COST ESTIMATION OF HIGHWAY PROJECTS IN DEVELOPING COUNTRIES: ARTIFICIAL NEURAL NETWORK APPROACH

Jamshid SODIKOV
Graduate Student
Department of Civil and Environmental Eng.,
Saitama University
255 Shimo-Okubo, Sakura-ku, Saitama
338-8570, Japan
E-mail: s03D2053@post.saitama-u.ac.jp

Abstract: Cost estimation of highway projects with high accuracy at the conceptual phase of project development is crucial for planning and feasibility studies. However, a number of difficulties arise when conducting cost estimation during the conceptual phase. Major problems faced are lack of preliminary information, lack of database of road works costs, data missingness, lack of an appropriate cost estimation methods, and the involvement of uncertainties. Given its significance, conventional tools such as regression analysis have been widely employed to tackle the problem. However, recent statistical studies show that errors in cost estimation have not decreased. This paper focuses on the development of a more accurate estimation technique for highway projects in developing countries at the conceptual phase using artificial neural networks.

Key Words: Developing countries, Cost estimation, Artificial neural network.

1. INTRODUCTION

Cost estimation is an essential component of infrastructure projects. Accurate estimation will assist project managers to choose adequate alternatives and to avoid misjudging of technical and economic solutions. The accuracy of cost estimation increases toward the end of the project due to detailed and precise information. The conceptual phase is the first phase of a project in which the need is examined, alternatives are assessed, the goals and objectives of the project are established and a sponsor is identified (Wideman R. M, 1995).

Major difficulties which arise while conducting cost estimation during the conceptual phase are lack of preliminary information, lack of database of road works costs, and lack of up to date cost estimation methods. Additional difficulties arise due to larger uncertainties as result of engineering solutions, socio-economical, and environmental issues. Parametric cost estimation or estimation based on historic database during the conceptual estimate phase is widely used in developed countries. However, developing countries face difficulties related to the creation of a road work costs database, which may be used for cost estimation in either the conceptual stage or the feasibility study of a project cycle.

One of the earliest attempts to develop such a database was done by the World Bank Transport Unit which yielded in form of ROoad Costs Knowledge System (ROCKS) to be used in developing countries. It was designed to develop an international knowledge system on road work costs in order to establish an institutional memory, and obtain average and range unit costs based on historical data that could ultimately improve the reliability of new cost estimates and
reduce the risks generated by cost overruns ROCKS (2002). Having such database road agencies in developing countries can build road works cost benchmarks and help successfully evaluate new road projects. In order to utilize these data more efficiently and to solve these problems ongoing research investigates the possibilities of cost estimation in the conceptual stage by applying artificial neural networks (ANN).

Several researchers have addressed the problem associated with cost estimation in the earlier stage of project development. Pearce A. R. et al., (1996) developed a technique for generating range estimates to evaluate the risk of cost escalation in building construction using neural networks. They outlined a procedure for building a prototype model for cost estimation, and using ANN model to generate estimates for construction projects. Hegazy T. et al., (1998) used a neural network approach to manage construction cost data and develop a parametric cost estimating model for highway projects. They introduced two alternative techniques to train network’s weights: simplex optimization (Excel’s inherent solver function), and GAs (genetic algorithms). Adeli H. et al., (1998) indicated that highway construction costs are very noisy and this noise results from many unpredictable factors such as factors related to human judgment, random market fluctuations, and weather conditions. They were successful in introducing a regularization neural network model based on a solid mathematical foundation for estimating the cost of construction projects, making it more reliable and predictable. Recently, Gwang-Hee K. et al., (2004) examined different methods of cost estimation models in the early stage of building construction projects such as multiple regression analysis, neural networks and case – based reasoning. They concluded that neural networks performed the best prediction accuracy but case – based reasoning indicated better results in long run.

This paper is distinct from those above – mentioned research works by the integrity of an investigation, it discusses how to handle data missingness, analyzes the impact of a different set of variables on the project cost, and evaluates the proposed cost estimation model. It proposes a cost estimation technique for developing countries and criteria for using an artificial neural network.

A typical project circle moves from the concept development phase, to the design, advertisement, bid/award, and finally the construction phase. As project progresses the accuracy of the cost estimation increases because project details becomes more clearly. At an initial stage, estimate accuracy is between about ±25% and ±50% (Schexnayder, C. J. et al., 2003) due to less defined project details and other uncertainties due to both internal and external factors.

The recent work of researchers reveals that there are still many problems in cost estimation at the conceptual stage of project cycle. In the US, a cost variation study between actual and estimated costs by Flyvbjerg B.H. et al., (2002) of 258 transportation projects concluded that not only are costs underestimated in almost 9 out of 10 projects but also actual costs are, on average 28% higher than estimated costs. However the most interesting conclusion from this study is the fact that the same amount of underestimation exists today as existed 30 – 70 years ago. Another statistical study was undertaken by the World Bank, Transport Unit (ROCKS, 2002); in this study data from 65 developing countries were used to make comparisons between estimated costs at appraisal and actual cost at completion. Among these projects 62% were overestimated, and the rest underestimated. Accurate cost estimation at the early stages of project development is not
only a problem for developed countries but also developing countries. Therefore, there is need for better cost estimation techniques at the conceptual phase to be developed.

2. DATA DESCRIPTION AND ANALYSIS

The ROCKS database was used for this research. This database contains road works cost data from 65 developing countries. Among these countries Poland and Thailand have a relatively large number of projects. For this reason they were chosen in this study to investigate the relationship between project cost and other variables such as work activity, terrain type, road parameters etc.

The quality of the available database influences estimation accuracy regardless of the method used. Two methods were used to increase the quality of the ROCKS database. One was filling in missing data by using the expectation-maximization method and the other by generating synthetic data by using Road Construction Cost (RCC) models (Tsunokawa, 1983).

The Poland data set has 315 projects with the following predominant work activities: asphalt overlay, new construction (bituminous), new construction (concrete), reconstruction bituminous, and upgrading block to bituminous. Among them, new construction projects (38 projects) were selected for analysis with the following input variables: work duration, pavement width, shoulder width, ground rise and fall, average site clearing and grubbing (ACG), earthwork volume (EWV), surface class, and base material (Table 1). A unit cost was set as the output variable which has been represented in year 2000 US dollar. It is worth mentioning that as the ROCKS database does not contain ACG and EWV, the values of these variables were generated by using RCC models.

<table>
<thead>
<tr>
<th>Input Variables</th>
<th>Description</th>
<th>Unit</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWA</td>
<td>Predominant Work Activity</td>
<td>Category</td>
<td>New Construction Asphalt or Concrete</td>
</tr>
<tr>
<td>WD</td>
<td>Work Duration</td>
<td>month</td>
<td>14-30</td>
</tr>
<tr>
<td>PW</td>
<td>Pavement Width</td>
<td>m</td>
<td>7-14</td>
</tr>
<tr>
<td>SW</td>
<td>Shoulder Width</td>
<td>m</td>
<td>0-2</td>
</tr>
<tr>
<td>GRF</td>
<td>Ground Rise Fall</td>
<td>m/km</td>
<td>2-7</td>
</tr>
<tr>
<td>ACG</td>
<td>Average Site Clear/Grub</td>
<td>m²/km</td>
<td>12605-30297</td>
</tr>
<tr>
<td>EWV</td>
<td>Earthwork Volume</td>
<td>m³/km</td>
<td>13134-31941</td>
</tr>
<tr>
<td>SURFCLASS</td>
<td>Surface Class</td>
<td>Category</td>
<td>Asphalt or Concrete</td>
</tr>
<tr>
<td>BASEMATE</td>
<td>Base Material</td>
<td>Category</td>
<td>Crushed Stone or Cement Stab.</td>
</tr>
<tr>
<td>Output Variable</td>
<td>Unit Cost of New Construction Project</td>
<td>US Dollars</td>
<td>(2000)</td>
</tr>
</tbody>
</table>

The Thailand data set has 123 projects with the following predominant work activities: asphalt overlay, reconstruction, slurry seal, and widening (bituminous). Only Asphalt Overlay projects (42 projects) were chosen with the input variables work duration, pavement width, and shoulder width as shown in Table 2.
Table 2. Variables Description in Thailand Data Set (42 projects)

<table>
<thead>
<tr>
<th>Input Variables</th>
<th>Description</th>
<th>Unit</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWA</td>
<td>Predominant Work Activity</td>
<td>Category</td>
<td>Asphalt Overlay</td>
</tr>
<tr>
<td>WD</td>
<td>Work Duration</td>
<td>month</td>
<td>3-10</td>
</tr>
<tr>
<td>PW</td>
<td>Pavement Width</td>
<td>m</td>
<td>4-14</td>
</tr>
<tr>
<td>SW</td>
<td>Shoulder Width</td>
<td>m</td>
<td>0-8</td>
</tr>
</tbody>
</table>

The databases frequently suffer from some data items being missing. Missing data may reduce the precision of calculated statistics because of limited information (SPSS MVA, 1997). The reasons for missing data are: excessive response burden; survey perceived not to be important due to specific rationales; set up minimum and optional data for collection; timeliness and other. Data missing from the ROCKS database is mainly because of minimum versus optional data.

During data preparation two types of missing data were encountered: General missing (Poland dataset) and univariate missing (Thailand dataset). All data were clustered in such way that each group had similar properties. Such grouping was achieved by sorting all data by a predominant work activity and grouping into the similar work activity such as: asphalt overlay, reconstruction, new construction, upgrading, etc. In some cases, after clustering it was found that there were some work activities with few numbers of cases and that these cases were eliminated from the dataset. In the Poland dataset, two types of projects were left such as New Construction Asphalt Roads and New Construction Concrete Roads, in the Thailand dataset only Asphalt Overlay projects.

Comprehensive data analysis software SPSS was used for data imputation. SPSS Missing Value Analysis (MVA) includes several estimation methods as follows: Listwise; Pairwise; Expectation-Maximization (EM); Regression. The choice of estimation method depends on missing data pattern: missing completely at random (MCAR), missing at random (MAR), not missing at random (NMAR) and missing data type. The disadvantages and advantages of the above mentioned missing estimation methods is described in Joop Hox et al (1998) works. EM has the following advantage over other methods: it keeps data from being lost and each EM iteration always increases the likelihood. Due to this, missing data were imputed using the EM method.

The other challenge was to define which variables should be used during analysis in order to perform better cost prediction. One of the limitations was to consider the available variables in the ROCKS system. Hypothetically, three variables were selected for Asphalt Overlay including: work duration, pavement width, and shoulder width (excluding predominant work activity which is descriptive variable). For development and new construction cases, eight variables were selected to approximate these variables to project cost: work duration, pavement width, shoulder width, GRF, ACG, EWV, surface class and base material. Several composition variables were tested and those which had less influence on unit cost were eliminated. The range of variables can
be chosen according to (Figure 1) which was found as a result of trial and error, and also after considering ROCKS statistical reports on road cost details.

![Figure 1. Number of Variables vs. Project Type](image)

During data processing, it was found that new construction projects have a greater share of their cost takes up with site preparation and earthwork volume. For example, new construction projects in Poland dataset have up to 17% of cost attributed to site preparation, 21% to base and sub base and approximately 35% to surface build. Taking this into account, site preparation and earth work volume were introduced as key variables for new construction projects. In fact there are no such variables in the ROCKS data base. Thus, synthetic data were generated by using RCC models and obtained ACG and EWV values for new construction projects (the Poland dataset).

### 2.1 RCC Models.

New construction and reconstruction projects have the greatest share of construction quantities such as site preparation, earthwork and drainage. Prediction of these construction quantities can be done using the Road Construction Cost (RCC) model. RCC models were developed to predict physical quantities of an average site clearing/grubbing and earthwork volume. On the basis of a recommendation made by Tsunokawa K. (1987) following empirical relationship is used to predict the approximate area of site clearing and grubbing:

New construction:

\[
ACG = 1770 \exp (0.0278 \text{ GRF}) + 1610 \exp (-0.0144 \text{ GRF}) \text{ RW} \tag{1.1}
\]
Widening:

$$ACG=1610 \exp (-0.0144 \text{GRF}) \left[ \text{RW}_{\text{after}} - \text{RW}_{\text{before}} \right]$$  \hspace{1cm} (1.2)

Where:

ACG – the average area of site clearing and grubbing per unit length of road, in m$^2$/km;
GRF – the ground rise plus fall, in m/km;
RW – the roadway width (including carriageway and shoulders width), in meters;

The earthwork volume model takes into account geometric standards and terrain conditions, which highly influence the construction volume. Therefore, the concept of effective height was introduced in this model to make it sensitive to terrain type. The following relationship is used to predict the volume of earthworks:

New construction:

$$\text{EWV} = 1000 \left( \text{RW} + 1.039 \text{H} \right) \text{H}$$  \hspace{1cm} (1.3)

Widening:

$$\text{EWV} = 1000 \text{H} \left[ \text{RW}_{\text{after}} - \text{RW}_{\text{before}} \right]$$  \hspace{1cm} (1.4)

Where:

EWV – The volume of earthwork per unit length of road, in m$^3$/km
H – The effective height of earthwork, in meters, given by
$$\text{H} = 1.41 + 0.129 \text{G} + 0.0139 \text{GRF}$$
G – The rise plus fall differential, in m/km, given by
$$\text{G} = \text{GRF} - \text{RF}$$
RF – The road rise plus fall, in m/km

RF is usually intended to be smoother than GRF. Therefore, it was assumed in this study that, in general RF is approximately equal to 80% of GRF. ACG and EWV are developed in relation to terrain type and road width. In the ROCKS database terrain is presented in categorically, for example: flat, rolling, hilly and mountainous. In order to convert these categorical values to numerical values, the following range of GRF values were assumed in this study (Table 3)

<table>
<thead>
<tr>
<th>SN</th>
<th>Assumed range of GRF</th>
<th>Terrain Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 &lt; GRF &lt; 20 m/km</td>
<td>Flat</td>
</tr>
<tr>
<td>2</td>
<td>20 m/km ≤ GRF &lt; 50 m/km</td>
<td>Rolling</td>
</tr>
<tr>
<td>3</td>
<td>50 m/km ≤ GRF &lt; 80 m/km</td>
<td>Hilly</td>
</tr>
<tr>
<td>4</td>
<td>80 m/km ≤ GRF</td>
<td>Mountainous</td>
</tr>
</tbody>
</table>

According to these ranges random numbers were generated with a normal distribution rule for each terrain type. These values were used to calculate ACG and EWV variables using equations 1.1 and 1.3 respectively for new construction in the Poland data set.
3. Multiple regression model (MR)

Regression estimation models are well established and widely used in cost estimation. They are effective due to a well-defined mathematical, as well as being able to explain the significance of each variable and relationship between independent variables. Basically, regression models are intended to find the linear combination of independent variables which best correlates with independent variables. The regression equation is expressed as follows:

\[ Y = C + b_1X_1 + b_2X_3 + \ldots + b_kX_k \]  

(1.5)

Where:
C – regression constant; \( b_1, b_2, \ldots, b_k \) – regression estimates; \( X_1, \ldots, X_k \) – independent variables; \( Y \) – dependent variable.

The variables descriptions are mentioned in table 1 and table 2 for Poland and Thailand respectively. The goal was to estimate regression coefficients of equation (1.5). SPSS statistical tool was employed to perform regression analysis. The adjusted \( R^2 \) value is equal to -0.82 for both cases Poland and Thailand. To test significance of model ANOVA (F test) was performed.

4. ARTIFICIAL NEURAL NETWORK

Artificial neural networks have been applied to almost every application area where a data set is available and a good solution is sought. Artificial neural networks can cope with noisy data, missing data, imprecise or corrupted data, and still produce a good solution (Nikola K., 1998). Among supervised learning artificial neural networks multilayer perceptron (MLP) with backpropagation algorithm (further used as ANN) achieved popularity in numerous research areas and this type ANN was tested in this paper. One of the reasons for this was that it is a pioneer artificial neural network which showed promise to solve ill-defined, complicated, complex problems which are hard to describe in mathematical formula or expression.

The choice of ANN architecture depends on a number of factors such as the nature of the problem, data characteristics and complexity, the number of sample data, etc. ANN architecture was chosen after several trial and errors. Some recommendations from previous research works, for example, Hegazy T. et. al (1994) heuristically suggested that the number of hidden nodes should be set as one-half of the total input and output nodes. The other suggestions by Nikola K., (1998) involved on how to choose network parameters in a situation where the training set is clustered in groups with similar features. The number of these groups can be used to choose the number of hidden layers, the minimum number of hidden nodes should be \( h \geq (p-1) / (n+2) \), where \( p \) is the number of the training examples and \( n \) is the number of the inputs of the networks. In a situation where the training data are sparse and do not contain any common features, the number of connections might need to be close to the number of training examples in order for the network to reach a convergence. The greater the number of hidden nodes in a network, the more characteristics of the training data it will capture, but more will be time-consuming the learning procedure. Transfer function, learning rule, stop criteria and training characteristics were chosen after literature survey and also FAQ on ANNs maintained by Waren S. (2002).
Taking into account above – mentioned recommendations two ANN models were built, one with 2 hidden layers, and other with 1 hidden layer. In the Thailand dataset following architecture was used 4-10-6-1 which represents 4 nodes input layer, 10 nodes in the first hidden layer, 6 nodes in the second hidden layer, 1 node in output layer. In the case of the Poland data set, it is similar but with one hidden layer (Table 4).

### Table 4. Artificial Neural Network Architecture

<table>
<thead>
<tr>
<th>Description</th>
<th>Thailand Dataset 4-10-6-1</th>
<th>Poland Dataset 8-4-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagram</td>
<td><img src="image.png" alt="Diagram" /></td>
<td><img src="image.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Transfer Function</td>
<td>Tanh</td>
<td>Tanh</td>
</tr>
<tr>
<td>Learning Rule</td>
<td>Quickprob</td>
<td>Quickprob</td>
</tr>
<tr>
<td>Stop criteria</td>
<td>100 Epochs</td>
<td>100 Epochs</td>
</tr>
<tr>
<td>Training</td>
<td>Generic</td>
<td>Generic</td>
</tr>
</tbody>
</table>

### 5. RESULTS

NeuroSolution software was used for analysis. The data set in both cases (Poland and Thailand dataset) were divided into 3 groups: training, cross validation, test set with following percentage ratios:

1. Training set – 60%
2. Cross validation – 20%
3. Test set – 20%

The training set was used to train the network in order to choose its parameters (weights), the cross validation set was used for generalization, that is to produce better output for unseen examples. Finally, the test set was used to measure the performance of the selected ANN model. A practical issue is the underfitting/overfitting dilemma. Underfitting may occur when models are too simple and have insufficient training data. In such cases both training error and testing error are large. Overfitting occurs when model is too complex, have insufficient or noisy training data, resulting training error to be small but large testing error. In order to tackle these problems stop criteria and weight reset were used during the training network.
Evaluation of cost estimation accuracy is carried out by comparing actual cost and estimated cost, the common approaches being to calculate: MRE (The Magnitude of Relative Error) and MdMRE(The Median Magnitude of Relative Error), and MMRE (The Mean Magnitude of Relative Error) which can be expressed as follows:

\[
\text{MRE} = \frac{|A - E|}{A}
\]  


The mean of all MREs is the MMRE, which defined by formula (1.7):

\[
\text{MMRE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|A_i - E_i|}{A_i}
\]

A result presented in Table 5 is a comparison of two models including ANN and MR model. MMRE value for the Poland dataset is 24% and for Thailand dataset 26% which obtained by ANN model. In the case of MR model the error relatively higher than ANN model by 12% and 4% respectively. Both models ANN and MR indicated better accuracy level for cost estimation at the conceptual stage relative to existing conditions where the errors of cost estimation higher than 50% (statistical studies of Flyvbjerg B.H. et al., 2002 and World Bank, Transport Unit, ROCKS, 2002). However the ANN model was revealed as superior compared to MR model when the amount of data is limited and it is difficult to find relationship between variables.

<table>
<thead>
<tr>
<th>Table 5. Comparison of error between ANN and MR models (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMRE</td>
</tr>
<tr>
<td>ANN</td>
</tr>
<tr>
<td>Poland Dataset</td>
</tr>
<tr>
<td>Thailand Dataset</td>
</tr>
</tbody>
</table>

To produce these solutions numerous trials were performed and it was a time consuming part of this study. During these trials, error categorization was been set up for conceptual estimate. Schexnayder, C. J (2003) proposed the error of cost estimation at the conceptual phase is approximately between ±25 and ±50. In this study was proposed error categorization based on MMRE (table 6). According to this table one can choose prediction accuracy level and reduce time for training and choosing network parameters. High prediction accuracy requires more time to train the network and search for sophisticated ANN models.

<table>
<thead>
<tr>
<th>Table 6. Error Categorization (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMRE</td>
</tr>
<tr>
<td>Good</td>
</tr>
<tr>
<td>Fair</td>
</tr>
<tr>
<td>Poor</td>
</tr>
<tr>
<td>Less than 25</td>
</tr>
<tr>
<td>25 – 50</td>
</tr>
<tr>
<td>More than 50</td>
</tr>
</tbody>
</table>
The limitation made in this study was that only two types of road works were investigated: new construction and asphalt overlay projects. Some assumptions were accepted during data analysis such as the type of missing data and use of synthetic data. Nevertheless, this study can be reproduced and applied to other road works and with a complete dataset. Some recommendations were developed during the study such as how to choose the number of variables corresponding to project type and prediction accuracy levels.

6. SENSITIVITY ANALYSIS

Once the ANN model was built and tested, the next step is to evaluate the influence of each input variable to output variable. This serves as feedback, indicating which input channel has a significant effect. One might decide to remove input variables with less significance to reduce complexity of the ANN model and to save time on training. Sensitivity analysis is the method which discovers the cause and effect relationship between input and output variables. After numerous trials, it was found that pavement width, earthwork volume, work duration, average site cleaning and grubbing, surface class, and base material have a relatively high influence on an average new construction project’s cost (Figure 2). In the case of overlay projects work duration and pavement width had a high influence on project average cost (Figure 3).

Figure 2. Sensitivity about the Mean (Poland case)
7. CONCLUSION

Cost estimation done during the conceptual phase of the project cycle usually calculated approximately which leads to great inaccuracy. The reason for that is to have rough idea of a project in financial terms. On the other hand, project details are not fixed, well defined due to lack of information. In such cases ANN could be a suitable tool for estimating project cost. In this paper, an attempt was made to prove that cost estimation inaccuracy at the conceptual phase can be reduced to half of what it is at the present time. In conclusion, ANN could be an appropriate tool to help solve problems which come from a number of uncertainties such as cost estimation at the conceptual phase. Future work will be focused on developing an ANN model of cost estimation by incorporating other methods including fuzzy logic, case based reasoning, and other up to date techniques.

ACKNOWLEDGMENTS

I would like to thank Professor Koji Tsunokawa for supervising and providing all of the necessary conditions to conduct this research study. A special thanks to Dr. AliAkbar Golshani and Vivek Gupta for invaluable comments and recommendations.

REFERENCES


SPSS (1997), *Missing Value Analysis 7.5*, MaryAnn Hill / SPSS Inc.


ROCKS (2002), Road Costs Knowledge System (v. 2.01), (The World Bank, Transport Unit – TUDTR).
