*Fifth International Conference on Construction in the 21st Century (CITC-V)*
*"Collaboration and Integration in Engineering, Management and Technology"*
*May 20-22, 2009, Istanbul, Turkey*

# Cost Estimation of Trackworks of Lightrail and Metro Projects

Murat Gündüz, Erhan Öztürk
*Department of Civil Engineering, Middle East Technical University, Ankara, Turkey*
*gunduzm@metu.edu.tr, e124958@metu.edu.tr*

## Abstract

The main objective of this work is to develop models using multivariable regression and artificial neural network approaches for cost estimation of the construction costs of trackworks of light rail transit and metro projects at the early stages of the construction process in Turkey. These two approaches were applied to a data set of 16 projects by using seventeen parameters available at the early design phase. According to the results of each method, regression analysis estimated the cost of testing samples with an error of 2.32%. On the other hand, artificial neural network estimated the cost with 5.76% error, which is slightly higher than the regression error. As a result, two successful cost estimation models have been developed within the scope of this study. These models can be beneficial while taking the decision in the tender phase of projects that includes trackworks.

## Keywords

Artificial neural network, Early cost estimation, Metro, Light rail transit, Regression

## 1. Introduction

In today's world, due to the growing of population and its accumulation in the city centers, the public transportation comes out one of the most important issues, which will be handled by infrastructural investments to cities. When we consider the previous experiences and many other researches on ways of mass passenger transport in city centers, it is obvious that the most efficient solution to the public transportation is Light Rail Train (LRT) and Metro systems. These systems are used over centuries in developed countries of the world unlike developing countries. There is a considerable gap in terms of the availability of length of LRT or Metro line per citizen between developed and developing countries. That's why in order to compensate this gap and provide a modern service to their societies; recently the municipalities of developing countries start to make huge investments to these public transportation systems. At this point, an accurate early cost estimation of these systems becomes more critical for many parties including owners while taking investment decisions, because they have very limited budget. Besides, deviation from the pre-defined budget often brings a quick response from the public, the press, and sometimes even the state legislature. When this occurs, the municipality or state loses credibility over society and at the end the projects becomes less efficient than the design stage (Wilmot and Mei, 2005). On the other hand, if the owner can produce realistic budgets their image is enhanced and society gains. In addition to these; when we consider budget in terms of the contractors, the accuracy of estimation of construction costs in a construction project is a critical factor in the success of the project. The cost estimation models, which in the early stage estimate the construction costs with minimum project information, are useful in the preliminary design stage of a construction project. Improved cost estimation techniques, which are available to project managers, will facilitate more effective control of time and costs in construction projects (Hegazy, 2002).

Cost estimation is an area, which has received much attention from civil and cost engineers over the years. In an ideal situation all necessary cost information is present, allowing calculating the costs accurately. For the construction sector, this information comes from the geotechnical investigations, topographic measurements, structural design, and methods to be use. Collecting and combining of all of these components of detailed design stage, which are generated by separate specialized parties, takes considerable amount of time. However, sometimes reliable cost estimation is required within a very limited time period in order to decide the feasibility of the projects; it cannot be justified to generate detailed design drawings for every possible business development opportunity. Since these design stages are too time consuming, other fast yet accurate methods are required (Verlinden *et al.*, 2008). Therefore, parametric cost estimation methods, which are very useful in the early stage of a project's life cycle, has been introduced, when little information is known about the project's scope. These parametric cost estimation models include historical data that are currently used in practice as well as new data specific to a new project. One of the widely used parametric modeling types is regression, or multiple regression analysis. This is a very unique technique which can be used both analytical and predictive purposes by considering the affect of potential new items to the overall estimate reliability, although it is not appropriate when describing non-linear relationships, which are multidimensional, consisting of a multiple input and output problem (Tam and Fang, 1999). Another type is artificial neural network (ANN), is a computer system that simulates the learning process of the human brain. ANNs are widely applied in many industrial areas, including construction. The applicability of ANNs to construction has been extensively studied (Boussabaine, 1996). In addition, researchers have explored the application of ANNs to improve the accuracy of cost estimating beyond that of the regression model (Garza and Rouhana, 1995). In this study, these two techniques are used in order to evaluate realized project data.

The aim of this study is to establish and compare cost estimation models in order to assist cost prediction of trackworks of Light rail and Metro systems in Turkey regardless of the type of the infrastructure system of the project. In other word, the developed model for railway superstructure does not depend on feature or type of the section of the line such as TBM (tunnel boring machine) tunnel, depressed open/close or at grade line. For this reason, the data of completed LRT and Metro projects in which include trackworks in their scope were collected via site visits and related municipality and contractor interviews.

The study reported herein is based on realized data of actively working and under construction LRT and Metro Projects in Turkey. These data are gathered from various companies, which are responsible for construction of track works of above-mentioned projects. Trackworks data of 16 projects were analyzed by parametric cost estimation models, which are regression and neural networks.

In the following sections, the concept of multivariable regression and artificial neural network methodology are reviewed. Then, data collection and identification stages are demonstrated. After that, a stepwise procedure of multivariable regression and neural network analysis is represented. Finally, the summary of the study and the principal conclusions drawn from the comparison of the results of this study are provided.

## 2. Multivariable Regression and Artificial Neural Network Methodology

According to the recent citations from literature its usage area includes the following diverse applications: software development costs, roads in rural part of developing countries, query costs in data bases, urban water supply projects and design for manufacturability. Mason and Smith (1997) showed that professional cost estimators regularly use regression to build their cost models. Because of its strong mathematical background, regression analysis, being a cost estimation technique, has been used since the 1970's. However, Verlinden *et al.*, (2008) stated that although applied frequently, some drawbacks of regression techniques should be taken into account. Firstly, there is no general approach to help the cost engineer in choosing the model best fits the realized data for his specific problem. Secondly, when using regression techniques, the type of relationship between variables must be assumed as a priority. Thirdly, the number of input variables is limited to some extent. Regression models should be generated by considering above mentioned facts.

When evaluating the regression models also there are parameters needed to be checked. The main two important parameters are significance level (P value) and coefficient of determination ($R^2$ value). Because, $R^2$ value expresses the variability in the output that can be explained by the variables included in the model and P value shows the significance of the variables included in the model.

In this study, statistical software, Minitab, were used to develop the regression model. Multicollinearity of the variables is checked and the step-wise technique, based on the p values limitation, is followed by using this program.

As an alternative for regression techniques, artificial neural networks (ANNs) are currently used to generate cost estimations. Application of ANNs to enhance the accuracy of cost estimation by not being stuck within the limitations of regression has discovered by considerable number of researches. Verlinden *et al.*, (2008) observed that ANNs are applied in many fields such as financial services, biomedical applications, time-series prediction, text mining, decision-making and many others. Although, there are numerous applications of ANN, they all share an important common aspect: the processes to be predicted are correlated with a large number of explanatory variables and there may be high-level non-linear relationships between those variables. The most important aim of the ANNs is to find those nonlinear relationships to achieve better estimation.

Kim *et al.*, (2004) describes artificial neural network as a computer system that simulates the learning mechanism of the human brain. The main structure of ANNs is based on a number of neurons, which are grouped in one or several hidden layers. Neurons in these layers are connected to each other by a weighed function called transfer function. According to the contribution of each neuron to the final output, the output weight of neurons changes in every iteration process.

The design parameters determine to the performance of the ANNs considerably and will differ depending on the field of application. The number of hidden neurons and number of hidden layers have a great influence on detection capacity of ANNs for dependency between variables. However, there is no solid rule in determining of these parameters. This feature is one of the biggest drawbacks of ANNs shown in the literature. In addition, the parameters called learning rate and the momentum rate that affects the weight updating rule of ANNs are not also fixed values. All these parameters are decided by trial and error procedure, which takes considerable amount of time. However, in literature, several proposals that makes possible to limit the range of these parameters are present. So, training algorithm of ANNs can be chosen with a reasonable effort.

Hegazy *et al.*, (1994) proposed that one hidden layer is sufficient to generate an arbitrary mapping between inputs and outputs and the number of neurons in the hidden layer is 0,75m, m, or 2m + 1, where m is the number of input neurons. That’s why, ANN models, which contains three different numbers of hidden neurons, were performed in this study. He also proposed that the coefficients of the momentum and learning rate can be set to 0.9 and 0.7, respectively. In the light of this proposal, Kim *et al.*, (2004) were conducted ANN analysis by changing these parameters in a range which covers Hegazy’s proposal and got reasonable results. That’s why, in this study these coefficients were set between 0.5 and 0.9 (in steps of 0.1) to examine their effect and establish the best NN model. Numerous ANN models were evaluated by changing the number of neurons in the hidden layer according to previously proposed rule and by changing the coefficients of momentum and learning in steps of 0.1.

## 3. Data Collection and Identification

The first and one of the most important steps in collection of data was to decide on method. The second step in such a study is to decide on sample size, which affects the analysis validity directly. Therefore, the sample size is tried to be expanded as many as possible. A target project list was formed by conducting a small investigation for the existing and under construction projects, which have trackworks in their scope.

The data for this study were collected from 16 urban rail projects physically (in place) during one-year period. Several of these projects are LRT and the others are metros. Because of the reason that trackway construction is common both LRT and metro projects, these systems have been analyzed in the same manner in this study. Totally, 7 metro and 9 LRT project data are achieved to gather.

Variables that best describe the trackway cost is tried to select with a special attention. While selecting these variables, the experiences of the professionals working on this subject are taken into consideration. For the majority of the projects, which were selected, data of trackworks was collected successfully. In collection stage, it was very important to explain the scope of this study. Due to the reason that this study is dealing with neither the structural parts of the projects, nor the electrification and scada systems of the line. Therefore, the total cost data represents the required money to construct a trackway from the top of the sub-grade level to the top of the rail. It should be noted that all cost data is taken from the companies in the same currency, which is US Dollars.

## 4. Multivariable Regression Analysis Steps

The application of regression analysis was performed using Minitab, which is a statistical program with a spreadsheet-like data worksheet. It is capable of manipulating and transforming this data and can produce graphical and numerical summaries. Minitab also allows performing a wide variety of statistical computations. In this study, regression analysis is used to investigate and model the relationship between a response variable and one or more predictors. Minitab provides various least-squares and logistic regression procedures. Least squares procedures are used when response variables are continuous and logistic regression are used when response variables is categorical (Meyer and Krueger, 1998). Due to the fact that all variables are continuous in data set, the least square procedure is applied while evaluating the data.

At this point it is better to emphasize that in order to validate the prediction performance of regression analysis, the total data of 16 projects were divided into two groups, which are training set and the validation set. In the validation set, arbitrary chosen (by lottery) 2 project data is stored and these data were not used in the application of analysis. In other words, the training set analyzed by regression consists of remaining 14 projects data.

### 4.1 Correlation of Variables

The least square regression analysis is not applied if the total number of variables is greater than the number of observations, because residuals degree of freedom goes below zero. It can easily be seen that in data set of this study, the number of observations is equal to 16 and the number of the variables is 17. Moreover, when two observations are removed from the data set for the validation purpose, the gap is increased. That’s why, instead of removing variables based on the experience of which may have no effect on the cost, as it is proposed in the literature, correlation of independent variables have been investigated to find the linear relationship between each other, if exists. By using Minitab Pearson product moment, correlation coefficients between each pair of variables were calculated. Pearson product moment correlation coefficient measures the degree of linear relationship between two variables. The correlation coefficient assumes a value between -1 and +1. If one variable tends to increase as the other decreases, the correlation coefficient is negative. Conversely, if the two variables tend to increase together the correlation coefficient is positive. For two variables, the correlation coefficient r is calculated by using Minitab. As a result, several variables are highly correlated with each other, because of their high correlation values. That’s why; the variables LTT (Total Length of Main Trackway), NTW (Number of Thermite Welding) and MS (Maximum Superelevation) are eliminated.

### 4.2 Best Subset Procedure

“The best subsets regression procedure can be used to select a group of likely models for the analysis of variable selection. The general method is to the smallest subset that fulfills certain statistical criteria. The reason that one would use a subset of variables rather than a full set is because the subset model may actually estimate the regression coefficients and predict future responses with smaller variance than the full model using all predictors” (Gündüz, 2002). In the data analysis of this study, the best subset regression is decided to be used instead of using the full set of data for regression analysis to reduce the steps of regression and eliminate more variables, which do not contribute to the closeness of fitness of the final model. That’s why; best subset procedure is applied to data set in two parts. Consequently, by using correlation and best subset procedures 6 variables were eliminated (see Table 1). Least square regression analysis will be conducted with the remaining 11 variables and 14 observations.

**Table 1: List of Eliminated Variables**

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Total Length of Main Trackway (Meters) | Number of Simple Turnout | Sleeper Spacing (cm) | Number of Thermite Welding | Commercial Speed (Km/hr) | Maximum Superelevation (Cm) |
| LTT | NST | SS | NTW | CS | MS |

As it is stated previously, the first regression analysis is performed with 11 variables. The evaluation of these variables is done by stepwise manner and unnecessary parameters, which do not fit the model well, have been dropped off the model by considering their p values. This procedure is called parsimonious modeling. Pankratz (1983) states that the principle of parsimony is important, because parsimonious models generally produce better forecasts in general. The same elimination procedure was followed in this study. The P value of the each eliminated variable and the coefficient of determination ($R^2$) of each model from R.1 to R.5 is given in the Table 2.

**Table 2: List of P Values of Eliminated Variables**

| Regression models | Number of variables in the model | Eliminated variable | P value of eliminated variable | R2 values of the models |
|---|---|---|---|---|
| R.1 | 11 | MVC | 0.959 | 0.994 |
| R.2 | 10 | NC | 0.205 | 0.994 |
| R.3 | 9 | MOS | 0.208 | 0.988 |
| R.4 | 8 | MHC | 0.294 | 0.982 |
| R.5 | 7 | WC | 0.324 | 0.977 |
| R.6 | 6 | MSL | 0.161 | 0.972 |

The regression model R.7, was performed by using the remaining 5 variables. Because of the reason that P values of variables included in model R.7 are below or too close to 0.1, it is selected as final model, which has a $R^2$ value of 0.963.

Minitab results of Final Regression Model (R.7) as follows:

The regression equation is Cost = 529190 + 704 LBT + 707 LDF - 3860 WR + 293 NS + 325 HPC

S = 3087067   R-Sq = 96.3%   R-Sq (adj) = 93.9%

It is good to remember that these 2 project (observation 10 and observation 14) data were not used while generating the model. For testing the model, previously separated data of 2 projects were used. In accordance with the previously defined statistically significant variables and final regression equation, the predicted cost for these two projects were calculated (see Table3) by entering the values of the variables present in the equation.

**Table 3: Prediction Performance of Final Model**

| Project Number | Predicted Values (USD) | Real Project Values (USD) | Percent Error | MAPE |
|---|---|---|---|---|
| 10 | 27,799,471 | 26,640,000 | -4.17% | -2.32 |
| 14 | 21,379,299 | 21,280,000 | -0.46% | |

## 5. Neural Network Analysis Steps

The application of ANN analysis was performed using Neural Power, which is a general, integrated, easiest-to-use and powerful ANN program. It can be used in almost all study fields such as multi-nonlinear regression, forecasting, curve fit, pattern recognition, decision making and problem optimization, time series analysis and market predictions. The parameters of ANN, which were defined in previous sections were reorganized and changed after each trial to find the best architecture thorough the Neural Power. The RMSE is a quadratic scoring rule, which measures the average magnitude of the error and shows the difference between forecast and corresponding observed values, each squared and then averaged over the sample. Then, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful tool when large errors are particularly undesirable (Eumetcal, http://www.eumetcal.org.uk.). In this study, RMSE value of 0.01 was used as stopping criteria of the iteration.

The number of hidden layer neurons has been decided according to the Hegazy's proposal mentioned above. In this study, three sets of ANN models (S.1, S.2, S.3) with one hidden layer were performed for the analysis and numbers of hidden layer neurons were decided in accordance with the 0,75m, m, or 2m + 1, coefficients. In each set, the learning rate and the momentum parameters of ANN, were set between 0.5 and 0.9 (in steps of 0.1) to examine their effect and establish the best NN model.

The best architecture of each ANN group (see Table 4) was selected by examining prediction performance of them. The prediction error and MAPE of the best of each group for the projects number 10 and number 14 (testing projects) can be seen in Table 5. It should be remembered that the prediction results are scaled with (1/1000). According to the prediction performance represented by the MAPE, the model S3.C produced reasonable predictions within an average absolute error of 5.761%. Thus, the model S3.C was selected as best architecture (see Figure 1) for the data set of this study and analysis was finalized.

**Table 4: Network Architecture of Best ANN of Each Group**

| | Network Architecture | | |
|---|---|---|---|
| **Network Characteristics** | **S1.A** | **S2.B** | **S3.C** |
| Network architecture | 17-33-1 | 17-17-1 | 17-13-1 |
| Learning algorithm | BP | BP | BP |
| Learning rate | 0.6 | 0.5 | 0.5 |
| Momentum rate | 0.6 | 0.7 | 0.5 |
| Stopping criteria | 0.01 | 0.01 | 0.01 |
| Number of iteration | 2517 | 3245 | 2983 |

**Table 5: Prediction Performance of Each Model**

| | Prediction Error | | |
|---|---|---|---|
| **Project No** | **S1.A** | **S2.B** | **S3.C** |
| 10 | 9.136% | 8.845% | 5.656% |
| 14 | 12.220% | 12.152% | 5.867% |
| MAPE | 10.678 | 10.498 | 5.761 |

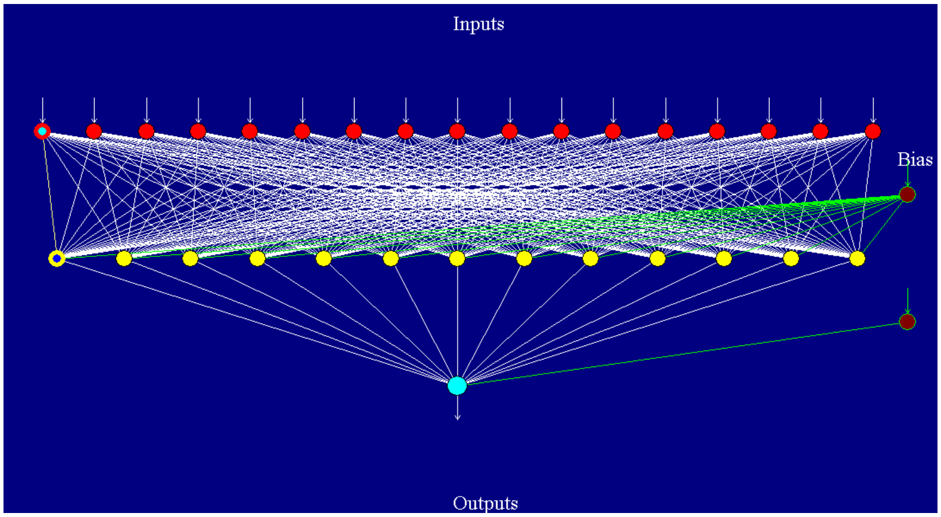

**Figure 1: ANN Architecture of S3.C**

## 6. Conclusion

The main objective of this work was to develop models using multivariable regression and artificial neural network approaches for cost estimation of the construction costs of trackworks of Turkish light rail transit and metro projects at the early stages of the construction process. These two approaches used a data set of 16 projects. The approach was shown to be capable of providing accurate estimates of trackworks cost by using seventeen parameters available at the early design phase.

According to the results of each method, regression analysis was estimated the cost of testing samples with an error of MAPE of 2.32%. On the other hand, artificial neural network was estimated the cost with 5.761% error, which is slightly higher than the regression error. As a result, two successful models have been developed within the scope of this study. These models can be beneficial while taking the decision in the tender phase of projects that includes trackworks.

The MAPE results have showed us, the regression has fit to the data set well. In addition to this, the prediction performance of the ANN is highly satisfactory also. According to many studies present in literature, the estimation performances of ANNs are usually presented as superior to regression analysis. That's why; the results of analysis of this research may be case specific due to the number of LRT and metro projects available in Turkey. In order to ensure the performances of these to model in this specific subject of area, further studies should be done with an expanded data set in the future. Because of the reason that neural networks train themselves by using observations and the performance of a neural network model of cost estimation inevitably depends on the quality and the quantity of data. As the number of observations increases the estimation error of ANNs decreases. Therefore, it is possible to develop a solid estimation model with ANN for trackway projects in Turkey with trustworthy, high quality, full-scale cost data of various projects. However; establishing the best ANN model needed considerable amount of time, because of the trial and error procedure, while defining ANN parameters to find architecture for best estimation. Therefore; to be more effective in finding the parameters of the ANN than the trial and error method, other applications such as back propagation network model incorporating a genetic algorithm (GA) may be used in future research to estimate the trackway cost in the early project stage.

## 7. References

Boussabaine, A.H. (1996). "The use of artificial neural networks in construction management: A review". *Construction Management and Economics*, Vol. 14, No. 5, pp. 427–436.

Eumetcal. http://www.eumetcal.org.uk, last access 15 June 2008.

Garza, J., and Rouhana, K. (1995). "Neural network versus parameter-based application". *Cost Engineering*, Vol. 37, No. 2, pp. 14–18.

Gündüz, M. (2002). "Change order impact assessment for labor intensive construction". PhD Thesis, University of Wisconsin Madison.

Hegazy, T. (2002). *Computer-Based Construction Project Management*. Upper Saddle River, NJ: Prentice-Hall Inc.

Hegazy, T., Fazio, P., and Moselhi, O. (1994). "Developing practical neural network applications using back-propagation". *Microcomputer in Civil Engineering*, Vol. 9, pp. 145–159.

Kim, G.H, Sung-Hoon, A., and Kyung-In, K. (2004). "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning". *Building and Environment* , Vol. 39, pp. 1235 – 1242.

Mason, A.K and Smith, A.E (1997). "Cost estimation predictive modeling. Regression versus neural networks". *Engineering Economist*, Vol. 42, No. 2, pp. 137-161.

Meyer, R., and Krueger, D. (1998). *A Minitab Guide to Statistics*, Prentice Hall Inc.

Pankratz, A. (1983). *Forecasting with Univariate Box-Jenkins Models*, John Wiley and Sons, New York.

Tam, C.M, and Fang, C.F. (1999). "Comparative cost analysis of using high performance concrete in tall building construction by artificial neural networks". ACI Structural Journal, Vol. 96, No. 6, pp. 927–936.

Duflou, B. J.R. Collin, P. and Cattrysse. D. (2008). "Cost estimation for sheet metal parts using multiple regression and artificial neural networks: A case study". *International Journal of Production Economics*, Vol. 111, pp. 484 - 492.

Wilmot, C., and Mei, B. (2005). "Neural network modeling of highway construction costs". *ASCE, Journal of Construction Engineering and Management*, Vol. 131, No. 7.