Artificial Neural Network Analysis for Cost Estimation of Building Projects in India

* Ar. Ankita Gupta ¹, Ar. Piyali Debnath ²
Colliers International, Project Management Consultants, Mumbai, India. ¹
National Institute of Technology, Faculty of Architecture and Planning, Calicut, India ²
E-mail ¹: guptaankitain@gmail.com (Corresponding author), E-mail ²: piyalidebnath163@gmail.com

Abstract
In Construction Management, it is difficult to predict the cost estimate during the preliminary stage of the project because of limited information and unknown factors. Artificial Neural Networks can help in the prediction of estimate because of their simplicity and adaptability to non-linear problems. Due to their self-organizing nature they can be used to solve the problems even with low level programming. This makes them useful in interpreting and generalizing inadequate input information. ANN’s are crude derivatives of the biological neural network with single layered or multi-layered neuron in the form of input layer, hidden layer and output layer. The neural network first has to undergo training from historical data in order to make predictions or show results. The size of the data set, number of hidden neurons and the neural network architecture determines the success of the results. Selecting the right data set becomes imperative in this case. For the purpose of cost estimation, the cost drivers were taken as inputs and their estimated costs were taken as the target value. The cost drivers were selected carefully through literature review and survey to provide more accurate results for the estimate. The main drivers identified were: type of building, location, seismic zone, project complexity, ground condition, soil condition, plot area, plinth area, built-up area, number of stories, number of basement, principal structural material, type of foundation, level of design complexity, modular design, market conditions, construction conditions, risk factor, impact of risk, estimated duration of work, specification, quality of work and detailed cost estimate of project. A Pareto analysis performed on the significant drivers showed that the Duration of Work, Complexity of the Building, Plinth Area and Built-up Area, Height and Specifications were the most important cost drivers in a construction project. A problem was formulated based on these drivers with numerical and categorical data. The data set was trained with a neural network using the MATLAB software using feed forward backpropagation. Training was carried out till the greatest correlation and least Mean Squared Error was obtained after multiple iterations. This trained data was used to predict the cost for a new project. The output of the testing was 87% accurate despite the small data set used. 

Key words: Artificial Neural Network, Cost estimation, Construction projects, Building projects.

1. Introduction
In Construction Management, it is a difficult challenge to produce an accurate project cost estimate that takes into account the risks and uncertainties which may result in costs overruns. Cost estimation generally involves predicting labour, material, utilities or other costs over time given a small subset of factual data on “cost drivers.” The different types of costs involved in building construction projects are as follows:

**Category I.** Direct charges, which are identifiable as being associated with a particular project—standard labour, project labour, material, installation, salaries, and miscellaneous costs;

**Category II.** Allocated costs, which are charged to the project as the result of applying various formulas to the direct charges—standard overhead, other manufacturing costs, construction overhead, and general sales and administrative expenses.

In project management, cost estimation is not just a simple arithmetic based process. Estimation is an experience based task which includes assumptions, inclusions and exclusions. Cost estimation plays a crucial role in the initial part of a project as it enables decision making for investors, helps in the quantitative assessment of project and selection of alternatives. This cost estimate is used to set up a budget during the front end of the project. The same budget is used as a baseline to assess the performance of the project during the later stages and to avoid cost overruns. In the worst cases, cost overruns can cause the interruption of construction activities, and even the abandonment of projects, situations that occur when the owner does not have the financial capacity to meet the unexpected costs (Herrera, 2020).
Artificial Neural Network (ANN) can be used here in cost forecasting because it is an analogy based process that works well with limited data in the same way as the human brain. Our brain has been developed after thousands of years of evolution, and holds several advantages over the von Neumann or modern parallel computers. These include massive tolerance, learning and generalization ability, adaptivity and low energy consumption (Jain, 1996). Neural networks use a process which is analogous to the brain, wherein the existing data undergoes “training”. By using an adaptive learning rate for training, a trained neural network is developed for a better test accuracy. (Takase, 2018) The trained network becomes an “expert” in the category of information that it has been given to analyse. This “expert” can then be used to provide projections for new scenarios. Artificial neural networks are purely data driven models which transit from a random state to a final model through iterative training. Artificial neural networks are non-parametric statistical estimators which have proven to be universal approximators, and thus have potential for use in cost estimation modelling.

1.1 Aim
This research aims to examine the performance of cost estimation modelling through neural networks to predict the initial cost estimate for building construction projects in India.

1.2 Need
Conceptual cost estimates play a crucial role in initial project decisions, even when scope is not finalized and very limited design information is available during early project stages. Some categories of problems and challenges faced in the project environment may depend on so many subtle factors that a computer algorithm cannot be created to calculate the initial cost. There exists a need to predict reliable results from complicated or imprecise data for obtaining the initial cost estimate of a project. Artificial Neural Networks (ANN) are a family of statistical learning models inspired by the way biological nervous systems that process information. Unlike some conventional statistical systems, ANN adapt continuously to new data and possess an ability to learn from arbitrary and noisy data permits (Abraham, 2005). Hence they can lend themselves towards cost estimation in the early stage of the project. ANN method is appropriate for cost prediction due to the following reasons: (a) it is able to determine the interdependencies between data when considering significant variables in construction, (b) it can deal with non-linear relationships (c) it can handle incomplete data sets (Arafa, 2010, Amen,2022).

1.3 Objectives
The objectives of the paper are as follows:
• Analysis of the cost drivers based on historical data and identification of significant cost drivers.
• Formulating a problem based on the identified data set for cost estimation.
• Finding the cost estimate for a test project based on input from simulated and actual data sets using artificial neural network analysis tool for projects in residential, office, commercial and institutional category.
• Analysis of artificial neural network analysis tools in cost estimation in terms of accuracy, variability, model creation and model examination.

2. Methodology
Step 1- Identification and collection of data set for cost estimation. The data set is a key factor in cost estimation of the project. If the data set is not reliable, the results from the networks calculations will not be reliable. This dataset is collected through literature study and actual survey.

Step 2- Identifying the variables in the data set and classifying them into the following types.
   Categorical data. Dependent Category: Dependent or output variable whose possible values are taken from a set of possible categories; for example, yes or no, or red, green, or blue.
   - Dependent Numeric: Dependent or output variable whose possible values are numeric.
   - Independent Category: Independent variable whose possible values are taken from a set of possible categories; for example, yes or no, or red, green, or blue.
   Numeric data. Independent variable whose possible values are numeric.

Step 3- Setting the input variables for cost estimation of the project. These input variable will differ based on typology, complexity, location, materials, labor and other specifications of the project.

Step 4- Estimating the cost using neural network analysis technique for the sample set. Regression analysis is a statistical technique used to measure the extent to which a change in one quantity (variable) is accompanied by a change in some other quantity (variable).

Step 5- Comparing the results of ANN with the actual results based on accuracy, variability, model creation and model examination.
Step 6- Determining the relevance of the quantitative tools for the project environments.

3. Artificial Neural Network – An overview
- An Artificial Neural Network (ANN) provides a method to characterize synthetic neurons to solve complex problems in a manner which is similar to the human brain. (Huang, 2009).
- An artificial neural network is composed of many artificial neurons that are linked together according to a specific network architecture.
- The objective of the neural network is to transform the inputs into meaningful outputs.
- The ANNs have a high degree of self-organizing ability. In its crude imitation of the biological neural network, it enables low level programming to solve complex problems especially those which are non-linear, non-analytical, non-stationary and stochastic.
- The types of neural networks are classified based on learning systems which include a) Supervised, b) Unsupervised and c) Deep learning. A feedforward neural network is a simple type of artificial neural network with supervised learning where the units do not form a cycle between the connections of the units. It is the first and simplest type of ANN. (Mishra, 2018)

4. Identifying the Cost Drivers for A Construction Project
In order to write a neural network program for cost estimation in a project (with reasonable accuracy), the factors which have an impact on the overall cost have been listed.
The cost of building materials can themselves account for 60% of the cost of residential buildings (Olotuah, 2002). However, it is important to understand that there are several factors which may affect the cost of construction. They can range from accountable factors like quantity of materials to unaccountable factors like bribes resulting from corruption. It is necessary to narrow down the factors to streamline the neural network and get a logical output.
The following steps are used to sort through cost drivers and decide the input parameters:

4.1 Identification of Significant Factors Based on Literature Study
Table-1 demonstrates the list of factors elicited through literature study. (Stoy, 2007), (Ferry, 1999), (Toh, 2012), (Chan. S, 2005), (Ji S.H, 2019). Based on their frequency of occurrence, they were classified as significant and non-significant factors.

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Parameters</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Historical Database</td>
<td>Non-significant</td>
</tr>
<tr>
<td>2</td>
<td>Accuracy and experience of estimator</td>
<td>Significant</td>
</tr>
<tr>
<td>3</td>
<td>Knowledge of Client</td>
<td>Significant</td>
</tr>
<tr>
<td>4</td>
<td>Weather</td>
<td>Non-significant</td>
</tr>
<tr>
<td>5</td>
<td>Design Team Resources</td>
<td>Non-significant</td>
</tr>
<tr>
<td>6</td>
<td>Type of Estimation Method Used</td>
<td>Significant</td>
</tr>
<tr>
<td>7</td>
<td>Site Investigation/ Geotechnical Data</td>
<td>Significant</td>
</tr>
<tr>
<td>8</td>
<td>Structural Materials used</td>
<td>Significant</td>
</tr>
<tr>
<td>9</td>
<td>Cost of Construction Materials</td>
<td>Significant</td>
</tr>
<tr>
<td>10</td>
<td>Availability of Labour</td>
<td>Significant</td>
</tr>
<tr>
<td>11</td>
<td>Labour Wage Rate</td>
<td>Non-significant</td>
</tr>
<tr>
<td>12</td>
<td>Contractor’s / Sub Contractor’s Experience</td>
<td>Non-significant</td>
</tr>
<tr>
<td>13</td>
<td>Project Manager’s Experience</td>
<td>Non-significant</td>
</tr>
<tr>
<td>14</td>
<td>Size of project</td>
<td>Significant</td>
</tr>
<tr>
<td>15</td>
<td>Location of Project</td>
<td>Significant</td>
</tr>
<tr>
<td>16</td>
<td>Typology of Project</td>
<td>Significant</td>
</tr>
<tr>
<td>17</td>
<td>Regulatory Requirements</td>
<td>Non-significant</td>
</tr>
<tr>
<td>18</td>
<td>Political Influence</td>
<td>Non-significant</td>
</tr>
<tr>
<td>19</td>
<td>Environmental Conditions</td>
<td>Non-significant</td>
</tr>
<tr>
<td></td>
<td>Social and Religious Impact</td>
<td>Non-significant</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>21</td>
<td>Construction Site Conditions</td>
<td>Non-significant</td>
</tr>
<tr>
<td>22</td>
<td>Availability of services</td>
<td>Non-significant</td>
</tr>
<tr>
<td>23</td>
<td>Insurance</td>
<td>Non-significant</td>
</tr>
<tr>
<td>24</td>
<td>Number of Stories/ Height</td>
<td>Significant</td>
</tr>
<tr>
<td>25</td>
<td>Gross Floor Area, Plot Area, Built-up Area</td>
<td>Significant</td>
</tr>
<tr>
<td>26</td>
<td>Shape of the Building</td>
<td>Non-significant</td>
</tr>
<tr>
<td>27</td>
<td>Probability of Force Majeure (Earthquake, Flood or Cyclone, Fire)</td>
<td>Significant</td>
</tr>
<tr>
<td>28</td>
<td>Ground Conditions</td>
<td>Significant</td>
</tr>
<tr>
<td>29</td>
<td>Design Complexity</td>
<td>Significant</td>
</tr>
<tr>
<td>30</td>
<td>Market Conditions</td>
<td>Significant</td>
</tr>
<tr>
<td>31</td>
<td>Type of Foundation</td>
<td>Significant</td>
</tr>
<tr>
<td>32</td>
<td>Availability of Labour and Equipment Required</td>
<td>Significant</td>
</tr>
<tr>
<td>33</td>
<td>Labour Productivity</td>
<td>Non-significant</td>
</tr>
<tr>
<td>34</td>
<td>Quality of Work Required</td>
<td>Significant</td>
</tr>
<tr>
<td>35</td>
<td>Specification of Finishes</td>
<td>Significant</td>
</tr>
<tr>
<td>36</td>
<td>Total Duration of Project</td>
<td>Significant</td>
</tr>
<tr>
<td>37</td>
<td>Project Program</td>
<td>Non-significant</td>
</tr>
<tr>
<td>38</td>
<td>Changes in design/ Scope Change</td>
<td>Non-significant</td>
</tr>
<tr>
<td>39</td>
<td>Changes in Material</td>
<td>Non-significant</td>
</tr>
<tr>
<td>40</td>
<td>Change in Schedule</td>
<td>Non-significant</td>
</tr>
<tr>
<td>41</td>
<td>Risk Factor</td>
<td>Significant</td>
</tr>
<tr>
<td>42</td>
<td>Impact of Risk</td>
<td>Significant</td>
</tr>
<tr>
<td>43</td>
<td>Volume of HVAC work</td>
<td>Non-significant</td>
</tr>
<tr>
<td>44</td>
<td>Overheads</td>
<td>Significant</td>
</tr>
<tr>
<td>45</td>
<td>Mark-up and Profit</td>
<td>Non-significant</td>
</tr>
<tr>
<td>46</td>
<td>Amount of Rework</td>
<td>Non-significant</td>
</tr>
<tr>
<td>47</td>
<td>Inflation Rates</td>
<td>Non-significant</td>
</tr>
<tr>
<td>48</td>
<td>Delay in construction</td>
<td>Significant</td>
</tr>
</tbody>
</table>

4.2 Finding the Rank of The Cost Drivers Based on Relative Importance Index from Survey

A survey was carried out amongst construction industry professionals, project managers and architects to identify the critical cost drivers amongst the significant factors which were identified. Table-2 and Table-3 exhibits the weightages and ranking of the previously identified cost drivers consecutively, which are derived by analysing the surveyed data by using RII (Relative importance index) by using Eq. 1.

$$RII = \frac{\sum W}{A \cdot N}$$

Where, W—weighting given to each statement by the respondents and ranges from 1 to 5; A—Higher response integer (5); and N—total number of respondents.

Ranking is done based on relative importance index.
The 48 cost drivers which were initially identified were narrowed down to 18 critical factors based on the weightages assigned to the surveyors. This was carried out to create a compact list of parameters as input for the artificial neural network, as a higher number of parameters could lead to inaccuracy of test result.

Table 2: Identified factors from literature study.

<table>
<thead>
<tr>
<th>Type of Building</th>
<th>Location of Building</th>
<th>Sector</th>
<th>Ground Conditions</th>
<th>Soil Conditions</th>
<th>Pbl Area, Plinth Area, Built-up Area</th>
<th>Number of Stories (or Height)</th>
<th>Number of Basements</th>
<th>RII</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 2: Ranking of the significant cost drivers.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>RII</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sector</td>
<td>0.6066666667</td>
<td>18</td>
</tr>
<tr>
<td>Seismic Zone</td>
<td>0.6333333333</td>
<td>17</td>
</tr>
<tr>
<td>Construction Conditions</td>
<td>0.6533333333</td>
<td>16</td>
</tr>
<tr>
<td>Level of Design</td>
<td>0.66</td>
<td>15</td>
</tr>
<tr>
<td>Complexity (Modular or Non-Modular Design)</td>
<td>0.66</td>
<td>14</td>
</tr>
<tr>
<td>Risk Factor (High, Medium, Low)</td>
<td>0.66</td>
<td>13</td>
</tr>
<tr>
<td>Number of Basements</td>
<td>0.68</td>
<td>12</td>
</tr>
<tr>
<td>Soil Conditions</td>
<td>0.72</td>
<td>11</td>
</tr>
<tr>
<td>Ground Conditions</td>
<td>0.7266666667</td>
<td>10</td>
</tr>
<tr>
<td>Type of Foundation (Isolated, Raft or Pile)</td>
<td>0.7266666667</td>
<td>9</td>
</tr>
<tr>
<td>Market Conditions</td>
<td>0.7733333333</td>
<td>8</td>
</tr>
<tr>
<td>Principal Structural Members (Masonry, R.C.C or Steel)</td>
<td>0.78</td>
<td>7</td>
</tr>
<tr>
<td>Location of buildings</td>
<td>0.7866666667</td>
<td>6</td>
</tr>
<tr>
<td>Specifications (for finishes)</td>
<td>0.7933333333</td>
<td>5</td>
</tr>
<tr>
<td>Number of Stories (or Height)</td>
<td>0.7933333333</td>
<td>4</td>
</tr>
<tr>
<td>Plot Area, Plinth Area, Built-up Area</td>
<td>0.8</td>
<td>3</td>
</tr>
<tr>
<td>Type of Building</td>
<td>0.8066666667</td>
<td>2</td>
</tr>
<tr>
<td>Complexity</td>
<td>0.82</td>
<td>1</td>
</tr>
<tr>
<td>Duration of Work</td>
<td>0.8266666667</td>
<td>1</td>
</tr>
</tbody>
</table>
Fig. 1 depicts a Pareto analysis deployed to elicit the major issues to focus on to sort the cost estimate effectively. For example, targeting first 5 cost drivers would automatically take care of ~80% of the cost of construction.

From the rank chart based on survey, it is found that the following drivers have maximum impact on the cost of construction:

- Duration of Work
- Complexity of the Building (In terms of plan, elevation, interior spaces etc.)
- Plinth Area and Built-up Area
- Height
- Specifications (in terms of finishes, furniture etc.)

5. Problem Formulation for Initial Cost Estimation Using Artificial Neural Networks

A problem was formulated for cost estimation by dividing the critical cost drivers, identified after survey and ranking into numerical and categorical data with description (Refer to Table 4).

<table>
<thead>
<tr>
<th>Project characteristics</th>
<th>Type of information</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
</table>
| Type of building       | Categorical         | Number | 1-Residential  
|                        |                     |      | 2-Educational  
|                        |                     |      | 3-Institutional 
|                        |                     |      | 4-Assembly     
|                        |                     |      | 5-Mercantile   
|                        |                     |      | 6-Business and Industrial 
|                        |                     |      | 7-Storage      
|                        |                     |      | 8-Healthcare   |
| Location               | Categorical         | Number | 1-Central Business District  
<p>|                        |                     |      | 2-Metropolitan  |</p>
<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seismic Zone</td>
<td>Numerical</td>
<td>1-Regional</td>
</tr>
<tr>
<td>Sector</td>
<td>Categorical</td>
<td>1-Government 2- Public (Statutory and Departmental) 3-Private Sector 4-Public-Private</td>
</tr>
<tr>
<td>Complexity</td>
<td>Categorical</td>
<td>1-Low 2-Medium 3-High</td>
</tr>
<tr>
<td>Ground Condition</td>
<td>Categorical</td>
<td>1-Plane Land 2-Gentle Slope 3-Steep Slope</td>
</tr>
<tr>
<td>Soil Condition</td>
<td>Categorical</td>
<td>1-Soft Soil (Sand and Gravel) 2-Silt and Clay 3-Rock 4-Black Cotton Soil</td>
</tr>
<tr>
<td>Plot Area</td>
<td>Quantitative</td>
<td>sq.m.</td>
</tr>
<tr>
<td>Plinth Area</td>
<td>Quantitative</td>
<td>sq.m.</td>
</tr>
<tr>
<td>Built-Up Area</td>
<td>Quantitative</td>
<td>sq.m.</td>
</tr>
<tr>
<td>Number of Stories</td>
<td>Quantitative</td>
<td>Number</td>
</tr>
<tr>
<td>Number of Basement</td>
<td>Quantitative</td>
<td>Number</td>
</tr>
<tr>
<td>Principal Structural Material</td>
<td>Categorical</td>
<td>1-Masonry 2-R.C.C 3-Steel</td>
</tr>
<tr>
<td>Type of foundation</td>
<td>Categorical</td>
<td>1-Isolated 2-Pile Foundation 3-Raft Foundation</td>
</tr>
<tr>
<td>Level of Design Complexity</td>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>Modular design</td>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>Market Conditions</td>
<td>Categorical</td>
<td>1-Good 2-Fair 3-Bad</td>
</tr>
<tr>
<td>Construction Conditions</td>
<td>Categorical</td>
<td>1-Good 2-Fair 3-Bad</td>
</tr>
<tr>
<td>Risk factor based on relevance</td>
<td>Categorical</td>
<td>1-Low Relevance 2-Average Relevance 3-High relevance 4-Extremely High Relevance</td>
</tr>
<tr>
<td>Risk factor based on probability of unknown circumstances</td>
<td>Categorical</td>
<td>1-Low Risk 2-Medium Risk 3-High Risk</td>
</tr>
<tr>
<td>Estimated duration of work</td>
<td>Quantitative</td>
<td>Number (in months)</td>
</tr>
<tr>
<td>Specification</td>
<td></td>
<td>1-Poor 2-Average 3-Rich</td>
</tr>
<tr>
<td>Quality of Work</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detailed Cost Estimate of Project**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The Risk quantification model estimates the probable consequences in terms of cost, schedule and quality. It is measured by rating the level of relevance with respect to the consequences.*
6. Data Set and Training of Artificial Neural Network

A dataset was prepared based on the problem formulation sheet as input for the neural network (Figure-5) considering 10 construction projects which includes residential, commercial, institutional, healthcare and retail buildings.

<table>
<thead>
<tr>
<th>PARAMETERS</th>
<th>Project A</th>
<th>Project B</th>
<th>Project C</th>
<th>Project D</th>
<th>Project E</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYPE OF BUILDING</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>LOCATION</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SEISMIC ZONE</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>PROJECT COMPLEXITY</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GROUND CONDITION</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PLINTH AREA</td>
<td>9104.85</td>
<td>1150</td>
<td>1800</td>
<td>5040</td>
<td>5027</td>
</tr>
<tr>
<td>BUILT-UP AREA</td>
<td>108849</td>
<td>3467</td>
<td>4900</td>
<td>13100</td>
<td>10054</td>
</tr>
<tr>
<td>NUMBER OF STORIES</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>NUMBER OF BASEMENTS</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PRINCIPAL STRUCTURAL MATERIAL</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>TYPE OF FOUNDATION</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LEVEL OF DESIGN COMPLEXITY</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MARKET CONDITIONS</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>CONSTRUCTION CONDITIONS</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>RISK FACTOR</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>IMPACT OF RISK FACTOR</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ESTIMATED DURATION OF WORK</td>
<td>24</td>
<td>12</td>
<td>8</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>SPECIFICATION</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>QUALITY OF WORK</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>COST ESTIMATE</td>
<td>2363006505</td>
<td>407032054</td>
<td>563540744</td>
<td>936459256</td>
<td>69616000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PARAMETERS</th>
<th>Project F</th>
<th>Project G</th>
<th>Project H</th>
<th>Project I</th>
<th>Project J</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYPE OF BUILDING</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LOCATION</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>SEISMIC ZONE</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>PROJECT COMPLEXITY</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GROUND CONDITION</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>PLINTH AREA</td>
<td>166</td>
<td>2668</td>
<td>7343.92</td>
<td>110</td>
<td>2832</td>
</tr>
<tr>
<td>BUILT-UP AREA</td>
<td>720</td>
<td>8403</td>
<td>56953.46</td>
<td>271.55</td>
<td>34215</td>
</tr>
<tr>
<td>NUMBER OF STORIES</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>NUMBER OF BASEMENTS</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PRINCIPAL STRUCTURAL MATERIAL</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>TYPE OF FOUNDATION</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>LEVEL OF DESIGN COMPLEXITY</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MARKET CONDITIONS</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>CONSTRUCTION CONDITIONS</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>RISK FACTOR</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IMPACT OF RISK FACTOR</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ESTIMATED DURATION OF WORK</td>
<td>18</td>
<td>25</td>
<td>9</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>SPECIFICATION</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>QUALITY OF WORK</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>COST ESTIMATE</td>
<td>11495407</td>
<td>65402135.87</td>
<td>2796053034</td>
<td>9960000</td>
<td>696913416</td>
</tr>
</tbody>
</table>

Figure 2. Prepared data-set used as input in the Matlabs software
The trained neural network shows a mean squared error of 6.0364. MSE is one of the best ways to check the performance of the trained neural network. While performing the testing of the network, the network with the least MSE was selected as the trained network.

MSE can be defined mathematically as (highlighted in Eq. 2):

\[ MSE = \frac{1}{n} \sum_{i=1}^{n} (x_i - E(i))^2 \]  

- A well trained neural network should have a very low value of MSE.
- The mean squared error is defined as the average of squares of the ‘errors’. It is the difference between the attribute which is to be estimated and the estimator.
- The above figure indicates that the training has shown a positive outcome since the validation and the test curves are nearly similar (as seen in Fig. 3).
The coefficient of correlation or R values indicate whether the training data is a good or bad fit. It measures the direction and strength of the linear relationship between actual and predicted value.
An agreement can be seen between the actual and predicted lines, especially in the training and overall regression line, as indicated by the concentration of predicted values around the 45° line (Ongpeng, 2014).
An R value of 1 means a close relationship while 0 means a random relationship. A well trained neural network must have an R value of more than 0.8 (as seen in Fig. 4 the trained neural network shows an R value of 0.9805).

7. Prediction of Cost by The Trained ANN Model
Theoretically, testing the artificial neural network shall be same as training it. With the introduction of new data, the network needs to generalize the results from the test data and forecast the results. This evaluates the performance of the network after the training process.
For the purpose of testing the following new project dataset was taken:

Figure 4. Regression Results for Training, Test and Target Data.
### Table 1

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Test Project</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PARAMETERS</strong></td>
<td>Allims Jammu-Medical College</td>
</tr>
<tr>
<td>Type of Building</td>
<td>3</td>
</tr>
<tr>
<td>Location</td>
<td>2</td>
</tr>
<tr>
<td>Seismic Zone</td>
<td>4</td>
</tr>
<tr>
<td>Project Complexity</td>
<td>1</td>
</tr>
<tr>
<td>Ground Condition</td>
<td>2</td>
</tr>
<tr>
<td>Plinth Area</td>
<td>3856</td>
</tr>
<tr>
<td>Built-Up Area</td>
<td>17416</td>
</tr>
<tr>
<td>Number of Stories</td>
<td>6</td>
</tr>
<tr>
<td>Number of Basements</td>
<td>0</td>
</tr>
<tr>
<td>Principal Structural Material</td>
<td>2</td>
</tr>
<tr>
<td>Type of Foundation</td>
<td>2</td>
</tr>
<tr>
<td>Level of Design Complexity</td>
<td>2</td>
</tr>
<tr>
<td>Market Conditions</td>
<td>2</td>
</tr>
<tr>
<td>Construction Conditions</td>
<td>2</td>
</tr>
<tr>
<td>Risk Factor</td>
<td>1</td>
</tr>
<tr>
<td>Impact of Risk Factor</td>
<td>2</td>
</tr>
<tr>
<td>Estimated Duration of Work</td>
<td>18</td>
</tr>
<tr>
<td>Specification</td>
<td>1</td>
</tr>
<tr>
<td>Quality of Work</td>
<td>2</td>
</tr>
</tbody>
</table>

**Figure 5.** The test project dataset considered in the Matlabs software

For Fig. 5 (which shows the data of the test project for the trained ANN in Matlabs), the following results were obtained:

- Predicted value of Cost using the trained Artificial Neural Network: Rs. 691920130.2
- Actual Estimated Value of the project as per the BOQ: Rs. 614330223
- Percentage Difference in Estimated Value: +12.63%

Therefore on the basis of the input data set and training of the ANN, the results are 87% accurate.

[Note: Only one project has been used as a test case. This is because the data size is low. Increasing the data size would have improved the accuracy of the result.]

### 7.1 Remarks for the Artificial Neural Network model in Matlabs

A three layer function (Fig. 6) has been used for the designed artificial neural network: It consisted of the input layer, hidden layer and output layer.

**Figure 6.** Three layer function analysis.
For the same set of data, the network has been trained multiple times till we reach the least MSE (Mean Squares Error) and Maximum R value (Coefficient of Correlation). A feed forward back propagation type of the network has been designed as it is the simple and most commonly used to generate the best model (Kulkarni, 2017). The backpropagation algorithm gradually reduces the error between the model output and the target output by minimizing the mean square error (MSE) over a set of training set (Murat Günaydın, 2004). Each time we initiate a feed-forward network, the network parameters become different, due to assignment of different weights.

The MSE is a good overall measure of the success of the training process. If the network is still not trained properly then we can try increasing the number of hidden neurons. This is because the larger number of hidden neurons give the network more flexibility as there are more parameters to optimize. However, if the hidden layers are too large then the problem itself might become under-characterized.

Two different types of training algorithms were tried for this project: Bayesian and Lavenberg-Marquardt. The final neural network was trained using the Bayesian regularization method. The Bayesian function (trainbr) is used for difficult, small or noisy datasets. The algorithm requires more time to execute than the Lavenberg-Marquardt (trainlm) function but shows better results. For this ANN the Lavenberg-Marquardt function showed the best coefficient of correlation result as 0.8385 whereas the Bayesian function has shown a much better coefficient of correlation result- 0.9805 indicating a better fit.

The ANN showed a better result when the training data set was increased. Providing additional data helps in producing a network that can generalize well to new data and shows more accurate prediction values. Too much input data is also not good for the training of an artificial neural network as it can result in a common problem known as overfitting. In this case the error in the training set is driven to a small value because of the large data-set provided. But when a new data is added to the network, the error margin increases because the network has memorized the training samples but it has not learnt to generalize to new situations.

8. Conclusions
When different typologies of buildings across India were used for training, validation and testing of an Artificial Neural Network (taking a data-set with 15 samples or case studies) the following conclusions were drawn:

- The neural networks can successfully predict the early stage cost of construction for a building project.
- The accuracy of the predicted cost increases with the data size.
- With the increase in the sample size, the results showed a higher coefficient of correlation (R) and a lower Mean Squared Error (MSE).
- The neural network shows different training results and prediction outcomes if any of the parameters of the training data are altered or there is a change in weight from the original training. The training needs to be repeated till the best results are obtained.
- After repeated training iterations for this exercise, the training model with 10 hidden nodes gave the best results.
- For the training of construction cost data set in ANN, the Bayesian function (trainbr) shows better training results as compared to the Lavenberg-Marquardt (trainlm) function but the process takes more time to run (by a couple of seconds). The optimal neural network models involve a good combination of input data set, network architecture and training method.
- Apart from cost estimating in the decision making phase of the project, the ANN can also be used to find the lowest estimated bid of a supplier or the cost of insurance.

9. Limitations and Future Scope
The artificial neural network has a wide range of applications in the analogous field of cost estimation. It shows accurate results when the data sets, comprising of a large number of independent variables, have a non-linear structure or functional discontinuities.

Given the accurate results after successful training, the ANN can be considered as an alternative for parametric cost estimating especially during the earlier stages of the project. It is a great tool if we have a reliable historical data set. However the most tedious and expensive task in this process is to gather the right data for training and testing.

Artificial Neural Networks are yet to be used for live projects and a standard benchmark needs to be developed in order to determine the accuracy of the predicted results through ANN. They cannot be used where little or no information is available and they cannot be used for precise and arithmetic calculations (Mishra, 2018). Apart from the early stage cost estimation of projects, the ANN can also be used to find the lowest estimated bid of a supplier or the cost of insurance. (Egemen, 2007).
Acknowledgements
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interests
The authors declare no conflict of interest.

References