Joint CSCE Construction Specialty & CRC Conference 2025
Conférence conjointe spécialisée en construction de la SCGC et CRC-2025

Montreal, Quebec
July 28-31, 2025 / *28-31 juillet 2025*

DEVELOPMENT OF PARAMETRIC COST PREDICTION MODELS FOR LEGISLATIVE RECONSTRUCTION AND WIDENING PROJECTS

Shrestha, K.J.[1], Paul, P.[2], and Uddin, M.[3]

[1] Associate Professor, Department of Engineering, Engineering Technology, and Surveying, East Tennessee State University, Johnson City, TN 37614, USA, Email: shresthak@etsu.edu
[2] Graduate Research Assistant, Department of Engineering, Engineering Technology, and Surveying, East Tennessee State University, Johnson City, TN 37614, USA, Email: paulp1@etsu.edu
[3] Professor, Department of Engineering, Engineering Technology, and Surveying, East Tennessee State University, Johnson City, TN 37614, USA, Email: uddinm@etsu.edu

ABSTRACT: Accurate cost estimates for legislative roadway projects – such as reconstruction and widening projects – are essential for ensuring optimal use of available budget, proper planning of fiscal year projects, and successful execution of planned projects that support legislative priorities. Various factors such as project length, work type, and project location impact the overall project costs. However, addressing such factors to predict project costs can be challenging if the impact of the factors needs to be considered manually based on engineer's experience. If machine learning models can be developed to account for such factors, it would ease the cost-estimating process for the engineers. This study aims to develop parametric cost prediction models that enable engineers to estimate project costs quickly and reliably during the early stages of project development. To achieve the goal, this study collected and analyzed 11 years of legislative roadway infrastructure improvement projects to develop and validate cost prediction models. Multiple machine learning models are developed for each work type, for various geographic levels (state, region, and county), and using various project characteristics (e.g., project length, right of way cost, and route type). An innovative metric named "reliability index" is introduced as an indicator of the model's strength. Further, the reliability index and weight factors are used to aggregate cost prediction from multiple models to produce a single value. The final prediction is produced by adjusting this value for inflation. The validation of the model shows that 70% to 100% of the predictions complied with the AASHTO's guideline for estimation accuracy for the estimates produced during the project planning phase. The findings of the study are expected to aid highway engineers in predicting project costs quickly and reliably at the early phases of project development.

1. INTRODUCTION

Properly maintained transportation infrastructure is essential for economic growth and social development. However, aging transportation systems, continuous population growth, and increasing urbanization have significantly strained U.S. transportation networks. The American Society of Civil Engineers (ASCE) consistently rates U.S. infrastructure as subpar, which highlights the urgent need to upgrade roads, bridges, and transit systems across the country (ASCE, 2021). While the U.S. government has prioritized infrastructure improvement projects, such as, road expansions, bridge rehabilitations, and public transit upgrades, these initiatives are often capital-intensive and require accurate cost estimation to ensure

feasibility, availability of funding, and reduce unforeseen cost overruns. Although extensive research has been conducted on cost prediction models for new construction projects, limited research has focused on infrastructure improvement projects – despite their importance. Accurate cost estimation of infrastructure improvement projects in early phase of project development can be challenging because of various factors, such as, scope uncertainties, market fluctuations, project complexity, existing site conditions, integration with an existing transportation network, regulatory changes, and stakeholder demands. Inaccurate estimates can lead to severe consequences, including budget shortfalls, project delays, and reputational damage. Flyvbjerg et al. (2003) found that transportation projects worldwide frequently exceed initial budgets by 20% to 50%, with some cases surpassing 100%. These findings align with the U.S. Government Accountability Office report, which states that transportation project costs often exceed initial estimates by margins ranging from 2% to 211% (Sinnette 2004). Such overruns can trigger legal disputes, erode public trust, and strain public finances, which underscore the need for reliable cost prediction methodologies. To address such issue of cost overruns, many studies have developed cost estimating models aimed at improving the accuracy of cost estimates. However, existing studies have focused on developing models based on project types without due consideration for the rationale for the project. The legislative projects, which are often high-priority due to their alignment with legislative agendas, may exhibit different cost characteristics compared to non-legislative projects – even if they are of the project type (e.g., reconstruction). Moreover, many existing studies develop a single statewide model for all the projects in the state. The main goal of this study is to address this research gap by developing a multi-level machine learning models for estimating legislative project costs.

2. LITERATURE REVIEW

Parametric cost estimation has been widely studied in the context of transportation infrastructure projects, with researchers exploring its applications, methodologies, and challenges. In parametric cost estimating, the hidden relationships between project parameters and project cost are identified and extracted to enable future project cost estimation from project parameters. Parametric methods enable estimating project costs with limited project information, and hence it can be a powerful method for estimating project cost in early phase of project development (Piratla et al. 2024). However, uncertainties in the early phase of project development can often result in project cost errors ranging between -50% to +200%, according to the AASTHO's Practical Guide to Cost Estimating (2013) manual.

Many studies have been conducted to develop cost estimation models for infrastructure projects (Duverlie & Castelain 1999, Trost & Oberlender 2003, Karaca et al. 2020, Lowe et al. 2006, Gardner et al. 2017, Liu et al. 2011, Dominic & Smith 2014, Adel et al. 2016, Piratla et al. 2024). Various techniques used in these studies include case-based reasoning (CBR), regression models, and artificial neural network. For example, Duverlie and Castelain (1999) developed parametric models and case-based reasoning (CBR) models for cost estimating. Trost and Oberlender (2003) developed regression models by incorporating project complexity and site-specific factors, emphasizing the need for comprehensive data analysis. Karaca et al. (2020) highlighted the importance of top-down parametric approaches, especially for large, complex projects, by showing that carefully selected high-level cost drivers can yield more reliable conceptual estimates than traditional bottom-up methods. Lowe et al. (2006) developed a linear regression model with an R^2 value of 0.661 and a Mean Absolute Percentage Error (MAPE) of 19.3%. Their neural network model outperformed the linear regression model with an R^2 of 0.789 and a MAPE of 16.6%. Both models performed better than transitional estimation techniques which had MAPEs around 25%. Kim and Hong (2012) enhanced regression models by integrating it with CBR, accounting for variables like project specifications and environmental conditions, and improving planning-phase cost estimates.

Gardner et al. (2017) used bootstrap sampling and range estimation address variability by generating multiple cost scenarios that offered stakeholders a clearer understanding of potential fluctuations of the estimates. Liu et al. (2011) introduced hierarchical linear modeling and Multilevel Dirichlet Process Linear Models (MDPLM) to handle the complexities of highway project estimates and achieved better accuracy for complex projects with limited data. Dominic and Smith (2014) developed an empirical cost prediction model using artificial neural networks that demonstrated an average absolute percentage error of 3.67%, with 87%

of the model predictions falling within a ±5% range of actual project costs. Adel et al. (2016) similarly demonstrate the viability of using neural networks for conceptual highway cost models, combining a short list of high-impact project attributes such as project scope, project duration, length, width, and year of construction with genetic algorithm optimization. In another recent study, Piratla et al. (2024) developed a tool using machine learning and achieved accuracies between 61% and 84%for various project type. The study highlighted the trade-off between simplicity and accuracy, as factors like site-specific challenges and material availability remain uncertain during planning.

Overall, various parametric methods have pros and cons. At one end of the spectrum, overly simplified methods, such as cost per lane mile are likely to give less accurate results than more sophisticated models. At another end of the spectrum, to utilize more sophisticated models, engineers will need to have deep knowledge of specialized software tools that make it less accessible to engineers. Further, more sophisticated models, such as artificial neural network are also known as “black box” models, as it does not provide much insight into the inner workings of the model, which can make the end user less confident about utilizing such models. The middle of the road – which tend to be regression models – usually provides the balance of accuracy and ease of use. However, existing studies generally focus on developing one regression model to address all project characteristics – including geographical variations. Further, the traditional means of measuring the accuracy of models, such as R^2 and MAPE may not provide accurate picture of the model’s strength. As such, this study develops a new approach to predict legislative project costs using multiple regression models, and introduces a new metrics named “reliability index” to communicate the models’ strength or weakness.

3. METHODOLOGY

The overall methodology consists of 1) data collection, 2) data normalization, 3) model development, 4) reliability index, 5) prediction aggregation, and 6) prediction adjustment.

3.1 Data Collection

The research team collected project characteristics of 273 legislative projects from 2013 to 2024 from Tennessee Department of Transportation (TDOT). The project characteristics included project length, county, region, type of work (e.g., reconstruction, widening, etc.), route type, right of way (ROW) cost, and preliminary estimate.

3.2 Data Normalization

The preliminary estimates in the data are provided for the estimate year. As inflation impacts the project costs over time, when data from multiple years are to be analyzed together, the effect of inflation needs to be isolated before merging the data for further analysis. To achieve this, the data is normalized using annual inflation and time value of money equation. The relevant annual inflation can be computed using Highway Construction Cost Indexes (HCCIs) as HCCIs are representative of highway construction industry specific inflation (Shrestha, et al., 2017). Equation 1 presents the formula to compute the annual inflation from HCCI from desired model year, HCCI for estimate year, and the number of years between the model year and estimate year. For this study, National HCCI (NHCCI) is used for the annual inflation calculation as TDOT lacks TDOT-specific HCCIs.

$$[1]\ \text{Annual Inflation} = \left(\left(\frac{HCCI_{desired\ year}}{HCCI_{estimate\ year}} \right)^{\frac{1}{\text{Number of Years}}} - 1 \right) \times 100\%$$

3.3 Model Development

To develop Multiple Linear Regression (MLR) models, Minitab Statistical Software was used. The model development process included: 1) model specification, 2) model fitting and estimation, 3) assumptions and

model robustness, and 4) model evaluation and goodness of fit measures. Models were developed for three levels of geographic area: a) state level, b) region level, and c) county level. The state level model utilizes entire dataset, which each region level and county level models rely on data for the specific region or specific county. Further, separate models were developed for various work types (e.g., reconstruction, widening) in all geographic levels.

3.3.1 Model Specification

A MLR model was chosen due to its transparency and ability to establish relationships between multiple independent variables and a dependent variable. The general equation for the regression model is expressed in Equation 2.

[2] $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x + \epsilon$

where,

y	=	dependent variable
$x_1, x_2, \ldots\ldots x_n$	=	independent variables
β_0	=	y-intercept of the regression line
$\beta_1, \beta_1, \ldots \beta_n$	=	regression coefficients
ϵ	=	error term

By identifying key independent variables from domain knowledge, exploratory data analysis (EDA), and correlation analysis, the following equations have been developed:

$y = \beta_0 + \beta_1 \times$ Project Length

$y = \beta_0 + \beta_1 \times$ Project Length $+ \beta_2 \times \log$(ROW Cost)

$y = \beta_0 + \beta_1 \times$ Project Length $+ \beta_2 \times \log$(ROW Cost) $+ \beta_3 \times$ Route Type

$y = e^{\beta_0 + \beta_1 \times \log(\text{Length}) + \beta_2 \times \log(\text{ROW Cost})}$

$y = e^{\beta_0 + \beta_1 \times \log(\text{Length}) + \beta_2 \times \log(\text{ROW Cost}) + \beta_3 \times \text{Route Type}}$

Prior research underscores the importance of incorporating significant project attributes in cost modeling. Karaca et al. (2020)and Piratla et al. (2024) highlighted that area, length, width, and highway functional classification are significant predictors of project costs. Additionally, Goodrum et al. (2006) emphasize that right-of-way (ROW) constraints can lead to unplanned utility relocations, necessitating budget adjustments and increasing overall construction expenses. Liu et al. (2011) further established a significant interaction between ROW costs, project length, and estimated construction expenditures, reinforcing the necessity of integrating ROW considerations into cost estimation models. By systematically incorporating these influential factors, the developed equations aim to improve the reliability and applicability of parametric cost estimation.

3.3.2 Model Fitting and Estimation

The model parameters (β_0 coefficients) were estimated using the Ordinary Least Squares (OLS) method in the Minitab, which minimizes the sum of squared residuals (RSS) (Equation 3).

[3] $RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$

where,

y	=	actual observed value
$\hat{y}_i$	=	predicted value from the model

The stepwise regression technique was employed to iteratively add or remove predictors based on their statistical significance and impact on model accuracy. The p-values and adjusted R^2 values were used as criteria for including/excluding various parameters.

3.3.3 Assumptions and Model Robustness

Several key assumptions were tested to ensure the model's validity. Linearity was assessed to confirm a linear relationship between predictors and the dependent variable using scatter plots and residual analysis. The independence of observations was verified through the Durbin-Watson test, which detects autocorrelation in residuals. Homoscedasticity – to ensure constant variance of residuals across predictor levels – was examined using residual plots. Finally, multicollinearity was evaluated by computing the Variance Inflation Factor (VIF), where values exceeding 5 indicated high collinearity, necessitating variable removal or transformation.

3.3.4 Model Evaluation and Goodness-of-Fit Measures

To assess the performance of the model, several goodness-of-fit metrics were analyzed: a) coefficient of determination (R^2), b) ANOVA test, and c) residue analysis. R^2 measures the proportion of variance in the dependent variable explained by the independent variables. A higher R^2 generally tend to indicate better model performance with some caveats discussed in the next section. The overall significance of the model is assessed using ANOVA, which determines if the independent variables collectively explain a significant portion of the variation in the dependent variable. Various residual plots (e.g., residual vs. fitted plots) are used to visually assess whether the model's residuals follow the normality and randomness assumptions.

3.4 Reliability Index

Generally, it is more likely that the models based on larger number of data points are likely to be more reliable than the ones based on fewer data points. Thus, a higher value of R^2 alone could be misconstrued as a better model – especially if the model with higher value of R^2 is based on fewer data points. When developing models based on real data, it is possible that very few data points are available to develop a model. As such, this study introduces a new metrics named “reliability index.” The reliability index merges the R^2 value, and the relative size of the data points used to develop the model, to provide more meaningful reflection of the models’ strength. While there are no strict guidelines regarding the minimum dataset size required for developing a predictive model, statistical best practices suggest that a robust dataset should include at least 30 data points to ensure model validity and generalizability (Montgomery and Runger 2011).

Mathematically, a Reliability Index (RI) can be defined as Equation 4:

[4] $$Reliability\ Index = R^2 * Min\left(\frac{Data\ Point\ Count}{30}, 1\right)$$

The value of the reliability index ranges from 0 to 1.

3.5 Prediction Aggregation

To predict project cost for a new project, models of various geographic levels are available: a) state level, b) region level, and c) county level. For each of these geographic levels, multiple models are available (e.g., one with project length only, and one with project length and ROW). The predictions from these models are aggregated in two phases: a) aggregate multiple models within one geographic level and b) aggregate models from multiple levels. To aggregate multiple models within one geographic level, a reliability-index-weighted average of the predictions are computed using Equation 5.

[5] $$\text{Aggregated Predicted Cost at One Geographic Level} = \frac{\sum(\text{Predicted Cost} \times \text{Reliability Index})}{\sum \text{Reliability Index}}$$

Subsequently, to aggregate values from all geographic level, weighted average of the state, region, and county level values is calculated using Equation 6. The weights for the state, region, and county are 20, 30, and 50, respectively. The weights highlight the general notion that cost of projects that are close (e.g., from same county) are more likely to be similar that projects from larger geographic region (e.g., same region or same state).

$$[6]\ \text{Aggregated Predicted Cost} = \frac{\sum(\text{Predicted Cost} \times \text{Weight Factor})}{\sum \text{Weight Factor}}$$

3.6 Prediction Adjustment

Since the model is based on estimates normalized to a specific model year, the predictions will also be produced for the same model year. As such, the output from the model needs to be adjusted to ensure the costs are reflective of the project year. To achieve this, Equation 7 is used.

$$[7]\ E_c = E_m * (1+I)^{C_y - M_y}$$

where,

E_c = Estimated Cost for the Construction Year
E_m = Estimated Cost for the Model Year
I = Inflation Adjustment Factor
C_y = Planned Construction Year
M_y = Model Year

4. RESULTS

Data from 2013 to 2023 was used to develop models at various geographic levels. The R^2 values and reliability indexes were calculated for all the models of various geographic levels (Table 1). Some models yielded 100% R^2 values, a phenomenon attributed to the limitations of the dataset specific to those models. While a high R^2 value indicates that the model explains a significant proportion of the observed variance, this outcome may be misleading in cases where the dataset is constrained or lacks sufficient variability to capture the broader complexities of cost estimation. Such limitations can result in overfitting, where the model performs exceptionally well on the training data but fails to generalize to unseen or more diverse scenarios. Consequently, relying solely on R^2 to measure model accuracy may provide an incomplete or overly optimistic performance assessment. To address this issue, a reliability index was introduced as an improved metrics to evaluate model's strength. This index offers a comprehensive evaluation of model accuracy by incorporating the limitations of the dataset.

Table 1: R^2 values and Reliability Indexes of Models for Various Geographic Levels

Model	R^2 Value (%)		Reliability Index (%)	
	Maximum	Minimum	Maximum	Minimum
County Level	100	0.76	16.67	0.10
Region Level	100	0.03	72.98	0.02
State level	100	4.45	49.70	0.69

5. MODEL VALIDATION

A dataset consisting of 10 legislative project details for year 2024 was used to validate the performance of the model. The validation is presented in two forms: 1) compliance according to the AASHTO Cost Estimation Classification and 2) MAPE calculations.

All 10 projects comply with the prescribed error of -50% to +100% error that is anticipated in the 0% to 2% project maturity (Table 2). As project maturity increases, compliance percentages decline, with 40% of projects meeting the criteria in the Final Design Phase, where cost estimates are expected to be within -5% to 10%.

Table 2: Compliance of the Legislative Projects in reference to the AASHTO Cost Estimation Classification

AASHTO Cost Estimation Classification			Compliance Status of Legislative Projects (N = 10)	
Project Description Phase	Project Maturity (% project definition completed)	Estimate Range	No of complied Projects	% of Compliance
Planning	0 to 2%	-50% to 200%	10	100
	1 to 15%	-40% to 100%	7	70
Scoping	10 to 30%	-30% to 50%	6	60
Design	30 to 90%	-10% to 25%	5	50
Final Design	90 to 100%	-5% to 10%	4	40

Table 3 presents the MAPE values that provide a measure of relative error. Reconstruction projects show a lower MAPE of 16%, indicating a more precise cost estimation. In contrast, widening projects exhibit a significantly higher MAPE of 43%, suggesting more significant variability and potential inaccuracies in model predictions for this category.

Table 3: Different Statistics of the Model Validation

Work Type	Total No of Project	MAPE (%)
Reconstruction	3	16
Widen	7	43

The validation results highlight the model's strengths and limitations across different project phases and types. It performs well in the planning phase, achieving 100% compliance despite broad estimate ranges, but struggles to maintain accuracy as projects mature. Additionally, the model shows strong performance for reconstruction projects with MAPE 16% but faces difficulties with widening projects with MAPE 43%, likely due to more significant variability in scope and conditions.

6. CONCLUSION AND FUTURE WORK

This research developed multiple parametric cost prediction models for estimating legislative infrastructure improvement projects. By narrowing down the analysis to legislative projects only, it ensures that the models capture cost variations that are specific to the high priority legislative projects that advances legislative agendas for the state. Multiple models were developed for reconstruction and widening projects at various geographic levels (i.e., state, region, and county), and using various combination of project characteristics. A novel metric, termed “reliability index,” is introduced as an indicator of the overall reliability of the machine learning model. The reliability index is used to aggregate results for multiple models in the same geographic area. The aggregated results from each geographic area (state, region, and county) are then further aggregated using custom weight to ensure appropriate reflection of market conditions in the desired project location. This innovative approach enabled combining the strengths of multiple models for predicting cost for a specific project type. The result shows 70% to 100% compliance rate for AASHTO prescribed accuracy of the estimates for the planning phase. Future research should develop models for additional project types instead of focusing on legislative project type only. Further, a user-friendly tool should be developed to implement the findings of the study, which will enable state DOTs to quickly and reliably compute preliminary estimates for legislative projects. This framework and tool are expected to aid state DOTs in improving budgeting and planning of the legislative roadway projects.

7. ACKNOWLEDGMENT

The researchers would like to thank the Tennessee Department of Transportation (TDOT) and its staff members for funding this study (RES2024-08), providing valuable data, and providing insights about their current cost-estimating practices.

8. REFERENCES

Adel, K., A. Elyamany, A. M. Belal, and A. S. Kotb. 2016. “Developing Parametric Model for Conceptual Cost Estimate of Highway Projects.” *International Journal of Engineering Science and Computing*, 6 (7).

ASCE. 2021. “2021 Report Card for America’s Infrastructure.” Accessed November 9, 2024. https://infrastructurereportcard.org/.

Dominic, A. D. D., and S. D. Smith. 2014. “Rethinking construction cost overruns: Cognition, learning and estimation.” *Journal of Financial Management of Property and Construction*, 19 (1). https://doi.org/10.1108/JFMPC-06-2013-0027.

Duverlie, P., and J. M. Castelain. 1999. “‘Cost Estimation During Design Step: Parametric Method versus Case Based Reasoning Method.’” *The International Journal of Advanced Manufacturing Technology*, 15 (12): 895–906. https://doi.org/10.1007/s001700050147.

Flyvbjerg, B., M. K. S. Holm, and S. L. Buhl. 2003. “How common and how large are cost overruns in transport infrastructure projects?” *Transp Rev*, 23 (1). https://doi.org/10.1080/01441640309904.

Gardner, B. J., D. D. Gransberg, and J. A. Rueda. 2017. “Stochastic Conceptual Cost Estimating of Highway Projects to Communicate Uncertainty Using Bootstrap Sampling.” *ASCE ASME J Risk Uncertain Eng Syst A Civ Eng*, 3 (3). https://doi.org/10.1061/AJRUA6.0000895.

Goodrum, P. M., F. Kari, A. Smith, B. Slaughter, and C. N. Jones. 2006. *An Analysis of the Direct and Indirect Costs of Utility and Right-of-Way Conflicts on Construction Roadway Projects*.

Karaca, I., D. D. Gransberg, and H. D. Jeong. 2020. “Improving the Accuracy of Early Cost Estimates on Transportation Infrastructure Projects.” *Journal of Management in Engineering*, 36 (5). https://doi.org/10.1061/(asce)me.1943-5479.0000819.

Kim, B., and T. Hong. 2012. “Revised Case-Based Reasoning Model Development Based on Multiple Regression Analysis for Railroad Bridge Construction.” *J Constr Eng Manag*, 138 (1): 154–162. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000393.

Liu, M., J. E. Hummer, W. J. Rasdorf, D. A. Hollar, S. C. Parikh, J. Lee, and S. Gopinath. 2011. “Preliminary Engineering Cost Trends for Highway Projects.” *Report No. FHWA/NC/2010-10*. North Carolina Department of Transportation.

Lowe, D. J., M. W. Emsley, and A. Harding. 2006. “‘Predicting Construction Cost Using Multiple Regression Techniques.’” *J Constr Eng Manag*, 132 (7): 750–758. American Society of Civil Engineers (ASCE). https://doi.org/10.1061/(asce)0733-9364(2006)132:7(750).

Montgomery, D. C., and G. C. Runger. 2011. *Applied Statistics and Probability for Engineers*. *J R Stat Soc Ser A Stat Soc*. John Wiley & Sons.

Piratla, K. R., T. Le, M. S. Jamal, and Q. Do. 2024. “A Preliminary Cost Estimating Model for Transportation Projects.” *Report No. FHWA-SC-24-03*. South Carolina Department of Transportation, Federal Highway Administration.

Shrestha, K. J., Jeong, H. D. & Gransberg, D. D., 2017. “Multidimensional Highway Construction Cost Indexes Using Dynamic Item Basket.” J Const Eng Manag, 143 (8).

Sinnette, J. 2004. “Accounting for Megaproject Dollars.” Accessed November 9, 2024. https://highways.dot.gov/public-roads/julyaugust-2004/accounting-megaproject-dollars.

Trost, S. M., and G. D. Oberlender. 2003. “Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression.” *J Constr Eng Manag*, 129 (2): 198–204. https://doi.org/10.1061/(ASCE)0733-9364(2003)129:2(198).