Expert Systems with Applications 38 (2011) 2873–2877

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Parametric cost estimation system for light rail transit and metro trackworks

Murat Gunduz [a,*], Latif Onur Ugur [b], Erhan Ozturk [c]

[a] Department of Civil Engineering, Middle East Technical University, Ankara 06531, Turkey
[b] Kaman Junior Technical College, Ahi Evran University, Kirsehir, Turkey
[c] Metis Construction and Trade Co., Ankara, Turkey

ARTICLE INFO

Keywords:
Artificial neural network
Early cost estimation
Metro
Light rail transit
Regression

ABSTRACT

The main objective of this work is to develop early cost estimation models for light rail transit and metro trackworks using the multivariable regression and artificial neural network approaches. These two approaches were applied to a data set of 16 projects by using 17 parameters available at the early design phase. The 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 was slightly higher than the regression error. As a result, two successful cost estimation models have been developed depending on the findings of this paper. These models can effectively be utilized in the tender decision-making phase of projects with trackworks.

1. Introduction

In today's world, due to overgrowth of population and its accumulation in city centers, public transportation has become one of the most important infrastructural investments. The most efficient solutions to public transportation are light rail train (LRT) and metro systems. These systems have been used in developed countries for centuries. However, there is still a considerable gap in terms of the availability of the length of LRT or metro line per citizen between developed and developing countries. Municipalities of developing countries have lately started to make huge investments in aforementioned public transportation systems to compensate this gap as well as to provide their societies with modern services. At this point, an accurate early cost estimation of these systems while taking investment decisions becomes more critical for many parties including owners. Besides, deviation from the pre-defined budget often brings a quick response from the public, press, and sometimes even the state legislature. When this occurs, municipality or state loses credibility over society, as a result of which projects become less efficient than the design stage (Chester, Asce, & Bing, 2005). The accuracy of estimation of construction costs in a construction project is a critical factor for determining the success of a project. The cost estimation models, which in the early stage estimate construction costs with minimum project information, are useful in the preliminary design stage of a construction project. Furthermore, improved cost estimation techniques, which are available to project managers, will facilitate more effective control of time and costs in construction projects (Hegazy, 2002).

This study differs from others in the literature by its introduction into early cost estimation of trackworks. The sample data employed for the cost prediction comes from an intensive survey administered on the contractors and municipalities in Turkey.

2. Parametric cost estimation

Reliable cost estimation is required within a very limited time period in the feasibility stage of the projects when it is not possible to generate detailed design drawings. Since these design stages are too time consuming, other fast and accurate methods are required (Verlinden, Duflou, Collin, & Cattrysse, 2008).

Existing parametric cost estimation methods were introduced, when little was 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 multiple regression analysis. This is a unique technique which can be used for both analytical and predictive purposes by taking into consideration the effect of potential new items to estimate the overall reliability. However, this technique is not appropriate for the purpose of describing non-linear relationships, which are multidimensional as they consist of a multiple input and output problem (Tam & Fang, 1999). Another parametric cost estimation method is artificial neural network (ANN), which 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 by

* Corresponding author. Tel.: +90 312 210 5422; fax: +90 312 210 5401.
E-mail address: gunduzm@metu.edu.tr (M. Gunduz).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.08.080

Boussabaine (1996). In addition, researchers have explored the application of ANNs to improve the accuracy of cost estimation beyond that of the regression model (Garza & Rouhana, 1995). In this paper, these two techniques are used in order to evaluate the realized project data. This paper will perform both of the previously mentioned methods for the cost estimation procedure of light rail transit and metro trackworks.

3. Scope

The aim of the present 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 used in the project. In other words, the model developed for railway superstructure does not depend on the feature or type of the section of the line such as TBM (tunnel boring machine) tunnel, depressed open/close or grade line. For this reason, the data were obtained by means of site visits and interviews with municipalities and contractors from completed LRT and metro projects which included trackworks in their scope.

The paper is based on data from all actively working LRT and metro projects as well as those that are still under construction in Turkey. Furthermore, the data were gathered from various companies, which were responsible for the construction of trackworks of the above-mentioned projects. All in all, the trackworks data of 16 projects were analyzed by means of parametric cost estimation models, namely regression and neural networks.

4. Literature review

Both regression techniques and ANNs are frequently used in cost estimation. Many studies can also be found in the literature comparing these methods. McKim (1993) presented the use of ANNs in cost estimating. The estimates obtained from his study were compared with the estimates produced by three other methods known as pump scaling factor estimates, exponent scaling using the 0.6 rule, exponent scaling using the best exponent, and the best equation. McKim (1993) observed that ANNs have great potential for estimating non-deterministic costing systems. Smith and Mason (1999) showed the high performance, stability and ease of use of cost estimation modeling for ANNS and regression techniques. The authors concluded that if little knowledge about the relationships between dependent and independent variables is present, ANNs outperform the more classical regression techniques. However, if the relationship between different variables can be identified, the regression model will have advantages in the evaluation of the model. Zhang and Fuh (1998) developed an artificial neural network model for early cost estimation of packaging products. As an outcome, the authors revealed the cost affecting parameters of a product design. The correlation between these parameters and the final cost of the product was discovered by using a back-propagation artificial neural network algorithm depending on historical data. Garza and Rouhana (1995) examined appropriate areas for the use of ANNs for cost estimation purposes. In their study; the cost estimation of carbon steel pipes was done by using the parametric model for the purpose of comparing ANNs and their performances. Their study revealed that ANNs have considerable estimation capabilities. However, an ANN has a number of disadvantages too, such as a complex neural network architecture design and parameter setting, both of which require trial and error.

Adeli and Wu (1998) formulated a regularization neural network to estimate very noisy highway construction costs. The authors observed that as the number of attributes increased, the construction cost could be estimated with more accuracy. In another study, a neural network model for parametric cost estimation of highway projects was proposed by using a spreadsheet simulation. Hegazy and Ayed (1998) used ANNs in order to analyze the data from 18 construction projects with ten variables to forecast the final construction cost. The authors tried to optimize the ANN prediction performance by using back-propagation training, simplex optimization and genetic algorithms. Depending on their findings, it was concluded that back-propagation training were the most applicable to their data set. Kim, Sung-Hoon, and Kyung-In (2004) stated that adequate estimation of construction cost is a key factor in construction projects. The authors examined the performance of three cost estimation models. The examinations were based on multiple regression analysis (MRA), artificial neural networks (ANNs), and case-based reasoning (CBR) of 530 historical cost data. ANN estimating model gave more accurate results than both the MRA estimating model and CBR. Gunaydin and Dogan (2004) used the neural network methodology to estimate costs in early phases of building design processes. To this end, cost and design data from thirty projects were used for training and testing. Neural network methodology with eight design parameters was utilized in estimating the square meter cost of reinforced concrete structural systems of 4–8 storey residential buildings in Turkey; and an average cost estimation accuracy of 93% was achieved. Sonmez (2004) established a conceptual cost estimation model for building projects by using the data for thirty continuing care retirement projects built by a contractor in the United States. He showed the benefit of using both regression and ANNs to reveal the relationship among such variables as the total building area, combined percentage area of health center, number of floors, and percent area of structured parking. He constructed the parsimonious models, which could be defined as generating models to be used for the purpose of getting more satisfactory predictions by avoiding the unnecessary variables. In order to eliminate the non-contributing variables, a step wise regression process was applied by considering the p-values of each variable. After establishing the first regression model, the variables that had the highest p-value were eliminated one by one, as a result of which, the final regression model was developed with a reasonable R^2 value (closeness of fit). In addition to the regression model, ANNs were established to compare the prediction performance of these two models. Ugur (2007) studied the costs of multiple reinforced concrete residential buildings by using ANNs. The network developed by the author has a multi layer and back-propagation structure. Building elevations, unit numbers in each flat, normal flat area, heights of flats, total number of flats, empty areas in the outer surface, total areas of the outer surfaces and average areas of the units in normal flats were the variables of this study in which design parameters for minimum costs were determined by means of the ANN structure.

5. Data collection and identification

One important step in data collection was to decide on methodology and sample size. A target project list was formed by conducting a small-scaled investigation into the completed and under-construction projects, which had trackworks in their scope. The data for this study were collected from 16 urban rail projects (7 Metro and 9 LRT) physically (in place) within a period of one year. These 16 projects covered all the completed and ongoing urban rail projects in Turkey by the time of data collection. Due to the fact that trackway construction is common in both LRT and metro projects, these systems have been analyzed in the same manner.

Variables that best describe the trackway cost have been selected with special attention. The list of variables and their definitions can be seen in Table 1. While selecting these variables, the

Table 1
Variables and their definitions.

Variable #	Variable abbreviation	Variable description	Variable #	Variable abbreviation	Variable description
x1	LTT	Total length of main trackway (m)	x10	NS	Number of sleepers and concrete support blocks
x2	LBT	Length of ballasted trackway (m)	x11	HPC	Hourly passenger capacity (passenger/h/direction)
x3	LDF	Length of direct fixation trackway (m)	x12	MOS	Maximum operation speed (km/h)
x4	NC	Number of crossover	x13	CS	Commercial speed (km/h)
x5	NST	Number of simple turnout	x14	MSL	Maximum slope of the line (%)
x6	SS	Sleeper spacing (cm)	x15	MS	Maximum superelevation (cm)
x7	WC	Workmanship cost (USD)	x16	MHC	Minimum horizontal curvature (m)
x8	WR	Total weight of rail (tons)	x17	MVC	Minimum vertical curvature (m)
x9	NTW	Number of thermite welding	y	C	Cost (USD)

experiences of the professionals working on this subject were taken into consideration. The project holders were promised confidentiality in the phase of data collection. The term 'total cost data' in this study represents the required cost (in USD) to construct a trackway from the top of the subgrade level to the top of the rail.

6. Multivariable regression analysis steps

The application of regression analysis was performed by using Minitab, a statistical software with a spreadsheet-like data worksheet. Regression analysis was performed to investigate and model the relationship between the response variable and the predictors. Due to the fact that all variables were continuous in the data set, the least square procedure was applied while analyzing the data.

In order to validate the prediction performance of the regression analysis, two randomly selected projects were separated from the database.

7. Correlation of variables

It is common knowledge that 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 becomes negative. In the database of this study, the number of observations and the variables were 16 and 17, respectively. Moreover, when two observations were removed from the data set for the purpose of validation, the gap increased. The correlations of the independent variables were also investigated to eliminate the highly correlated variables.

The Pearson product moment correlation coefficient measures the degree of linear relationship between two variables. Thus, by using the Pearson product moment, the correlation coefficients between each pair of variables were calculated. Furthermore, the correlation coefficient r was calculated by the statistical package. The correlation coefficient assumes a value between -1 and $+1$. The variable pairs with high correlation values, found as a result of the Pearson correlation procedure, can be seen in Table 2. One of the highly correlated variables (LTT, NTW, MS) was excluded to overcome the multicollinearity problem.

8. 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 use the smallest subset that fulfills certain statistical criteria. The reason for choosing a subset of variables rather than a full set is that the subset model may actually estimate the regression coefficients and predict future responses with smaller variance than the full model using all predictors (Gunduz, 2002). In the data analysis stage of this study, the best subset regression was decided to be used instead of using the full set of data in order to eliminate the variables which poorly defined the dependent variable.

Table 2
List of variable pairs with high correlation values.

Variable pairs	Pearson correlation value
LTT – WR (x1−x8)	0.922
LTT – NTW (x1−x9)	0.915
WR – NTW (x8−x9)	0.927
MS – MOS (x15−x12)	0.902

The dataset was run with the best subset analysis and it was observed that the variables NST, SS and CS added very little to the procedure of defining the dependent variable. Therefore these variables were not used in the analysis stage of the study.

With the help of correlation and best subset analysis, the number of variables was reduced to eleven. The evaluation of these eleven variables was done in stepwise manner. Thus, the unnecessary parameters, which did not fit well into the model, were automatically dropped off the model according to their p-values. This procedure is called parsimonious modeling. Pankratz (1983) points out that the principle of parsimony is important, because parsimonious models generally produce better forecasts. In parsimonious models, a backward elimination method is used for the initial RM. According to this technique, variables that do not contribute to the model are eliminated one by one at each step. The regression statistic, significance level (p-value) is used for determining the variables to be eliminated. In general, the variables corresponding to the coefficients with p-values close to or less than 0.10 are considered to have significant contribution to the model (Ontepeli, 2005). The same elimination procedure was followed in this paper. The p-value of the each eliminated variable and the coefficient of determination (R^2) of each model from R.1 to R.6 is given in the Table 3.

In model R.1, the variable MVC had the highest p-value, which equaled to 0.959, which was very high when previously mentioned criteria of the p-values were considered. Therefore, the MVC variable did not probably have a major contribution to the model and thus it was removed from the model. So, the regression model R.2 was performed by the remaining 10 parameters.

In model R.2, the variable NC had the highest p-value (0.205) and was removed from the model. In the models R.3, R.4, R.5 and R.6, the same procedure was followed. As a result, the variables MOS, MHC, WC and MSL were removed from the analysis.

Table 3
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

Table 4
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%	

The regression model R.7, was performed by using the remaining 5 variables. Because the *p*-values of the variables included in model R.7 were below or close to 0.1, it was selected as the final model with an R^2 value of 0.963 (Eq. (1)).

$$\text{Cost} = 529190 + 704\text{LBT} + 707\text{LDF} - 3860\text{WR} + 293\text{NS} + 325\text{HPC} \quad (1)$$

For the purpose of validating the model, previously separated data of two projects were used. The predicted cost for these two projects was calculated in accordance with the previously defined statistically significant variables and final regression equation, as shown in Table 4.

9. Neural network analysis

The application of ANN analysis was performed by means of Neural Power, which is integrated, easiest-to-use and powerful ANN software. This software 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 in order to find the best architecture through the ANN software.

The RMSE (Root Mean Squared Error) is a quadratic scoring rule which measures the average magnitude of an 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 errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is the most useful tool when large errors are particularly undesirable (Eumetcal, 2009). In this paper, RMSE value of 0.01 was used as stopping criteria of the iteration.

Hegazy, Fazio, and Moselhi (1994) suggests that one hidden layer is sufficient to generate an arbitrary mapping between inputs and outputs; and that the number of neurons in the hidden layer is 0, 75 *m*, *m*, or $2m + 1$, where *m* is the number of input neurons. That's why, ANN models, which contain three different numbers of hidden neurons, were performed in this paper. He further points out that the coefficients of the momentum and learning rate can be set to 0.9 and 0.7, respectively. In the light of Hegazy et al.'s proposals, Kim et al. (2004) conducted an ANN analysis by changing these parameters in a range to cover Hegazy's proposal, and got significant results. That's why, in this paper these coefficients were set between 0.5 and 0.9 (in steps of 0.1) to examine their effect and to establish the best ANN model. Numerous ANN models were evaluated by changing the number of neurons in the hidden layer according to the previously proposed rule and by changing the coefficients of momentum and learning in steps of 0.1 (see Table 5).

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.

In the first set (S.1), which had a configuration of 17-33-1, 25 ANN models were developed by changing the learning rate and momentum parameters between 0.5 and 0.9 (in steps of 0.1). Among them, the best structure of ANN (S1.A) was determined to be 17-33-1 (0.6–0.6), which means that there were 17, 33, and 1 neurons in the input, hidden, and output layers, respectively, and the learning rate and the momentum coefficient of the back-propagation algorithm were both 0.6.

Table 5
Number of hidden neurons in each set of model.

Number of input neurons (m)	Number of hidden layer neurons			Number of output neurons
	S.1 2 m + 1	S.2 m	S.3 0.7 m	
17	33	17	13	1

In the second set (S.2), which had a configuration of 17-17-1, 25 ANN models were performed by changing the learning rate and momentum parameters between 0.5 and 0.9 (in steps of 0.1). Among them, the best structure of ANN (S2.B) was determined to be 17-17-1 (0.5–0.7). In other words, there were 17, 17, and 1 neurons in the input, hidden, and output layers, respectively, and the learning rate and the momentum coefficient of the back-propagation algorithm were 0.5 and 0.7, respectively.

In the final set (S.3), which had a configuration of 17-13-1, the other 25 ANN models were implemented by changing the learning rate and momentum parameters between 0.5 and 0.9 (in steps of 0.1). Among them, the best structure of ANN (S3.C) was determined to be 17-13-1 (0.5–0.5). That means there were 17, 13, and 1 neurons in the input, hidden, and output layers, respectively, and both of the learning rate and the momentum coefficient of the back-propagation algorithm were 0.5.

The best architecture of each ANN group (Table 6) was selected by examining their prediction performance. The predicted cost values of the best of each group for the validation projects can be seen in Table 7. In addition, the prediction error and MAPE (Mean Absolute Percentage Error) of S1.A S2.B and S3.C are presented in Table 8. 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

Table 6
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 7
Predicted cost values.

Project No.	Actual cost (USD * 1000)	Predicted cost (USD * 1000)		
		S1.A	S2.B	S3.C
10	26,640	29,318	29,225	28,237
14	21,280	24,242	24,224	20,101

Table 8
Prediction performance of each model.

Project no.	Prediction error		
	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

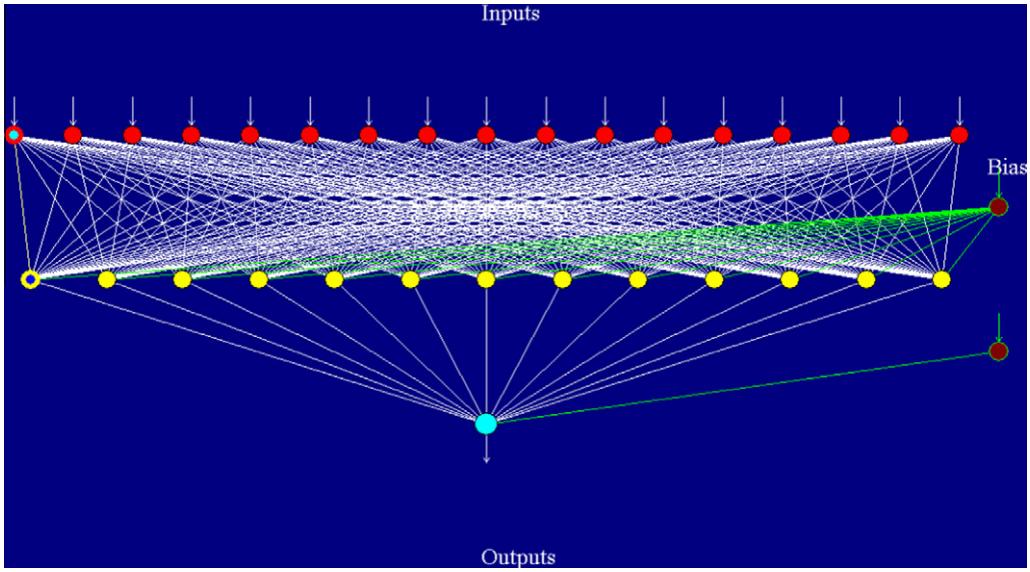

Fig. 1. ANN architecture of S3.C.

average absolute error of 5.761%. Thus, the model S3.C was selected as the best architecture (Fig. 1) for the data set of this paper and the analysis was finalized.

10. Conclusions

The main objective of this work was to develop multivariable regression and artificial neural network models for cost estimation of the construction costs of trackworks of light rail transit and metro projects at the early stages of the construction process. These two approaches used a data set of 16 projects and were shown to be capable of providing accurate estimates for costs of trackworks by using 17 parameters available at the early design phase.

According to the results of each method, regression analysis estimated the cost of the validation projects with an error of 2.32%. On the other hand, artificial neural network estimated the cost with an error of 5.761%, which is slightly higher than the regression error. Depending on the findings, two successful models have been developed within the scope of this paper. These models can be beneficial in the tender decision-making phase of projects that includes trackworks.

According to many studies present in literature, the estimation performances of ANNs are usually presented as superior to regression analysis. For the purpose of this study, all trackworks projects in Turkey were used. It is common knowledge that as the number of observations increases the estimation error of ANNs decreases. Therefore, more data from similar studies to be carried out in the near future will hopefully increase the efficiency of the ANN model as well as the regression model.

References

Adeli, H., & Wu, M. (1998). Regularization neural network for construction cost estimation. *Journal of Construction Engineering Management, 124*(1), 18–24.

Boussabaine, A. H. (1996). The use of artificial neural networks in construction management: A review. *Construction Management and Economics, 14*(5), 427–436.

Chester, G. W., Asce, M., & Bing, M. (2005). Neural Network Modeling of Highway Construction Costs. *ASCE, Journal of Construction Engineering and Management, 131*(7), 765–771.

Eumetcal (2009). http://www.eumetcal.org.uk. Last accessed 24.07.09.

Garza, J., & Rouhana, K. (1995). Neural network versus parameter-based application. *Cost Engineering, 37*(2), 14–18.

Gunaydin, H. M., & Dogan, Z. S. (2004). A neural network approach for early cost estimation of structural system of building. *International Journal Project Management, 22*, 595–602.

Gunduz, M. (2002). *Change order impact assessment for labor intensive construction*. PhD thesis, University of Wisconsin Madison.

Hegazy, T., & Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management, 124*(3), 210–218.

Hegazy, T. (2002). *Computer-based construction project management*. Upper Saddle River, NJ: Prentice-Hall Inc..

Hegazy, T., Fazio, P., & Moselhi, O. (1994). Developing practical neural network applications using back-propagation. *Microcomputer in Civil Engineering, 9*(2), 145–159.

Kim, G. H., Sung-Hoon, A., & Kyung-In, K. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment, 39*(10), 1235–1242.

McKim, R. (1993). Neural network application to cost engineering. *Cost Engineering, 35*(7), 31–35.

Ontepeli, M. B. (2005). *Conceptual cost estimating of urban railway system projects*. MS thesis presented to Middle East Technical University (METU), Ankara, Turkey.

Pankratz, A. (1983). *Forecasting with univariate Box-Jenkins models*. New York: John Wiley and Sons.

Smith, A. E., & Mason, A. K. (1999). Cost estimation predictive modeling: Regression versus neural networks. *Engineering Economist, 42*(2), 137–161.

Sonmez, R. (2004). Conceptual cost estimation of building projects with regression analysis and neural networks. *Canadian Journal of Civil Engineering, 31*(4), 677–683.

Tam, C. M., & Fang, C. F. (1999). Comparative cost analysis of using high performance concrete in tall building construction by artificial neural networks. *ACI Structural Journal, 96*(6), 927–936.

Ugur, L. O. (2007). *Analysis of construction of costs with artificial neural networks*. Thesis presented to Gazi University, Ankara, Turkey in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Verlinden, B., Duflou, J. R., Collin, P., & Cattrysse, D. (2008). Cost estimation for sheet metal parts using multiple regression and artificial neural networks: A case study. *International Journal of Production Economics, 111*(2), 484–492.

Zhang, Y. F., & Fuh, J. Y. H. (1998). A neural network approach for early cost estimation of packaging products. *Computers & Industrial Engineering, 34*(2–4), 433–450.