Resilient Structures and Sustainable Construction
Edited by Pellicer, E., Adam, J. M., Yepes, V., Singh, A., and Yazdani, S.
Copyright © 2017 ISEC Press
ISBN: 978-0-9960437-4-8

FORECASTING CONCEPTUAL COSTS OF BRIDGE PROJECTS USING NON-PARAMETRIC REGRESSION ANALYSIS

YUANXIN ZHANG and R. EDWARD MINCHIN JR.

M.E. Rinker, Sr. School of Construction Management, University of Florida, Gainesville, USA

The cost of bridge construction is influenced by numerous internal and external factors, making it very difficult to approximate. Years prior to a bridge project letting, state highway agencies (SHAs) must set a reliable budget for the proposed project, but the information available to them is very limited. Developing reliable cost estimates for bridge projects during the early pre-construction phases is very important and challenging for SHAs. This study employed a non-parametric regression analysis technique—multivariate adaptive regression splines (MARS)—to model the conceptual cost of bridge projects. This novel approach does not require detailed construction documents and does not require strict assumptions to be valid for the developed model or to be reliable for the model predictions. MARS was applied to the empirical data gathered from a Florida Department of Transportation database. The 10-fold cross-validation method was employed in this study to assess model performance. The criteria to gauge overall model fit, generalizability, and prediction error were evaluated. The results revealed that the developed model consistently performed well, based on Cross-Validated R-square (CVRSq), Generalized Cross-Validation (GSV), Generalized R-square (GRSq), and max error.

Keywords: MARS, Early phase cost, Construction management, Infrastructure, Budget.

1 INTRODUCTION

Bridges connecting roadways are important components of the surface transportation system (Fragkakis *et al.* 2010). Given the aging and dilapidated highway bridges in most states of the US, increasingly more bridges are needed to improve the current roadway system. A small percentage of inaccurate estimates of bridge construction costs can lead to a tremendous waste of taxpayers' money.

The construction cost of bridges has a significant effect on the total cost of the highway projects. The dramatic cost growth of a bridge can greatly increase the cost of the entire highway project. This means other projects become short of funding, especially considering the limited financial resources of State Highway Agencies (SHAs). The construction cost of bridges is influenced by numerous internal and external factors, making it very difficult to approximate accurately. Furthermore, SHAs often must create a budget several years before a bridge project letting. Moreover, the information available to them is limited at that moment, which makes this a challenging task. This study introduced a non-parametric regression technique to estimate the construction cost of bridges during early pre-construction phases. Unlike conventional regression analysis, this technique does not require strict assumptions and is sufficiently flexible to model bridge costs with high reliability.

2 LITERATURE REVIEW

There is limited previous research on estimating bridge costs using parametric approaches in the literature. The existing research shows various focal points and objectives. The identified research works are summarized subsequently in chronological order.

Saito *et al.* (1991) developed a set of cost estimation models for bridge replacement by regression analysis. The analysis of variance (ANOVA) was first used to assess the bridge attributes' effects. Nonlinear and log-transformed linear models were then formed to predict bridge replacement costs. They were subsequently validated and tested. Residual diagnosis substantiated that log-linear regression models performed better than nonlinear regression models. After that, the model validation process, using a separate dataset of bridge replacement, showed that these log-linear models yielded accurate estimations of bridge replacement cost.

Chengalur-Smith *et al.* (1997), using nonlinear regression analysis, modeled the cost of bridge rehabilitation based on the data gathered from the New York State Department of Transportation. Each set of models separately predicted the cost of major bridge components (e.g., deck, superstructure, and substructure) in rehabilitation. Two distinct means were employed to predict the cost of the entire bridge. One was to estimate the costs of the individual parts of the bridge. The other was to predict the unit costs and then obtain the total cost. A range of methods were attempted to identify the best-performing model. The performance of the former approach was consistently better.

Mackie *et al.* (2010) proposed an estimation approach for the repair cost and time of bridges damaged by earthquakes. The proposed approach considered the damage and repair of bridge components and assemblies in groups. Based on the linear relation between damage states and repair points, the intensity-dependent repair cost ratio and time were computed, and plots were generated showing disaggregated repair costs by quantities and by the damaged and repaired groups of components.

Fragkakis *et al.* (2010) also used regression analysis to model the conceptual cost of concrete bridge foundations. Their analysis was divided into three phases. The first stage classified and selected foundation types through a rule-based method. Prediction models were then formed by stepwise regression to estimate material quantities. The last step was to calculate the total cost of bridge foundations. The authors claimed that the parametric approach proposed in this research generated fast and reliable estimates during early stages of a project.

Li and Tang (2011) employed an artificial neural network (ANN) coupled with fuzzy logic to forecast costs of roadway bridges in mountainous areas. That research used the bridge features as the independent variables. Fuzzy logic was used to filter the cost data of the bridge projects for model training. The results showed that the deviation between predicted costs and actual costs was within 10%, which indicated it is viable to use ANN to predict bridge costs in mountainous areas.

Hollar *et al.* (2012) attempted to model the preliminary engineering (PE) cost of bridge projects using regression analysis. They found that underestimation of PE costs is a common practice and is on a significant scale. The critical factors contributing to PE costs of bridges were identified. The model development process was also exhibited, and the model performance was evaluated. The prediction error of 42.7% was higher than expected, but the regression model yielded useful information regarding the estimation of PE costs.

The previous research performed by others in bridge cost estimating in the literature demonstrated a variety of methods and solutions, and different levels of accomplishments were achieved. In addition, the authors started off with distinctive aspects and different project phases. It is interesting that most of the scholars adopted regression analysis to build prediction models. The non-parametric regression introduced in this study, to the authors' best knowledge, has not been employed in cost estimation, particularly in bridge construction cost estimation.

3 MULTIVARIATE ADAPTIVE REGRESSION SPLINES

Multivariate adaptive regression splines (MARS), proposed by Friedman (1991), is a nonparametric and nonlinear regression modeling approach. Because it is nonparametric, MARS does not require any statistical assumptions about the underlying relations between the dependent and independent variables, which makes it a very flexible procedure. Not only can it model nonlinear relations between independent and dependent variables but it also allows interactions among few independent variables (Hastie *et al.* 2009).

Both MARS and conventional linear regression determine optimal models based on the minimum sum of the squared distances between the observed (actual) and predicted data. The sum of squared distances between the actual data and points on the MARS model is smaller than that between the data and the points on the conventional linear regression model. This means that the MARS method potentially fits the data better than the conventional linear regression approach. MARS fits piecewise regressions in different intervals of the independent variables space to approximate intricate relationships. Since each variable can run as an independent regression in the corresponding interval, generalization of MARS models should be at least as good as the conventional linear regression models, unlike the typical nonlinear regression model (Friedman 1991).

The location and number of hinges are the determinant factors for a MARS model and dictate the goodness-of-fit of the MARS models. An increase in the number of hinges may improve the model fit, but could undermine models' ability in generalization, and vice versa. Hence, the MARS training process aims to find a balance with an appropriate number of knots (hinges) without comprising the model's good generalization capability.

The mathematical equation of a typical MARS model consists of a series of weighted basis functions (BF) (or hinge functions) and can be mathematically expressed as Eq. (1):

$$\hat{f}(\mathrm{x}) = \beta_0 + \sum_{i=1}^{i=M} c_i \, B_i(\mathrm{x}) \tag{1}$$

where $B_i(\mathrm{x})$ is a BF or hinge function, which can take on one of three forms: (1) a constant 1 when there is only one such term; (2) a hinge (or basis) function that takes either max $(x-h)_+$ or max $(x-h)_-$ (see Eq. (2) and (3)) where h is a constant; or (3) a product of more than one hinge (or basis) function. Each c_i is a constant coefficient. β_0 is the intercept.

$$(\mathrm{X_i} - \mathrm{h_i})_+ = \begin{cases} (\mathrm{X_i} - \mathrm{h_i}) & \text{if } \mathrm{X_i} > \mathrm{h_i} \\ 0 & \text{if } \mathrm{X_i} \le \mathrm{h_i} \end{cases} \tag{2}$$

$$(\mathrm{h_i} - \mathrm{X_i})_+ = \begin{cases} (\mathrm{h_i} - \mathrm{X_i}) & \text{if } \mathrm{X_i} < \mathrm{h_i} \\ 0 & \text{if } \mathrm{X_i} \ge \mathrm{h_i} \end{cases} \tag{3}$$

The MARS model fitting process comprises two major steps: the forward selection and the backward elimination passes. In the forward pass, MARS begins with a model containing only the intercept term, the mean of the response values. It then repetitively adds basis functions in pairs to the model. At each step, it selects the pair of basis functions that can minimize the sum of squared errors to the greatest extent. To improve model fit, interactions between the BFs are also examined. The order of interactions is decided before the training initiates. This process continues until the change in the sum of squared errors is negligible. The forward pass tends to overfit a model with a large number of BFs. To prevent overfitting and maintain good ability in generalization, the backward pruning pass is then executed. In this step, terms are removed one at a time, dropping the least effective term until the best submodel is found.

Generalized cross-validation (GCV) is a criterion to measure the goodness-of-fit and poses a penalty for adding excessive BFs. Therefore, GCV is used to determine the terms to be discarded. The lower the GCV, the better. Its mathematical equation is shown in Eq. (4) and (5):

$$\text{GCV}\ (M) = \frac{1}{N}\sum_{1}^{N} RSS \Big/ [1 - \frac{\varphi(M)}{N}]^2 \tag{4}$$

$$\varphi\ (M) = (\text{M} + 1) + \text{d} \times \text{M} \tag{5}$$

where N denotes the number of observations in the dataset; RSS stands for the sum of squared errors; $\varphi\ (M)$ is the cost-complexity measure of a model containing M basis functions. This term aims to penalize the addition of knots/BFs to the MARS model because adding new BFs increases the complexity of the model and introduces noise in the data; d is the user-defined cost for each BF optimization. The higher the d value, the more BFs are to be excluded.

4 DATA ACQUISITION

Data used in this research were collected from a Florida Department of Transportation (FDOT) database. Data from 468 bridge projects were retrieved, among which three were dropped because they would form a small group in terms of location (Florida Turnpike). Likewise, another 16 projects were excluded for the same reason based on their contract types. Finally, 14 more data points were deleted because of missing contractor's past performance rating (CPPR) scores. Eventually, data from 408 projects were utilized in this study for model development and validation.

The original dataset contained a huge amount of information regarding each bridge project completed by FDOT between 2006 and 2014. The initial independent variables used in this study were location, change in the number of contract days, days used, change order amount, bituminous adjustment amount, fuel adjustment amount, CPPR scores, and year of letting. This study avoided redundant information and the use of unordered categorical variables, which cause discontinuities and impact model performance (Flood and Issa 2010).

5 MODEL DEVELOPMENT AND ASSESSMENT

The MARS model development was implemented through the earth package in the statistics software - R. Figure 1 provides a set of plots regarding the model selection in cross-validation (upper left) and the residuals vs. fitted graph (upper right), residual heteroscedasticity (lower left), and QQ plot (lower right). The maximum number of terms was set at 21. The results showed that 16 out of 21 BFs were selected based on 17 decimals. Six out of eight variables were predictors entered into the final model. The dependent variable was transformed by a quadratic root to improve the linearity and model fit.

The final model developed in this study is shown as Eq. (6):

$$\begin{gathered}\text{Bridge Cost} = 31.1082 + 0.0380\ \text{BF1} + 0.0824\ \text{BF2} - 0.1441\ \text{BF3} + 0.0529\ \text{BF4} \\ - 16.0859\ \text{BF6} + 0.0044\text{BF7} - 0.0004\ \text{BF8} - 0.0002\ \text{BF9} + 0.0919\ \text{BF10} \\ - 0.0009\ \text{BF11} + 0.0536\ \text{BF12} + 0.0136\ \text{BF14}\end{gathered} \tag{6}$$

where BF1 = B(160-DaysChanged), BF2 = B(DaysChanged-160), BF3 = B(137-DaysUsed),

BF4 = B(DaysUsed-137), BF5 = B(418999-ChangeOrder), BF6 = B(CPPR-106),
BF7 = B(Dist-4) * B(DaysUsed-137), BF8 = B(57-DaysChanged) * B(DaysUsed-137),
BF9 = B(DaysChanged-57) * B(DaysUsed-137), BF10 = B(160-DaysChanged) * B(CPPR-106),
BF11 = B(160-DaysChanged) * B(106-CPPR), BF12 = B(DaysChanged-160) * B(Year-2010),
BF13 =B(680-DaysUsed) * B(ChangeOrder-418999),
BF14 = B(DaysUsed-680) * B(ChangeOrder-418999),
BF15 =B(137-DaysUsed) * B(Year-2010).

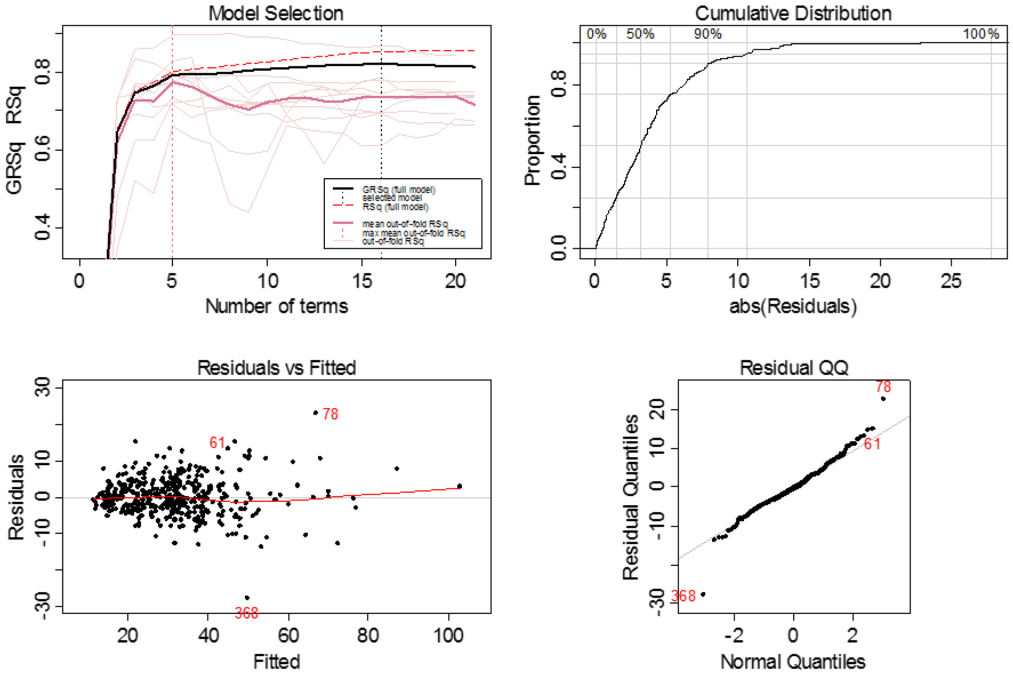

Figure 1. Model selection plots.

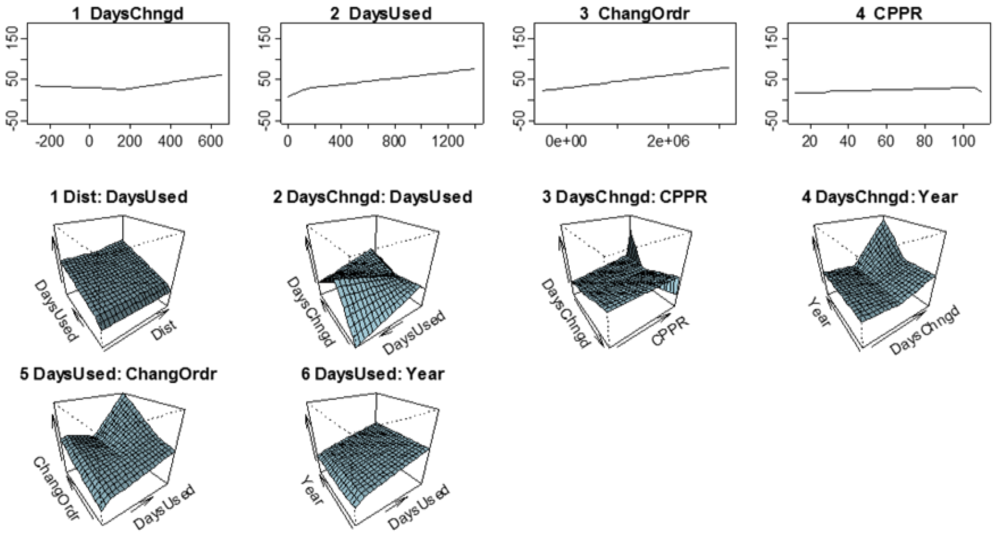

Figure 2. Graphical representation of the terms of cross-validated MARS model.

Figure 2 illustrates 10-fold cross-validated MARS models with two degrees of interaction allowed.

In 10-fold cross-validation, all the gathered data were randomly divided into 10 groups. Nine of the groups were used for model training, and the remaining one was held for model validation. This process was repeated a number of times to assess model performance. The R-square was 0.8534. The GCV was 33.0420, and the residual sum of squares (RSS) was 11024.1800. The generalized R-square (GRSq) was 0.8221, which is the normalized GCV. The cross-validation R-square (CVRSq) was 0.7361 with a standard deviation of 0.0640. The max error was -36 with a standard deviation of 26 based on the 10-fold cross-validation method.

6 CONCLUSION

This research aimed to assist SHAs in predicting the construction cost of bridges during early pre-construction phases with high reliability. Instead of using the conventional regression analysis, this study employed a non-parametric regression analysis, MARS, which has great flexibility and advantages in modeling bridge construction costs.

For the sake of time and accuracy, this study implemented the MARS model development and cross-validation in statistics software program R through the earth package. The final model included 16 of the 21 terms, which contained six independent variables. Based on the criterion for model fitting, the R-square was 0.8535, meaning the model explained the variance of the data very well. Regarding model generalizability, the GSV reported was 33.0412, and the GRSq was 0.8221. With regard to model predictive power, the CVRSq was 0.7361 with a standard deviation of 0.0640 and max error of -36 with a standard deviation of 26 based on the 10-fold cross-validation. The results indicated that the developed model performed consistently well.

References

Chengalur-Smith, I., Ballou, D. P., and Pazer, H. L., Modeling the Costs of Bridge Rehabilitation, *Transportation Research Part A: Policy and Practice,* 31(4), 281-293, 1997.

Flood, I., and Issa, R. A. A., Empirical Modeling Methodologies for Construction, *Journal of Construction Engineering and Management,* 36(1), 36-48,2010.

Fragkakis, N., Lambropoulos, S., and Tsiambaos, G., Parametric Model for Conceptual Cost Estimation of Concrete Bridge Foundations, *J Infrastruct Syst,* 17(2), 66-74, 2010.

Friedman, J. H., Multivariate Adaptive Regression Splines, *The Annals of Statistics*, 19(1), 1-67, 1991.

Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer New York, 2009.

Hollar, D. A., Rasdorf, W., Liu, M., Hummer, J. E., Arocho, I., and Hsiang, S. M., Preliminary Engineering Cost Estimation Model for Bridge Projects, *J. Constr. Eng. Manage.*, 139(9), 1259-1267, 2012.

Li, F., and Tang, H., Cost Prediction of Highway Bridge in Mountain Area Based on BP Neural Networks, *Journal of Civil Engineering and Management,* 4, 019, 2011.

Mackie, K. R., Wong, J., and Stojadinovic, B., Post-Earthquake Bridge Repair Cost and Repair Time Estimation Methodology, *Earthquake Eng. Struct. Dyn.*, 39(3), 281-301, 2010.

Saito, M., Sinha, K. C., and Anderson, V. L.,Statistical Models for the Estimation of Bridge Replacement Costs, *Transportation Research Part A: General*, 25(6), 339-350, 1991.