. (1990). On a Min-Max Procedure for the Existence of a Positive Solution for Certain Scalar Field Equations in $\mathbb R^ N$. *Revista Matematica Iberoamericana,* 6(1), 1-15.

Adab, H., Kanniah, K. D., & Solaimani, K. (2013). Modeling forest fire risk in the northeast of Iran using remote sensing and GIS techniques. *Natural hazards*, 65, 1723-1743.

El-Kholy, A. M. (2019). Exploring the best ANN model based on four paradigms to predict delay and cost overrun percentages of highway projects. *International Journal of Construction Management*, 1-19.

Karaca, I., Gransberg, D. D., & Jeong, H. D. (2020). Improving the accuracy of early cost estimates on transportation infrastructure projects. *Journal of Management in Engineering*, 36(5), 04020063.

Kumar, A. (2020). Examination of cost overrun in highway projects using artificial neural networks in Kerala. *Int. J. Innov. Sci. Res. Technol,* 5, 1382-1392.

Mahalakshmi, G., & Rajasekaran, C. (2019). Early cost estimation of highway projects in India using artificial neural network. *In Sustainable construction and building materials* (pp. 659-672). Springer.

Saharkar, U. R., & Bachav, A. (2020). Life Cycle Cost Analysis of Road by Using ANN Method. *Universal Research Reports*, 7(8).

Sandhya, W. T., & Judit, L. P. (2020). Cost Evaluation of Construction Using Artificial Neural Network (ANN). *International Research Journal of Modernization in Engineering Technology and Science*, 2(4).

Tijanić, K., Car-Pušić, D., & Šperac, M. (2019). Cost estimation in road construction using artificial neural network. *Neural Computing and Applications*, 1-13.