Expert Systems with Applications 38 (2011) 9913–9917

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Range estimation of construction costs using neural networks with bootstrap prediction intervals

Rifat Sonmez *

Department of Civil Engineering, Middle East Technical University, Ankara 06531, Turkey

ARTICLE INFO

Keywords:
Neural networks
Cost estimation
Bayesian regularization
Bootstrap method
Construction projects

ABSTRACT

Modeling of construction costs is a challenging task, as it requires representation of complex relations between factors and project costs with sparse and noisy data. In this paper, neural networks with bootstrap prediction intervals are presented for range estimation of construction costs. In the integrated approach, neural networks are used for modeling the mapping function between the factors and costs, and bootstrap method is used to quantify the level of variability included in the estimated costs. The integrated method is applied to range estimation of building projects. Two techniques; elimination of the input variables, and Bayesian regularization were implemented to improve generalization capabilities of the neural network models. The proposed modeling approach enables identification of parsimonious mapping function between the factors and cost and, provides a tool to quantify the prediction variability of the neural network models. Hence, the integrated approach presents a robust and pragmatic alternative for conceptual estimation of costs.

1. Introduction

In early stages of construction projects, detailed design drawings are not usually available and conceptual estimation of costs is required for making budgeting and feasibility decisions. Cost models provide a powerful alternative for conceptual estimation of construction costs. However, development of cost models can be challenging as there are several factors impacting costs, and there is usually sparse and noisy data available for modeling.

Regression models have been used commonly to quantify the impact of factors on project costs (Kaiser, 2006; Karshenas, 1984; Kouskoulas & Koehn, 1974; Lowe, Emsley, & Harding, 2006; Sonmez, 2008). Regression analysis requires the user to decide a priori on the class of relations (linear, quadratic, etc.) to be used in modeling. Determination of the class of relations between the factors and project costs may become complicated, especially when multiple cost components are considered as the dependent variables.

Neural networks (Adeli & Wu, 1998; Cheng, Tsai, & Sudjono, 2009; Duran, Rodriguez, & Consalter, 2009; Hegazy & Ayed, 1998; Kim, Seo, & Kang, 2005; Sonmez, 2004) and case-based reasoning (Chou, 2009; Dogan, Arditi, & Gunaydin, 2006; Wang, Chiou, & Juan, 2008) models have been proposed in recent years for modeling of costs as an alternative to regression analysis. Neural network and case-based reasoning cost models usually provide a point estimate for estimating costs. However, a single point prediction does not include any information regarding the level of variability included in the estimated costs. Inclusion of estimation variability is very crucial for management decisions as conceptual cost estimates usually include a high amount of uncertainty.

The level of uncertainties included in the cost estimates can be quantified by developing range estimates using simulation techniques (Touran & Wiser, 1992; Wang, 2002). However the impacts of parameters on project costs are not generally included in simulation techniques. Parametric range estimation can be performed by using prediction intervals in regression models but, in this method a priori decision on the class of relations is required (Sonmez, 2004, 2008). Within this content, the main purpose of this study is to develop a method for range estimation of costs, which can identify the impact of parameters on costs easily, and can also quantify the level of uncertainties included in the estimated costs. The remainder of the paper is organized as follows: Section 2 is devoted to description of the project data and formulation of the cost modeling problem. In Section 3, neural network models are described. Bootstrap prediction intervals are presented in Section 4. Finally, concluding remarks are made in Section 5.

2. Modeling of building costs

Construction cost models in general reflect experiences that are unique to a construction organization for a certain project type. In this study cost models are developed for continuous care retirement community (CCRC) projects. CCRCs are living units for

* Tel.: +90 3122102422.
E-mail address: rsonmez@metu.edu.tr

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.042

Table 1
Factors impacting cost.

No	Description
$X1$	Total gross residential, commons, nursing facilities, and structured parking area in m^2
$X2$	Construction cost index
$X3$	City cost index
$X4$	Number of stories
$X5$	Percent area of commons and nursing facilities in the total building area
$X6$	Percent structured parking area in total area
$X7$	Total gross building area per residential unit
$X8$	Site area in m^2
$X9$	Major demolition on site
$X10$	Site waste treatment
$X11$	Wood frame
$X12$	Steel frame
$X13$	Concrete frame
$X14$	Steel and concrete frame
$X15$	Masonry structure
$X16$	Wood exterior finish
$X17$	Vinyl exterior finish
$X18$	Masonry exterior finish
$X19$	Plaster exterior finish
$X20$	Number of elevator stops
$X21$	Project duration in months

seniors, and offer them access to coordinated social activities, dining and health care services. Models are developed using data of 20 CCRC projects compiled from a building contractor. The projects were built over a 13 year time frame, at 10 different locations in the United States. The data included information of 21 factors which are presented in Table 1. Factors used in this study are the variables related to building, site, and project conditions which might impact the project costs. The factors $X9, X10, \ldots, X19$ are binary variables, and are used to represent presence of a certain condition such as; major demolition on the site ($X9$). The cost components are the system costs, and are defined according to a cost breakdown structure. The contractor used 11 cost components to organize the CCRC costs, as shown in Table 2.

The task of cost modeling for CCRC projects is determination of the relations between the factors ($X1, X2, \ldots, X21$) and cost components ($Y1, Y2, \ldots, Y11$). Quantitative relations between the factors and cost components can be represented by a single overall model, or by an individual model for each cost component. In regression modeling the factors ($X1, X2, \ldots, X21$) are the independent variables and the cost components ($Y1, Y2, \ldots, Y11$) are the dependent variables. One of the main difficulties of cost modeling by regression analysis is determination of a proper model representing the relations between the factors and cost components adequately. Linear regression models without any interaction terms can be used to simplify the modeling process. However, linear models do not always guarantee adequate representation of the relations. An alternative approach, as implemented in this study, is to use neural networks to establish a mapping function between the factors and cost components.

3. Neural network models

Feed forward neural networks are used to develop an adequate cost model for the CCRC projects. The input buffer of the first neural network model consisted of 21 units, representing all of the factors ($X1, X2, \ldots, X21$) and, the output layer consisted of 11 units representing the cost components ($Y1, Y2, \ldots, Y11$), as shown in Fig. 1. Three neural networks with different number of hidden units were trained to determine the number of hidden units for the first neural network model. The neural networks had one hidden layer including 32 (Model-1a), 16 (Model-1b), and 8 (Model-1c) hidden units. Back propagation algorithm with an adaptive learning rate was used for training. In adaptive learning rate, the learning step size is kept as large as possible while maintaining a stable learning, by making the learning rate responsive to the complexity of the local error surface (Demuth & Beale, 2001). Leave-one-out cross validation was performed to evaluate the adequacy of the neural network models. One project data was not used during training, and the trained network was used to predict the total cost of that project. The procedure was repeated for all the projects, and predicted costs were compared with the actual estimated costs to assess the prediction performance. Mean absolute percent error (MAPE) was used as an error measure to evaluate the prediction performance. MAPE value for a cost model was the average of deviations between predicted total project cost and actual estimated total project cost in absolute values; expressed as proportion of the actual estimated cost. MAPE values for Model-1a, Model-1b and Model-1c are 32.4, 27.7 and 33.3 respectively as shown in Table 3. 16 hidden units are used for Model-1 based on the results.

Table 2
Cost components.

No	Description
$Y1$	Site development
$Y2$	Foundations and slab on grade
$Y3$	Structure
$Y4$	Enclosure
$Y5$	Interior finishes
$Y6$	Equipment and special construction
$Y7$	Conveying systems
$Y8$	Mechanical
$Y9$	Fire protection
$Y10$	Electrical
$Y11$	General requirements

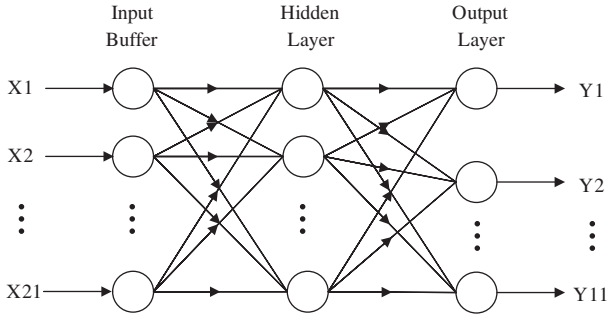

Fig. 1. Model-1.

Table 3
Prediction performance of Models 1a, 1b, and 1c.

Model	Nh*	MAPE
Model-1a	32	32.4
Model-1b	16	27.7
Model-1c	8	33.3

* Nh: Number of units in the hidden layer.

3.1. Elimination of factors

The second cost model (Model-2) consisted of eleven neural networks ($N1, N2, \ldots, N11$) with each having one unit in the output layer representing the cost components ($Y1, Y2, \ldots, Y11$) respectively. For each neural network model the factors which may have a potential impact on the cost component was included

Table 4
Factors included in the input buffers of neural networks for Model-2.

Factors	Neural networks										
	*N*1	*N*2	*N*3	*N*4	*N*5	*N*6	*N*7	*N*8	*N*9	*N*10	N11
*X*1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
*X*2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
*X*3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
*X*4		✓	✓	✓							
*X*5					✓	✓		✓	✓	✓	
*X*6					✓			✓			
*X*7					✓	✓		✓	✓	✓	
*X*8	✓										
*X*9	✓										
*X*10	✓										
*X*11		✓	✓								
*X*12			✓								
*X*13			✓								
*X*14			✓								
*X*15			✓								
*X*16				✓							
*X*17				✓							
*X*18				✓							
*X*19				✓							
*X*20							✓				
*X*21											✓
Nh*	7	6	10	9	7	6	5	7	6	6	5

* Nh: Number of units in the hidden layer.

in the input buffer. As an example, the factors identifying the exterior finish type ($X16, X17, \ldots, X19$) were included in the input buffer of $N4$; the neural network for enclosure costs. However, these factors were not included in the $N1$, since the exterior finish type is not expected to have an impact on the site development costs. The unimportant input variables are eliminated to achieve parsimonious models for improving the generalization capabilities of the neural network models (Sato et al., 2005; Shastri, Rabelo, Onjeyekwe, & Vila, 1998). The factors that were included in the input buffers of the neural networks ($N1, N2, \ldots, N11$) for Model-2 are given in Table 4. The important factors for each cost component were determined by a senior cost estimator based on experience. The neural networks ($N1, N2, \ldots, N11$) of Model-2 were also trained using back propagation algorithm with an adaptive learning rate.

3.2. Bayesian regularization

An ideal neural network model for cost estimation is the one that generalizes well and is able to predict cost of new projects accurately. To achieve a neural network that generalizes well, the size of the network parameters can be constrained by regularization technique. The idea behind regularization is that the true underlying function is assumed to have a degree of smoothness, and when network parameters are kept small, the network response will be smooth (Gencay & Qi, 2001). Thus, the neural network should be able to represent the true function, rather than capturing noise. Regularization neural networks were recommended for construction cost estimation as the noise in the data can be taken into account with this technique (Adeli & Wu, 1998).

In a typical feed forward neural network, the objective function used for training is the mean sum of squares of the network errors:

$$\mathrm{Ft} = \mathrm{mse} \tag{1}$$

With regularization the objective function becomes:

$$\mathrm{Fr} = \gamma \mathrm{mse} + (1-\gamma)\mathrm{msw} \tag{2}$$

where; γ is the performance ratio, and msw is the mean sum of squares of the network weights and biases. One of the main difficulties of regularization is to determine the optimum value for the performance ratio. If the performance ratio is too small, the neural network may not adequately fit the training data, on the other hand if the ratio is too large, the network may over fit to the data. Determination of regularization parameters can be automated. One approach to this process is to use Bayesian techniques (MacKay, 1992). In Bayesian framework, the weights and biases of the network are assumed to be random variables with specified distributions. The regularization parameters can be estimated with statistical techniques. A combination of Bayesian regularization and Levenberg–Marquart training (Foresee & Hagan, 1997) has been implemented to develop the third model (Model-3). Model-3 is very similar to the Model-2 however, in this model Bayesian regularization is used to improve generalization of the neural networks ($N1, N2, \ldots, N11$).

3.3. Model results

The MAPE prediction performance values of Model-1, Model-2 and Model-3 are presented in Table 5. Leave-one-out cross validation was performed to determine MAPE for each model. MAPE values for Model-1, Model-2, and Model-3 were 27.7, 13.5, and 11.7 respectively. Paired t-tests were performed to test the significance of the difference between Model-1 and Model-2, and between Model-2 and Model-3. The levels of significance of the t-tests (P values) are given in Table 6. A P value of 0.002 for the first test indicates that, there was significant evidence to reject the hypothesis that the MAPE for Model-1 was equal to MAPE for Model-2 at the $\alpha = 0.1$ significance level. On the other hand, a P value of 0.181 for the second test indicates that, there was not significant evidence to reject the hypothesis that the MAPE for Model-2 was equal to MAPE for Model-3 at the $\alpha = 0.1$ significance level. The difference between MAPE values of Model-1 and Model-2 was statistically significant, however the difference between MAPE values of Model-2 and Model-3 was not statistically significant.

Elimination of the factors that do not have a potential impact on the cost components improved prediction performance of the neural network models. Model-2, which only included the important factors had a better prediction performance than the Model-1, which included all of the factors. When all of the factors are included in the input buffer for each cost component, neural network model does not generalize adequately. The results also indicate that Model-2 is sufficiently parsimonious since implementation of regularization did not significantly improve the prediction performance of the model. The accuracy levels of Model-2 and Model-3 are both acceptable as they are within the suggested range of −15 to +25 for early building cost models (AbouRizk, Babey, & Karumanasseri, 2002).

4. Bootstrap prediction intervals

Bootstrap is a procedure that involves random re-sampling of the existing data with replacement. In sampling with replacement every sample is returned to the data set after sampling. So a particular data point from the observed data set could appear zero times, once, twice, or, more in a given bootstrap sample. The size of each bootstrap sample is chosen to be equal to the size of the original

Table 5
Prediction performance of models 1, 2, and 3.

Model	MAPE
Model-1	27.7
Model-2	13.5
Model-3	11.7

Table 6
Paired *t*-test results for MAPE.

Test	*P* Value
Model-1 vs. Model-2	0.002
Model-2 vs. Model-3	0.181

Table 7
Range estimates for the case project.

No	Description	Probability level		
		15%	50%	85%
*Y*1	Site development	398,210	952,200	2,082,400
*Y*2	Foundations and slab on grade	238,040	879,520	1,449,300
*Y*3	Structure	759,760	2,757,600	3,542,600
*Y*4	Enclosure	539,460	1,579,000	3,134,000
*Y*5	Interior finishes	2,821,900	4,644,700	5,431,700
*Y*6	Equipment and special construction	194,400	411,570	532,950
*Y*7	Conveying systems	238,050	293,820	304,270
*Y*8	Mechanical	998,410	2,490,000	3,515,600
*Y*9	Fire Protection	159,160	323,380	427,100
*Y*10	Electrical	716,130	1,574,000	2,199,500
*Y*11	General conditions	1,781,400	2,088,200	2,884,400
	Total project cost	14,666,650	17,557,030	20,100,180

sample. The purpose of bootstrap is to mimic the process of sampling observations from the population by resampling data from the observed sample (Efron & Tibshirani, 1993). Bootstrap sampling is commonly used to establish a level of uncertainty for the estimated parameters. Bootstrap method can also be utilized to improve prediction performance of neural networks when sparse data is available for training (Tsai & Li, 2008).

The bootstrap sampling method was implemented to develop range estimates for a case project. Model-2 is used for range estimating since it demands significantly less computation time when compared to Model-3, and the predictive accuracy of the model was not statistically different than the predictive accuracy of the Model-3. The case project consisted of a CCRC project under development, which was not included in the training data set. The following algorithm was implemented for range estimation with bootstrap sampling (Davison & Hinkley, 1997):

For $r = 1, 2, \ldots, 1000$

(1) sample $i_1^*, i_2^*, \ldots, i_{20}^*$ randomly with replacement from $\{1, 2, \ldots, 20\}$;
(2) for j 1, 2, ..., 20, set $X1_j^* = X1i_j^*, X2_j^* = X2i_j^*, \ldots, X21_j^* = X21i_j^*$
(3) for j 1, 2, ..., 20, set $Y1_j^* = Y1i_j^*, Y2_j^* = Y2i_j^*, \ldots, Y11_j^* = Y11i_j^*$
(4) train neural networks $N1_r, N2_r, \ldots, N11_r$ using the data sets:

$$\begin{array}{cccccccc} X1_1^*, & X2_1^*, & \ldots, & X21_1^*; & Y1_1^*, & Y2_1^*, & \ldots, & Y11_1^* \\ X1_2^*, & X2_2^*, & \ldots, & X21_2^*; & Y1_2^*, & Y2_2^*, & \ldots, & Y11_2^* \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \\ X1_{20}^*, & X2_{20}^*, & \ldots, & X21_{20}^*; & Y1_{20}^*, & Y2_{20}^*, & \ldots, & Y11_{20}^* \end{array}$$

(5) determine the estimated the cost components for the sample project $Y1E_r, Y2E_r, \ldots, Y11E_r$ using the trained neural networks $N1_r, N2_r, \ldots, N11_r$;

where, the star notation (*) indicates that * is not the actual data set, but rather a resampled version of the actual data set.

An empirical probability distribution function for each cost component is obtained using cost estimates made based on 1000 bootstrap samples. Empirical probability distribution function for the estimated total cost is obtained similarly; by adding cost component estimates for each bootstrap sample to obtain 1000 estimates for the total cost. Range estimates for the cost components and for the total cost for the case project is developed using the empirical probability distribution functions. Table 7 provides range estimates for 70% probability level. The 70% probability level indicates that there is a 70% chance that the total cost of the case project will be between $14,666,650 and $20,100,180. There is a 15% chance that the total cost will be less than $14,666,650, and similarly there is an 85% chance that the total cost will be less than $20,100,180.

Range estimates for the cost components provide an indication for the level of uncertainties included in the estimated costs. As an example the 85% probability estimate for the conveying systems is $304,270. This estimate is $66,220 or, 28% more than the 15% probability estimate of $238,050 for conveying systems. On the other hand, for site development the 85% probability estimate of $2,082,400 is $1,684,190 or, 423% more than the 15% probability estimate of $398,210. Comparison of range estimates indicated that, the uncertainties included in the site development cost estimates are significantly larger than the uncertainties included in the conveying systems cost estimates.

5. Conclusions

In this paper a combination of neural networks and bootstrap method is presented for conceptual range estimation of costs. Neural networks present a powerful tool for modeling of the complex relations between factors and costs. Bootstrap technique enables a pragmatic method for quantification of the prediction variability. The nonparametric bootstrap presented in this research avoids restrictive assumptions about the form of the underlying populations and enables a pragmatic method for development of prediction intervals for neural network models. Integration of range estimating technique with the neural networks provides crucial information to management especially during early stages of construction projects, and would hopefully improve the decision making process.

Elimination of the unimportant factors and Bayesian regularization were implemented to improve generalization capabilities of the neural network models. Elimination of the unimportant factors improved generalization hence, results indicate that inclusion of all of the factors in the input buffer may result in over fitting especially when sparse and noisy data is available for training. Prediction performance comparisons revealed that back propagation neural networks including the important factors was sufficiently parsimonious, since Bayesian regularization did not improve generalization significantly.

References

AbouRizk, S. M., Babey, G. M., & Karumanasseri, G. (2002). Estimating the cost of capital projects: an empirical study of accuracy levels for municipal government projects. *Canadian Journal of Civil Engineering, 29*, 653–661.

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

Cheng, M. Y., Tsai, H. C., & Sudjono, E. (2009). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications*. doi:10.1016/j.eswa.2009.11.080.

Chou, J. S. (2009). Web-based CBR system applied to early cost budgeting for pavement maintenance project. *Expert Systems with Applications, 36*, 2947–2960.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.

Demuth, H., & Beale, M. (2001). *Neural network toolbox user's guide*. Natick, MA: The MathWorks Inc.

Dogan, S. Z., Arditi, D., & Gunaydin, H. M. (2006). Determining attribute weights in a CBR model for early cost prediction of structural systems. *Journal of Construction Engineering and Management, 132*(10), 1092–1098.

Duran, O., Rodriguez, N., & Consalter, L. A. (2009). Neural networks for cost estimation of shell and tube heat exchangers. *Expert Systems with Applications, 36*, 7435–7440.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Foresee, F. D., & Hagan, M. T. (1997). Gauss–Newton approximation to Bayesian regularization. In *Proceedings of the 1997 international joint conference on neural networks* (Vol. 3, pp. 1930–1935). Houston: Texas.

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.

Gencay, R., & Qi, M. (2001). Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. *IEEE Transactions on Neural Networks, 12*(4), 726–734.

Karshenas, S. (1984). Predesign cost estimating method for multistory buildings. *Journal of Construction Engineering and Management, 110*(1), 79–86.

Kaiser, M. J. (2006). Offshore decommissioning cost estimation in the Gulf of Mexico. *Journal of Construction Engineering and Management, 132*(3), 249–258.

Kim, G. H., Seo, D. S., & Kang, K. I. (2005). Hybrid models of neural networks and genetic algorithms for predicting preliminary cost estimates. *Journal of Computing in Civil Engineering, 19*(2), 208–211.

Kouskoulas, V., & Koehn, E. (1974). Predesign cost-estimation function for buildings. *Journal of Construction Division, 100*(CO4), 589–604.

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management, 132*(7), 750–758.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation, 4*(3), 415–447.

Sato, F., Shimada, Y., Selaru, F. M., Shibatu, D., Meada, M., Watanabe, Gi, et al. (2005). Prediction of survival in patients with esophageal carcinoma using artificial neural networks. *Cancer, 103*(8), 1596–1605.

Shastri, V., Rabelo, L. C., Onjeyekwe, E., & Vila, J. (1998). Device-independent color correction for multimedia applications using neural networks and abductive modeling approaches. *Expert Systems, 15*(2), 110–119.

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

Sonmez, R. (2008). Parametric range estimating of building costs using regression models and bootstrap. *Journal of Construction Engineering and Management, 134*(12), 1011–1016.

Tsai, T. I., & Li, D. C. (2008). Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems. *Expert Systems with Applications, 35*, 1293–1300.

Touran, A., & Wiser, E. (1992). Monte Carlo technique with correlated random variables. *Journal of Construction Engineering and Management, 118*(2), 258–272.

Wang, H. J., Chiou, C., & Juan, Y. K. (2008). Decision support model based on case-based reasoning approach for estimating the restoration budget of historical buildings. *Expert Systems with Applications, 35*, 1601–1610.

Wang, W. C. (2002). SIM-UTILITY: Model for project ceiling price determination. *Journal of Construction Engineering and Management, 128*(1), 76–84.