

Preliminary Engineering Cost Trends for Highway Projects

by

Min Liu, Ph.D.

Joseph E. Hummer, Ph.D., P.E.

William J. Rasdorf, Ph.D., P.E.

of

North Carolina State University

Department of Civil, Construction, and Environmental Engineering

Campus Box 7908

Raleigh, NC 27695

Donna A. Hollar, P.E.

of

East Carolina University

Department of Construction Management

Rawl Building Room 329

Greenville, NC 27858-4353

Shalin C. Parikh, Graduate Student

Jiyong Lee, Graduate Student

Sathyanarayana Gopinath, Graduate Student

of

North Carolina State University

Raleigh, NC 27695

North Carolina Department of Transportation

Research and Development Group

Raleigh, NC 27699-1549

Final Report
Project: 2010-10

October 2011

Technical Report Documentation Page

1. Report No. FHWA/NC/2010-10	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Preliminary Engineering Cost Trends for Highway Projects		5. Report Date October 21, 2011	
		6. Performing Organization Code	
7. Author(s) Liu, M., Hummer, J., Rasdorf, W., Hollar, D., Parikh, S., Lee, J., Gopinath, S.		8. Performing Organization Report No.	
9. Performing Organization Name and Address North Carolina State University Department of Civil, Construction, and Environmental Engineering Campus Box 7908, Raleigh, NC 27695		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No.	
12. Sponsoring Agency Name and Address North Carolina Department of Transportation Research and Analysis Group 1 South Wilmington Street Raleigh, North Carolina 27601		13. Type of Report and Period Covered Final Report August 16, 2009 – December 31, 2010	
		14. Sponsoring Agency Code 2010-10	
15. Supplementary Notes			
16. Abstract <p>Preliminary engineering (PE) for a highway project encompasses two efforts: planning to minimize the physical, social, and human environmental impacts of projects and engineering design to deliver the best alternative. PE efforts begin years in advance of the project's construction letting, often five years or more. An efficient and accurate method to estimate PE costs would benefit transportation departments. Typically, departments estimate PE costs as a percentage of construction costs disregarding other project-specific parameters. By analyzing 461 North Carolina Department of Transportation bridge projects and 188 roadway projects let between 2001 through 2009, the research team developed statistical models linking variation in PE costs and PE duration with distinctive project parameters. The development of a user interface application aids agency users in executing the models to predict a project's PE cost ratio. Modeling strategies included multiple linear regression, hierarchical linear models, Dirichlet process linear models, and multilevel Dirichlet process linear models (MDPLM). The 461 bridge projects exhibited a mean PE cost ratio of 27.8% (ratio of PE cost over estimated construction cost) and a mean PE duration of 66.1 months. Mean PE cost ratio for the 188 roadway projects was 11.7% with a mean PE duration of 55.1 months. Project parameters utilized in the predictive models included project scope classification such as widening or new location, dimensional variables (project length, structure length, detour length, and number of spans); geographical region; and estimated costs for construction and right of way. The MDPLM minimized the mean absolute prediction error for bridges' PE cost ratio, but interpretation of variable effects and sensitivity is difficult because of the multilevel structure. Regression modeling results are also reported since sensitivity interpretation from them is more direct.</p>			
17. Key Words preliminary engineering cost estimation preliminary engineering duration Dirichlet process linear model		18. Distribution Statement	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 133	22. Price

DISCLAIMER

The contents of this report reflect the views of the authors and not necessarily the views of North Carolina State University or East Carolina University. The authors are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of either the North Carolina Department of Transportation or the Federal Highway Administration at the time of publication. This report does not constitute a standard, specification, or regulation.

ACKNOWLEDGEMENTS

The research team acknowledges the North Carolina Department of Transportation for supporting and funding of this project. We extend our thanks to the project Steering and Implementation Committee members:

Calvin Leggett, P.E. (Chair)
Terry Gibson, P.E.
Deborah Barbour, P.E.
Victor Barbour, P.E.
Majed Al-Ghandour, P.E.
Missy Pair, P.E.
Art McMillan, P.E.
David Smith, P.E.
Mike Bruff, P.E.
Moy Biswas, Ph.D., P.E.
Neal Galehouse, P.E.

The authors especially thank those who were instrumental in assisting with the project. Key NCDOT units providing guidance include the Program Development Branch (Calvin Leggett, Majed Al-Ghandour, Lisa Gilchrist, and Bill Martin), the Structure Inventory and Appraisal Unit (Cary Clemmons), the Schedule Management Office (Kim So, Rose Simson, and Anna Twohig), and the Financial Management Division (Burt Tasaico). Many additional NCDOT personnel provided suggestions and insights through participation in meetings and responding to inquiries. Their contributions positively influenced this research effort.

In addition to the authors, several graduate and undergraduate students contributed to the successful completion of this research: Ingrid Arocho and Lisa Moll of North Carolina State University, and Richard Hibbett of East Carolina University.

The research team also thanks the Southeastern Transportation Center whose early support was instrumental in transitioning this research topic into a funded project.

EXECUTIVE SUMMARY

The research goal was to investigate how to accurately estimate preliminary engineering (PE) cost and PE duration for NCDOT highway projects. Preliminary engineering (PE) for a highway project encompasses two efforts: planning to minimize the physical, social, and human environmental impacts of projects and engineering design to deliver the best alternative. PE efforts begin years in advance of the project's construction letting, often five years or more. An efficient and accurate method to estimate PE costs would benefit transportation departments. Typically, departments estimate PE costs as a percentage of construction costs disregarding other project-specific parameters.

During this project we completed a comprehensive study of the factors affecting PE costs and PE duration, built a database containing 8.5 years of NCDOT highway project data, and developed a computer application to help NCDOT estimate PE costs more accurately and efficiently. During our investigation, we discovered that our analyses would be most effective if the highway projects were segregated into two databases – bridge projects and roadway projects.

The research team acquired data from ten sources to build a database containing information on 461 bridge projects. All projects were let for construction between January 2001 and June 2009. Through correlation analyses and ANOVA, twenty-eight independent variables were identified. Members of the research team analyzed the bridge project data to develop separate predictive regression models for the PE cost ratio and the PE duration. Modeling strategies included multiple linear regression (MLR), hierarchical linear models (HLM), Dirichlet process linear models (DPLM), and multilevel Dirichlet process linear models (MDPLM). Mean absolute percentage error (MAPE) was used to rank predictive performance when each candidate model was applied to a validation set.

The 461 bridge projects exhibited a mean PE cost ratio of 27.8%. The recommended MLR model achieved a MAPE of 0.1889 and included eight variables with interactions. Since four of the eight variables selected were categorical, we investigated a HLM approach. The HLM allows data to be divided into subgroups based on categorical variable values. Then, using MLR, a model is uniquely fit to each subgroup. The HLM results for PE cost ratio prediction failed to outperform the baseline MLR when applied to our validation set (MAPE_{HLM} of 0.2425 compared with MAPE_{MLR} of 0.1889). The recommended MLR model is incorporated in the user interface application and includes the following eight variables (4 numerical and 4 categorical):

- Right of Way Cost to STIP Estimated Construction Cost
- Roadway Percentage of Construction Cost
- STIP Estimated Construction Cost
- Bypass Detour Length
- Project Construction Scope
- NCDOT Division
- Geographical Area of State
- Planning Document Responsible Party

We were unable to identify the PE authorization date 45 of the 461 bridge projects, thereby reducing the number of projects to 416 for bridge PE duration analyses. The mean PE duration for the 416 bridge projects was 66.1 months. With MLR, the best model for predicting PE duration achieved a MAPE of 0.4542. When applying a HLM for PE duration estimation, the MAPE improved to 0.2131. The HLM hierarchy used four tiers with successive subdivisions by categorical variable values. The tier structure, selected by maximizing information gain, is geographical area of the state (tier 1), NEPA document classification (tier 2), bridge project construction scope (tier 3), and planning document responsible party (tier 4). The five numerical variables selected for the model included:

- Right of Way Cost to STIP Estimated Construction Cost
- Project Length
- Roadway Percentage of Construction Cost
- STIP Estimated Construction Cost
- Number of Spans

Data acquisition for roadway projects was more difficult in comparison with bridge project data acquisition as the number of completed projects was considerably lower (188 roadways to 461 bridges). However, the value range of project characteristics was more expansive than for bridges. Duration data were especially difficult to acquire and were only available for 113 of the 188 roadway projects. The mean PE cost ratio for the 188 roadways was 11.7%. The mean PE duration for the 113 roadways was 55.7 months. We used MLR to fit models to the roadway data to predict the PE cost ratio and to predict the PE duration. The HLM developed to capitalize on categorical variable groupings had limited success because of the small number of projects within each subgroup of the hierarchy.

Through validation, the best regression model for PE cost ratio of roadway projects yielded a MAPE of 0.2773 and included the following four numerical variables:

- Right of Way Cost to STIP Estimated Construction Cost
- Project Length
- Length of Structures within Project
- Roadway Percentage of Construction Cost

This model is included in the user interface application.

In predicting PE duration of roadway projects, the best regression model included the following six variables (3 numerical and 3 categorical):

- Right of Way Cost to STIP Estimated Construction Cost
- Project Length
- STIP Estimated Construction Cost
- Project Construction Scope
- STIP Prefix
- NCDOT Division

The MAPE achieved through validation was 0.1375.

Regression analyses supported our assumption that project data are heterogeneous, meaning the relationships between the PE cost ratio and other variables are complicated and multiple relationships exist. To make a predictive model for such heterogeneous data, we extended the DPLM to a multilevel model, the MDPLM. MDPLM can deal with more complex situations than traditional regression techniques. The MDPLM is a convenient way of modeling in that it reduces human efforts. It is not necessary to describe how data are distributed or structured, nor specify the random effects. The MDPLM can fit complex data with less variance. We applied the MDPLM approach to data from 505 bridge projects let during a slightly different time interval, between January 1999 and June 2008. MDPLM achieved a mean absolute relative error (MARE) of 0.208. The MDPLM model used thirteen variables. Applying the DPLM modeling technique to 181 roadway projects yielded a MARE of 2.814. The roadway model includes 11 variables. The user interface application utilizes both the bridge MDPLM and the roadway DPLM to prepare PE cost ratio predictions.

The primary deliverable created from this research was a user interface application developed using Microsoft Visual C++ with MFC libraries. The interface acts as a hub to facilitate a user entering project data, directing that data to the selected modeling library, generating prediction results by executing the model, and exporting those results into a Microsoft Word report format. Four modeling libraries exist; two for bridge PE cost analyses (MDPLM and regression) and two for roadway PE cost analyses (DPLM and regression). For each project type (bridge or roadway), the user may specify either a Dirichlet process or a regression model to generate predicted PE cost ratio. The interface application is available on CD. We recommend an initial training session be scheduled involving the intended NCDOT users and the research team. The current interface does not support modeling for PE duration prediction.

TABLE OF CONTENTS

Disclaimer.....	ii
Acknowledgements.....	iii
Executive Summary.....	iv
Table of Contents.....	vii
List of Figures.....	x
List of Tables.....	xi
1.0 Introduction.....	1
1.1 Background.....	1
1.2 Preliminary Engineering (PE) Defined.....	1
1.3 Significance of the Research.....	1
1.4 Research Tasks.....	2
2.0 Literature Review.....	3
2.1 NCDOT Historical Performance on PE Budgeting.....	3
2.2 PE Estimation Efforts for the Transportation Industry.....	3
2.2.1 Total PE as a Percentage of Construction Costs.....	4
2.2.2 Planning Component of PE.....	5
2.2.3 Design Component of PE.....	5
2.2.4 PE Provider: In-house versus Consultant.....	6
2.3 Construction Cost Estimating for Transportation Projects.....	7
2.3.1 Estimate Development Timeline.....	7
2.3.2 Estimating Practices in the Transportation Industry.....	9
2.3.3 Research Efforts on Construction Estimating Techniques.....	10
2.3.4 NCDOT Construction Cost Estimating Practices.....	10
2.4 Applicable Statistical Analysis Techniques.....	12
2.4.1 Factor Analysis by Principal Component Analysis (PCA).....	12
2.4.2 Multiple Regression Techniques and Models.....	13
2.4.3 Multilevel Hierarchical Regression Modeling.....	14
2.5 Summary.....	14
3.0 Bridge Projects: PE Cost Analyses.....	16
3.1 Database Compilation.....	16
3.2 Validation Sampling.....	17
3.3 Response Variable for PE Cost Analyses.....	17
3.4 Independent Variables for Prediction.....	19
3.4.1 Variable Sensitivity.....	19
3.4.2 Variable Correlations.....	21

3.5	Multiple Linear Regression (MLR) Modeling.....	22
3.6	Hierarchical Linear Model (HLM)	23
4.0	Bridge Projects: PE Duration Analyses	27
4.1	Background on PE Duration Estimation.....	27
4.2	Database Compilation.....	27
4.3	Validation Sampling	29
4.4	Independent Variations for Prediction	29
4.4.1	Variable Sensitivity.....	30
4.4.2	Variable Correlations	31
4.5	Hierarchical Linear Model (HLM)	32
4.5.1	Tier Selection using Information Gain Theory	32
5.0	Roadway Projects: PE Cost and PE Duration Analyses	37
5.1	Database Compilation.....	37
5.2	Validation Sampling	38
5.3	Response Variables: PE Cost Ratio and PE Duration.....	38
5.3.1	PE Cost Ratio.....	38
5.3.2	PE Duration.....	39
5.3.3	Back Transformation of Response Variables.....	39
5.4	Independent Variables for Prediction	40
5.4.1	Variable Sensitivity and Correlations with PE Cost Ratio Response	42
5.5	Multiple Linear Regression (MLR) Modeling.....	42
5.5.1	MLR Baseline for PE Cost Ratio Prediction.....	43
5.5.2	MLR Baseline for PE Duration Prediction	43
5.6	Hierarchical Linear Model (HLM)	44
5.6.1	HLM for PE Cost Ratio	44
5.6.2	HLM for PE Duration	48
6.0	Estimating PE Cost Ratio using Multilevel Dirichlet Process Linear Modeling	50
6.1	Introduction.....	50
6.2	Background.....	50
6.2.1	Heterogeneous Population	50
6.2.2	Dirichlet Process Linear Mixture Model (DPLM).....	50
6.3	Multilevel Dirichlet Process Linear Model (MDPLM)	51
6.3.1	Model Construction.....	52
6.3.2	Covariate Extension of Layers	52
6.3.3	Overfit Model of Layers	53
6.3.4	The Layer of Prediction	54

6.4	Bridge Model for Predicting PE Cost Ratio.....	55
6.5	Roadway Model for Predicting PE Cost Ratio	57
7.0	User Interface Application	59
7.1	Interface Programming	59
7.2	Estimate Initialization Inputs	59
7.3	Inputs Based on Project Type	62
7.4	Estimate Results and Archived Summary Report.....	65
8.0	Findings and Conclusions	67
8.1	Modeling Results: Bridge Projects	67
8.1.1	MDPLM Results for Bridge PE Cost Ratio Prediction.....	67
8.1.2	Regression Results for PE Cost Ratio Prediction	67
8.1.3	Regression Results for PE Duration Prediction	68
8.2	Modeling Results: Roadway Projects	69
8.2.1	DPLM Results for Roadway PE Cost Ratio Prediction	69
8.2.2	Regression Results for PE Cost Ratio Prediction	70
8.2.3	Regression Results for PE Duration Prediction	71
8.3	Conclusions.....	71
9.0	Recommendations	74
9.1	Future Needs	74
9.2	Future Opportunities	74
10.0	Implementation and Technololgy Transfer Plan.....	75
10.1	Research Products	75
11.0	References.....	76
12.0	Appendices.....	81
12.1	Regression Results for Bridge PE Cost Ratio Models	82
12.2	Regression Results for Roadway PE Cost Ratio Models	95
12.3	Sample Functional Cost Estimate Worksheet	97
12.4	Analyses of Financial Details	98

LIST OF FIGURES

Figure 2.1 DOT Reported PE Costs as a Percentage of Construction Costs	5
Figure 2.2 Timeline of Construction Cost Estimates for DOT Transportation Projects.....	8
Figure 2.3 NCDOT Estimating Timeline.....	11
Figure 3.1 Distribution of PE Cost Ratio for Bridge Projects.....	18
Figure 3.2 Distribution of Cubed Root of PE Cost Ratio for Bridge Projects	18
Figure 3.3 Comparison of Cost Trends for Bridge Projects.....	22
Figure 4.1 Mean PE Duration versus Year of Letting	28
Figure 5.1 Distribution of Transformed Response Variable for PE Cost Analyses.....	39
Figure 6.1 MDPLM Schematic	52
Figure 6.2 System-wide Representation of the MDPLM.....	54
Figure 6.3 MDPLM (Bridge Model) Expected Error Measurements	56
Figure 6.4 MDPLM (Bridge Model) Layer-wise Predictive Results.....	57
Figure 7.1 User Interface - Login Screen.....	60
Figure 7.2 User Interface – Project Startup Screen.....	60
Figure 7.3 User Interface – Project Information Screen	61
Figure 7.4 Excerpt from a Preliminary Estimate Worksheet	62
Figure 7.5 User Interface – Roadway Project Details.....	63
Figure 7.6 User Interface – Bridge Project Details	64
Figure 7.7 User Interface - Results Popup Screen	65
Figure 7.8 User Interface - Archived Estimate Summary Report	66
Figure 8.1 MDPLM (Bridge Model) Layer-wise Predictive Results.....	67
Figure 8.2 DPLM (Roadway Model) Predictive Results	69
Figure 12.1 R-2405 Breakdown of PE Expenditures over the PE Phase (Trend 1).....	100
Figure 12.2 R-3427 Breakdown of PE Expenditures over the PE Phase (Trend 1).....	101
Figure 12.3 R-2907 Breakdown of PE Expenditures over the PE Phase (Trend 2).....	101
Figure 12.4 R-2207 Breakdown of PE Expenditures.....	102
Figure 12.5 R-2248 Breakdown of PE Expenditures.....	103
Figure 12.6 R-3807 Breakdown of PE Expenditures.....	104
Figure 12.7 R-3303 Breakdown of PE Expenditures.....	105
Figure 12.8 R-3415 Breakdown of PE Expenditures.....	106
Figure 12.9 R-2904 Breakdown of PE Expenditures.....	107
Figure 12.10 R-2823 Breakdown of PE Expenditures.....	108
Figure 12.11 R-2643 Breakdown of PE Expenditures.....	109
Figure 12.12 R-2616 Breakdown of PE Expenditures.....	110
Figure 12.13 R-2604 Breakdown of PE Expenditures.....	111
Figure 12.14 R-2600 Breakdown of PE Expenditures.....	112
Figure 12.15 R-2555 Breakdown of PE Expenditures.....	113
Figure 12.16 R-2552 Breakdown of PE Expenditures.....	114
Figure 12.17 R-2538 Breakdown of PE Expenditures.....	115
Figure 12.18 R-2517 Breakdown of PE Expenditures.....	116
Figure 12.19 R-2302 Breakdown of PE Expenditures.....	117
Figure 12.20 R-2236 Breakdown of PE Expenditures.....	118
Figure 12.21 R-2213 Breakdown of PE Expenditures.....	119

LIST OF TABLES

Table 2.1 Findings by State Auditor Regarding Costs of NCDOT Highway Projects	3
Table 2.2 Estimate Classifications and Associated Accuracy Levels	9
Table 2.3 Listing of Innovative Techniques for Improving Construction Costs Estimates	10
Table 2.4 NCDOT Estimate Types and Associated Contingencies	11
Table 2.5 Summary Table of Relevant Research.....	15
Table 3.1 Bridge Projects Database	17
Table 3.2 Variance Values for Back Transformation of Response Variables	19
Table 3.3 Independent Variables for Bridge Projects	20
Table 3.4 Categorical Variables: Sensitivity on Cubed Root of PE Cost Ratio.....	21
Table 3.5 Numerical Variables: Correlation with Cubed Root of PE Cost Ratio	22
Table 3.6 Information Gain for Hierarchical Subgroup Combinations	24
Table 3.7 Hierarchical Organization of Bridge Projects	25
Table 4.1 Scope of Categorical Independent Variables used in PE Duration Estimation.....	29
Table 4.2 Scope of Numerical Independent Variables used in PE Duration Estimation	30
Table 4.3 One-Way ANOVA Results.....	30
Table 4.4 Correlation Coefficient Values	31
Table 4.5 Hierarchical Level Table for PE Duration	33
Table 4.6 Information Gain Results for Tier Classification.....	34
Table 5.1 Roadway Projects Database.....	37
Table 5.2 Roadway Dataset Differences for PE Cost and PE Duration Analyses	38
Table 5.3 Variance Values for Back Transformation of Response Variables	40
Table 5.4 Independent Variables for Roadway Projects	41
Table 5.5 Categorical Variables: Sensitivity on Cubed Root of PE Cost Ratio.....	42
Table 5.6 Numerical Variables: Correlation with Cubed Root of PE Cost Ratio	42
Table 5.7 Baseline MLR Model for Predicted Cubed Root of PE Cost Ratio.....	43
Table 5.8 Baseline MLR Model for Predicted Cubed Root of PE Duration.....	43
Table 5.9 Information Gain Comparison for PE Cost Ratio Hierarchical Schemes	45
Table 5.10 Hierarchical Organization for Predicting PE Cost Ratio	46
Table 5.11 HLM Model for Predicted Cubed Root of PE Cost Ratio	47
Table 5.12 Information Gain Comparison for PE Duration Hierarchical Schemes.....	48
Table 5.13 Hierarchical Organization for Predicting PE Duration	48
Table 5.14 HLM Model for Predicted Cubed Root of PE Duration	49
Table 6.1 MDPLM (Bridge Model) Listing of Variables	56
Table 6.2 MDPLM (Bridge Model) Comparison of Error Measurements	56
Table 6.3 Roadway Model Listing of Variables	57
Table 6.4 DPLM (Roadway Model) Error Measurements.....	58
Table 6.5 DPLM (Roadway Model) Averaged Error Measurements	58
Table 8.1 MDPLM Performance for Final Bridge Model at Layer 4	67
Table 8.2 Performance of PE Cost Ratio Regression Models for Bridges	68
Table 8.3 Performance of PE Duration Regression Models for Bridges	69
Table 8.4 DPLM Performance for Final Roadway Model.....	69
Table 8.5 Performance of PE Cost Ratio Regression Models for Roadways	70
Table 8.6 Performance of PE Duration Regression Models for Roadways	71
Table 8.7 Error Comparisons between Bridge and Roadway Models	71
Table 12.1 Full Bridge MLR Model Definition.....	82
Table 12.2 Reduced Bridge MLR Model with Year of Letting Omitted.....	82
Table 12.3 Bridge HLM Model Definition.....	83
Table 12.4 Bridge HLM (with Surrogate) Model Definition.....	87
Table 12.5 Bridge HLM (with Mean as Surrogate) Model Definition	91

Table 12.6 Roadway MLR (with Interactions) Model Definition	95
Table 12.7 Roadway HLM (with no tiers) Model Definition	95
Table 12.8 Roadway HLM (with 1 tier) Model Definition.....	96
Table 12.9 Listing of PE Phase Sub Activities	98

1.0 INTRODUCTION

This research addresses the need to accurately estimate preliminary engineering (PE) costs required to plan and design North Carolina Department of Transportation (NCDOT) highway projects. Additionally, the duration required to complete PE activities was investigated.

1.1 Background

Over the last thirty years, transportation projects have increased in number and complexity. Accuracy of project cost estimates has become a larger concern. Initial focus was placed on construction cost accountability. DOTs have adopted tighter financial controls on other project cost components such as right-of-way (ROW), utilities, mitigation, and PE costs. DOTs report cost accounting information to their governing bodies. External agencies routinely audit DOTs and investigate project costing.

Public reporting of project costs differentiates between ROW costs and construction costs. Comparisons between actual and budgeted costs are common. Similar reporting of project PE cost is increasing. PE cost estimates are frequently based on estimated project construction costs. A search of transportation literature identified that most DOTs use a constant or sliding percentage of estimated construction costs to develop a PE budget. The most frequent percentage cited is ten percent of estimated construction costs [WSDOT 2002].

For NCDOT highway projects, PE costs have been a significant portion of total project costs. Consistent with other agencies, PE costs are generally estimated by NCDOT to be about 10% of total project cost. However, there can be a wide range depending on project type and complexity. It is difficult to accurately estimate PE costs in the early project stages since an accurate definition of project scope has not yet been established. This is problematic; NCDOT is unable to plan and budget PE funds efficiently, which then affects total project cost control. It is important for NCDOT to avoid project cost escalation. One way to do so is by estimating PE cost more accurately.

Intuitively, factors such as project type, project complexity, whether PE efforts were performed in-house or by consultants, and when PE was conducted should have significant impacts on PE costs. Previously however, there had not yet been a comprehensive study defining the full set of factors and estimating their effects on NCDOT PE costs. In the literature, there is a large body of work on cost estimation in general, but nothing available specifically on PE cost estimation.

1.2 Preliminary Engineering (PE) Defined

For this study, PE was defined as the efforts required to plan and design a highway project for construction. PE begins when a specific highway project first receives funding authorization for planning and/or design activities. The delivery of the construction documents for project letting marks the end of PE.

Consistent with other researchers' definitions, PE in this study does not include ROW acquisition or construction [Turochy et al. 2001; WSDOT 2002]. In general, highway projects have PE, ROW, and construction components. If projects require feasibility studies and/or mitigation, these costs are tracked separately and are not part of PE. PE also excludes any efforts undertaken before a specific project is identified or funding is authorized, and any efforts undertaken after a construction contract has been let.

1.3 Significance of the Research

PE costs usually comprise a significant portion of the total project costs. Accurate PE cost estimation can help NCDOT make the best possible programming and budgeting decisions. This research benefits

NCDOT by identifying a method to improve the accuracy of PE cost estimates. With better PE cost estimates, funding allocations can be proactive, matching the specific needs of each project.

Continuing to estimate PE costs using a fixed percentage method is inefficient over the project cycle. Some projects require less PE funding while others require more. Under-allocation or over-allocation necessitates management actions to redistribute PE funds. Avoiding such redistributions improves total project cost control.

Additionally, by having a project-specific PE cost estimate generated at the beginning of each project's preconstruction phase, the PE budget status becomes trackable as a performance metric.

The data collected and analyzed during this research were specific to the NCDOT. However, the research findings should prove helpful to other transportation agencies. City DOTs in North Carolina may be able to use the prediction equations directly. DOTs in other states, and other countries, may be able to apply the methods demonstrated here to develop their own equations.

1.4 Research Tasks

The goals of this research were to complete a comprehensive study of the factors affecting NCDOT PE costs and PE duration and to build tools to assist NCDOT in estimating PE costs accurately and efficiently. These goals were met through completion of the following tasks:

- Develop a comprehensive list of factors affecting PE cost and PE duration based on a literature review and NCDOT project data.
- Conduct statistical analyses of past NCDOT highway projects to identify the factors that have significant impacts on PE costs and PE duration.
- Develop databases of NCDOT highway projects.
- Build modeling tools for PE cost ratio estimation and PE duration estimation.
- Develop an easy-to-use software application to help NCDOT project managers estimate PE cost accurately and quickly.

2.0 LITERATURE REVIEW

The research team reviewed journals, agency reports, academic research studies, and NCDOT documents to assess the status of PE cost and PE duration estimating practices, factors influencing PE costs and PE duration, and applicable analysis techniques. Few studies were targeted specifically at PE cost estimating for transportation projects. The studies found focused on one phase of preconstruction such as environmental planning or technical design. For a small sample of highway projects, factors influencing costs were identified. DOT agencies reported PE estimating practices similar to that used by NCDOT. Some agencies (notably Virginia and Texas) do include PE budgets in their STIP documents. Most agencies report using a percentage of estimated project construction cost to establish PE budgets. Regression techniques have been implemented in construction cost estimating procedures, especially preliminary and early cost estimates. References to regression analysis applied specifically to PE cost estimating were not found.

Section 2.5 contains a summary table of the relevant research reviewed.

2.1 NCDOT Historical Performance on PE Budgeting

A search for NCDOT documentation identifying specific PE budgets located the 2008 North Carolina State Auditors' report on highway projects' cost and schedule performance [Merritt 2008]. The State Auditor reviewed total project costs and PE, ROW, and construction cost components for 292 highway projects. Construction of audited projects was completed between April 1, 2004 and March 31, 2007. Table 2.1 contains the State Auditor's assessment of project costs.

Table 2.1 Findings by State Auditor Regarding Costs of NCDOT Highway Projects

Project Cost Component	Aggregate Estimated Costs (in millions)	Aggregate Actual Costs (in millions)
PE	\$ 73.4	\$ 117.1
ROW	\$ 83.8	\$ 148.7
Construction	\$ 650.3	\$ 1,020.3
Total	\$ 807.5	\$ 1,286.1

The cost figures reported in Table 2.1 identify several cost trends:

- Actual costs exceeded estimated amounts for all cost components.
- PE expenditures increased 59 percent (\$43.7 increase compared to original \$73.4 estimate).
- Actual PE expenditures represented 18 percent of estimated construction costs (\$117.1 compared to \$650.3). Theoretically, if only 10.3 percent was budgeted (the average of PE percentages reported by WSDOT (2002)) NCDOT would have experienced insufficient PE funding to complete PE activities requiring additional PE funding authorizations.

2.2 PE Estimation Efforts for the Transportation Industry

PE can be broken into two components – planning and design. The planning component of PE includes all efforts required to prepare and deliver a project's environmental documents in the preconstruction phase. In the typical project cycle, planning is initiated before design. Design PE includes all efforts required to produce the project's construction documents. The summation of these components is a project's total PE. All PE tasks occur in the preconstruction phase. The personnel involved in planning and design PE functions may be involved in related actions during or after construction. For example,

design efforts related to construction change orders are not considered PE, but construction engineering. Similarly, environmental monitoring during construction is not PE, but construction compliance.

2.2.1 Total PE as a Percentage of Construction Costs

In a 2001 study, the Virginia Transportation Research Council (VTRC) reported on the current state of practice among nine DOTs with regard to cost estimating of highway projects during the planning phase. The premise of the study was that a highway project cost estimate included three elements: PE costs, right-of-way (ROW) and utility costs, and construction costs. The study sought to identify how these elements were estimated during the project planning phase, before preliminary design efforts began. VTRC defined PE as “the development of a project and the expenses to be incurred when a project advances from planning to design to when the project design is complete.” The VTRC researchers noted that, “ROW and PE are the states’ most difficult cost categories to estimate and often present the greatest challenges and deviations with the cost estimation process.” Most respondents reported that PE costs were estimated as a percentage of estimated construction costs, with percentages between five and twenty percent. Two of the nine DOTs reported using alternate techniques in certain circumstances. Texas estimates PE cost as a function of ROW width on some projects. Delaware utilizes a detailed form to guide how PE costs should be estimated based on project size. The PE cost of large projects can be estimated as a percentage of construction costs, whereas small projects should estimate required man-hours to determine PE costs. Moderate sized projects may utilize a combination of both estimating methods [Turochy et al. 2001].

As part of a comparative analysis of construction costs, Washington State DOT [WSDOT 2002] collected information from twenty-five DOTs whose members served on the AASHTO Subcommittee on Design. Survey participants were asked to identify their typical project PE cost as a percentage of construction cost. PE was defined as, “the work that goes into preparing a project for construction.” The average PE cost among respondents was 10.3 percent of construction costs and the range of costs reported was between four and twenty percent. NCDOT participated in the survey and reported PE costs of ten percent of construction costs.

Figure 2.1 summarizes geographically the PE costs acquired from the two surveys. Responses from twenty-eight DOTs were acquired and have been mapped in Figure 2.1 [Turochy et al. 2001; WSDOT 2002].

Building upon their 2001 study, VTRC assisted Virginia DOT (VDOT) during 2004 to find and implement a construction estimating tool. The estimating tool selected for statewide implementation was based on an existing spreadsheet application developed by the Fredericksburg District of VDOT. In the enhanced statewide tool, PE costs can be estimated for roadways and bridges separately, and then combined to provide a total PE estimate. For roadways, a cost curve relating PE costs to construction costs was derived using data from thirty completed VDOT roadway projects. The resulting ratio of PE costs to construction costs ranged from eight to twenty percent. PE costs were found to be inversely related to construction costs. To verify that the template’s PE cost curve was applicable for statewide use, an additional 135 completed VDOT roadway projects were included to update the derived PE cost curve. For bridges, a similar PE cost curve was derived and confirmed using data from twenty-three completed bridge projects [Kyte et al. 2004a, 2004b].

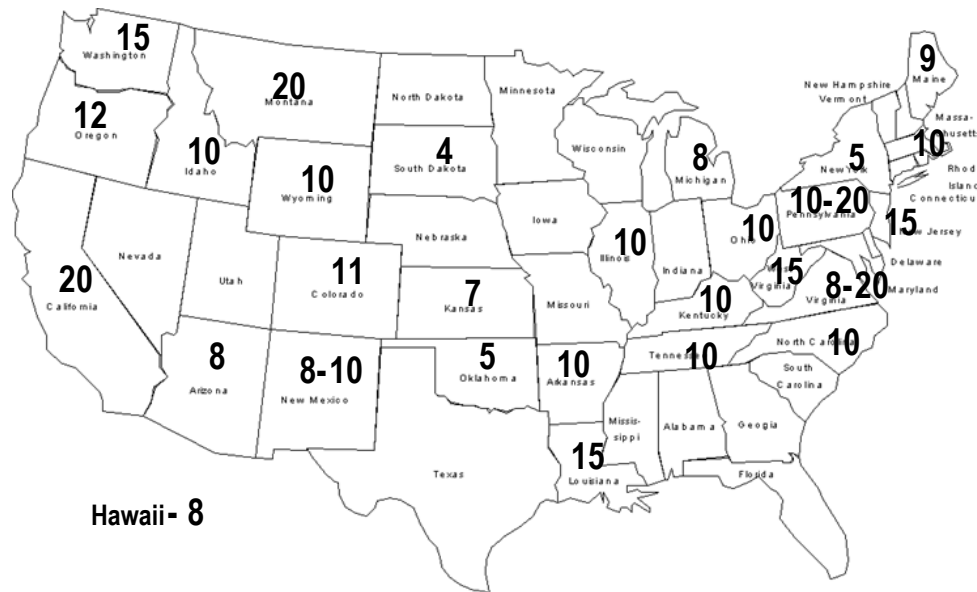


Figure 2.1 DOT Reported PE Costs as a Percentage of Construction Costs [Turochy et al. 2001; WSDOT 2002]

2.2.2 Planning Component of PE

Just one reference on estimating the PE costs associated with planning could be found in the literature. WSDOT noted in their 2002 survey that the preconstruction efforts required to meet environmental compliance requirements are highly variable between projects. Instead of asking survey respondents to quantify environmental compliance costs, WSDOT attempted to capture how these costs typically change during the preconstruction phase. Twenty-one of the twenty-five respondents (84 percent) indicated variability ranges from zero to ten percent. Three other respondents (12 percent) indicated higher variability in the eleven to twenty percent range [WSDOT 2002].

2.2.3 Design Component of PE

More information on the design component of PE can be found in the literature than the planning component. The work of Nassar et al. (2005) provides a significant contribution to the literature specifically addressing PE design costs of transportation projects. Nassar sought to create a model to estimate costs of design consultants' efforts. The model was based on data from 59 Illinois Department of Transportation (IDOT) projects. IDOT projects advertised for consultant design have a complexity factor assigned to assess the anticipated difficulty level of design. Nassar's research did not address how this value was assigned, but did note that process was "controversial" [Nassar et al. 2005].

A non-linear regression model using transformations of the variables below was analyzed:

- Initial planned construction cost (programmed costs)
- Complexity factors
- Percent of bridges projects
- Percent of roadways projects

The best fit model was a log transformation using only one independent variable, the initial planned construction cost, to predict consultant design costs [Nassar et al. 2005]. The prediction error of the model was not reported.

Gransberg and others (2007) investigated the correlation of design fees to construction “cost growth from the initial estimate” termed CGIE. Using 31 projects of the Oklahoma Turnpike Authority (OTA), Gransberg confirmed that an inverse relationship existed between design fees and construction cost growth. Their conclusion asserted that as design fees decrease, construction cost growth (from initial estimate to final closeout) increases. Correlating design fees to design quality, the results support the premise that allocating sufficient funding in design reduces the likelihood of construction cost increases from the initial estimate. Gransberg’s measurement of cost growth from the initial estimate was a departure from other studies that measured cost growth only from the bid price. The construction cost growth (CGIE) for all projects in the study was 9.65 percent. Thus, the difference between final construction cost and the initial estimate was less than ten percent.

Gransberg’s study provided quantitative data on design costs. However, the sample size was small. Of the 31 projects investigated, 13 were roadway projects and 18 were bridge projects. The average design cost for all projects was 5.2 percent. The roadway projects design costs averaged two percent, whereas the bridge projects exhibited design costs nearly four times higher (7.6 percent). The researchers concluded “bridge design projects should command a relatively higher design fee than roadway projects due to the increased complexity of design.” [Gransberg et al. 2007]

2.2.4 PE Provider: In-house versus Consultant

Wilmot et al. (1999) presented a concise summary of 17 studies that compared the design costs of in-house staff with that of consultants. These studies typically concluded that consultant design costs were greater than in-house costs. However, the magnitude of this difference varied significantly depending on the comparison methodology utilized. Wilmot et al. identified the difficulties inherent in earlier comparisons that used a similar project methodology. Though significant effort was made to identify and compare similar projects, no two projects are exactly alike.

Wilmot et al. proposed using a paired cost comparison on the same project by generating an estimated design cost to compare with the actual design cost. For in-house projects, the comparable consultant design cost would be estimated. Similarly, projects by consultants would be estimated as though completed in-house. Thus, every project compared would have two design costs, one actual and one estimated. These paired costs comparisons were analyzed using thirty-seven projects (twenty designed in-house and seventeen designed by consultants) completed between 1995 and 1997 by the Louisiana Department of Transportation and Development (LDOTD). After including cost factors such as overhead rates, space rental, and insurance in both paired costs, the overall comparison found that in-house costs are approximately 80 percent of consultants’ costs. This difference was found to be statistically significant at the 95 percent level. However, the difference was largely accounted for by the additional in-house efforts required to prepare and supervise the consultants’ contract. This supervisory effort amounted to an extra 50 hours of in-house time on average for the projects studied [Wilmot et al. 1999].

Two other incidental aspects of the Wilmot et al. study are worth noting. Using a paired cost comparison, the effect of project characteristics such as complexity, uniqueness, size, or type was eliminated. When comparing two costs for the same project, these characteristics would be equally accounted for in both cost figures. Thus, there is no way to infer how such characteristics may affect design costs. Wilmot et al. also discussed the difficulties in acquiring the necessary data from typical DOT databases. Most DOTs lack integrated databases making the data “less useful and less available.” Wilmot et al. suggests that DOTs would benefit from using “integrated client-server databases” [Wilmot et al. 1999].

VDOT had previously studied the cost difference between roadway design performed in-house compared with using consultants. (VDOT 1999 as reported by Kyte et al. 2004a). Consultant design costs were found to be fifty percent higher than in-house designs. This finding was incorporated into the VDOT statewide estimating tool. The estimator can enter the percentage of roadway design performed by

consultants and the tool automatically applies the fifty percent cost multiplier to the applicable portion of PE costs. The adequacy of the multiplier was verified using cost data from 29 consultant projects and 107 in-house projects [Kyte et al. 2004a].

The cited research of Wilmont et al. (1999) and Kyte et al. (2004a) above specifically addressed design functions. Both researchers agree that consultant design costs more than in-house design. There is no literature comparing planning costs based on provider. This research will study cost allocation for planning and design for selected NCDOT projects. A correlation between PE planning costs and PE provider is anticipated, similar to the findings reported above for design costs. Research findings will be used to validate this assumed correlation.

2.3 Construction Cost Estimating for Transportation Projects

Typically, PE costs are reported as a percentage of construction costs. Extensive research efforts to improve construction costs estimation practices have been undertaken. A full review of all such research is outside the scope of this PE estimation investigation. However, this section identifies innovative estimating techniques organized along a project timeline.

2.3.1 Estimate Development Timeline

Construction estimates are prepared at multiple times within a project's lifecycle. Few details are known at the beginning of this cycle, making accurate construction estimating difficult. The Cost Estimate Classification System developed by AACE International (2003) asserts that the degree of project definition should be "the primary characteristic to categorize estimate classes." As a project progresses through its lifecycle, more information becomes available and the degree of project definition increases [AACE International 2003]. The additional information allows for refined construction estimates. The timeline of Figure 2.2 aims to illustrate when, along the project definition spectrum, innovative techniques could be applied.

Figure 2.2 contains three horizontal bands representing estimate classification (top), project timeline with definition levels (center), and typical estimating practices within the transportation industry (bottom). The estimate classification band (top of Figure 2.2) combines the work of AACE International (2003); AbouRizk, Babey, and Karumanasseri (2002); and Schexnayder, Weber, and Fiori (2003). The naming convention and number of estimate classes vary between researchers. Similarly, the anticipated level of estimating accuracy for each researcher's classification varies. Table 2.2 provides a comparative summary of the estimate classifications and corresponding accuracy levels. The multiple classification systems are included in Figure 2.2 to frame the estimating techniques identified.

The center band of Figure 2.2 shows the level of project definition (as a percentage of total project definition) along the estimate timeline. Associating project definition levels with estimate classifications is subjective. However, for all classification schemes, project definition increases to the right along the timeline. Oval markers, labeled T1 through T10, are positioned along the timeline. Each marker represents a research technique developed to improve construction cost estimating. Table 2.3 lists each technique shown on the timeline. The position of the marker along the timeline addresses the quantity and quality of input information needed for each technique. All techniques aim to improve the accuracy of construction estimates. Section 2.3.3 provides further details on the techniques identified.

The typical estimating practices utilized by DOTs are identified in the bottom horizontal band of Figure 2.2. Byrnes (2002) and Schexnayder et al. (2003) surveyed all fifty state DOT agencies to determine current practices. The following section describes their findings on estimating personnel and methodologies employed at various stages along the estimate development timeline.

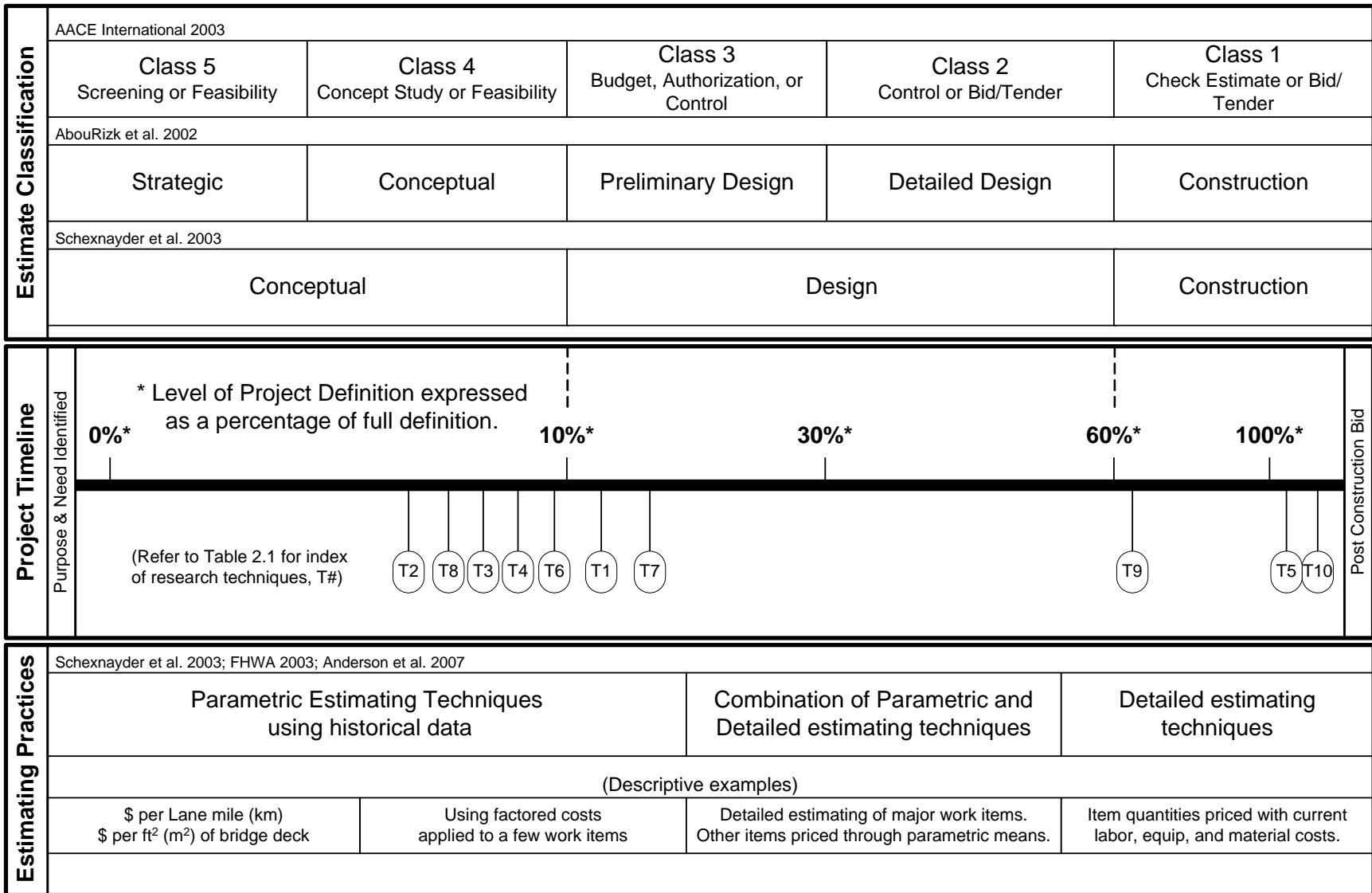


Figure 2.2 Timeline of Construction Cost Estimates for DOT Transportation Projects

Table 2.2 Estimate Classifications and Associated Accuracy Levels

Researcher	Estimate Classification	Estimate Accuracy Level (Percentage)
AACE International (2003)	Class 5 – Screening or Feasibility	Low: -20 to -100 High: +40 to +200
	Class 4 – Concept Study or Feasibility	Low: -15 to -60 High: +30 to +120
	Class 3 – Budget, Authorization, or Control	Low: -10 to -30 High: +20 to +60
	Class 2 - Control or Bid/Tender	Low: -5 to -15 High: +10 to +30
	Class 1 - Check Estimate or Bid/Tender	Low: -5 High: +10
AbouRizk, Babey, and Karumanasseri (2002)	Strategic	± 50
	Conceptual	± 30
	Preliminary Design	± 20
	Detailed Design	± 10
Schexnayder, Weber, and Fiori (2003)	Construction	± 10
	Conceptual	± 40
	Design	± 15 to ± 5
	Construction	± 5

2.3.2 Estimating Practices in the Transportation Industry

DOT personnel perform project estimating for most highway projects. Approximately half of the DOTs organize their estimators into a unit dedicated to estimating. Others accomplish estimating tasks using personnel assigned to design or contract administration units. In two-thirds of the DOTs, estimators have a minimum of ten years of experience. In 42 states where external consultants prepare cost estimates, DOT personnel review those estimates in detail [Schexnayder et al. 2003].

The estimating techniques used depend on the amount of information known at the time of estimate development, the amount of time available to prepare the estimate, and the experience of the estimator. Schexnayder et al. (2003) found that two-thirds of DOTs do not have a structured estimating manual. Less experienced estimators learn estimating techniques “on-the-job” from more experienced colleagues.

DOTs use three general approaches to estimating [FHWA 2003, Schexnayder et al. 2003]:

- Parametric estimating using historical cost figures.
- Detailed estimating using quantity takeoff techniques and pricing of labor, equipment, and materials.
- A combination of parametric and detailed techniques.

The level of project definition, as referenced along the estimating timeline, influences which estimating approach is used. Parametric estimating can be used when scoping information is very limited (project definition level is less than 5 percent). For example, if only location, length, and number of lanes are known, parametric estimating is effective [Anderson et al. 2007]. Parametric estimates rely on historical cost databases and defined relationships between cost items. Many DOTs use Trns*port, an AASHTO sponsored software package [FHWA 2003]. Trns*port’s Cost Estimating System module streamlines parametric estimating [Anderson et al. 2007]. When projects are fully scoped and detail design efforts are

underway (project definition level exceeds 50 percent), detailed estimating techniques are commonly used. Detail estimating of scope line items use quantity takeoffs with material, labor, and equipment pricing. Schexnayder et al. (2003) found DOTs performing detailed estimates do so only for the major work items that account for 65 to 80 percent of project costs. The remaining work items are estimated using parametric tools [Schexnayder et al. 2003].

2.3.3 Research Efforts on Construction Estimating Techniques

Innovative techniques for estimating construction costs have been developed as an improvement or alternative to the current methods DOTs employ. Refer to the center band of Figure 2.2 and to Table 2.3 for ten examples of techniques researchers have investigated. Each coded oval marker (T1 through T10) on the timeline corresponds to a research effort shown in Table 2.3. The positioning along the timeline indicates the level of project definition required to implement the technique.

Table 2.3 Listing of Innovative Techniques for Improving Construction Costs Estimates

Timeline Marker	Researcher
T1	Al-Tabtabai, H., Alex, A. P., Tantash, M. (1999). "Preliminary Cost Estimation of Highway Construction using Neural Networks." <i>Cost Engineering</i> (Morgantown, West Virginia), 41(3), 19-24.
T2	Cheng, M., Tsai, H., Hsieh, W. (2008) "Web-Based Conceptual Cost Estimates for Construction Projects using Evolutionary Fuzzy Neural Inference Model." <i>Automation in Construction</i> , In Press, Corrected Proof.
T3	Chou, J., Peng, M., Persad, K. R., O'Connor, J. T. (2006). "Quantity-Based Approach to Preliminary Cost Estimates for Highway Projects." <i>Transportation Research Record</i> , (1946), 22-30.
T4	Chou, J., Wang, L., Chong, W. K., O'Connor, J. T. (2005). "Preliminary Cost Estimates Using Probabilistic Simulation for Highway Bridge Replacement Projects." <i>Proceedings, Construction Research Congress 2005: Broadening Perspectives - Proceedings of the Congress</i> , San Diego, CA. April 5-7, 2005. American Society of Civil Engineers, 939-948.
T5	Gkritza, K., and Labi, S. (2008). "Estimating Cost Discrepancies in Highway Contracts: Multistep Econometric Approach." <i>Journal of Construction Engineering and Management</i> , 134(12), 953-962.
T6	Hegazy, T., and Ayed, A. (1998). "Neural Network Model for Parametric Cost Estimation of Highway Projects." <i>Journal of Construction Engineering and Management</i> , 124(3), p. 210-218.
T7	Kyte, C. A., Perfater, M. A., Haynes, S., Lee, H. W. (2004). "Developing and Validating a Tool to Estimate Highway Construction Project Costs." <i>Transportation Research Record</i> , (1885), 35-41.
T8	Molenaar, K. R. (2005). "Programmatic Cost Risk Analysis for Highway Megaprojects." <i>Journal of Construction Engineering and Management</i> , 131(3), 343-353.
T9	Shaheen, A. A., Fayek, A. R., AbouRizk, S. M. (2007). "Fuzzy Numbers in Cost Range Estimating." <i>Journal of Construction Engineering and Management</i> , 133(4), 325-334.
T10	Williams, T. P. (2005). "Bidding Ratios to Predict Highway Project Costs." <i>Engineering, Construction and Architectural Management</i> , 12(1), 38-51.

2.3.4 NCDOT Construction Cost Estimating Practices

Figure 2.3 illustrates the construction estimating practices of NCDOT using a similar timeline organization. The classification naming presented by AbouRizk et al. (2002) most closely correlates to NCDOT's project development process. Milestone symbols identify decision points, termed concurrence points, included in the typical project development process.

Concurrence points are defined by NCDOT (2008f) as:

- CP1 Purpose and need, and study area defined
- CP2 Detailed study alternatives carried forward
- CP2A Bridging decisions and alignment review
- CP3 Least environmentally damaging preferred alternative (LEDPA) selection
- CP4A Avoidance and mitigation
- CP4B Thirty percent hydraulic review
- CP4C Permit drawings review

NCDOT prepares five types of construction cost estimates throughout the project’s development [Lane et al. 2008]. The oval markers positioned below the estimate timeline correlate estimate type with concurrence points and project definition level. NCDOT utilizes a detailed estimating technique for major work elements when preparing estimates. As project definition increases, the uncertainty in major work items decreases. Thus, each estimate type has different contingency percentages to account for uncertainty in the roadway work items and the structure work items. Table 2.4 summarizes these contingency rates for the five estimate types. NCDOT construction estimates are generally within five percent of bid amounts [Lane et al. 2008].

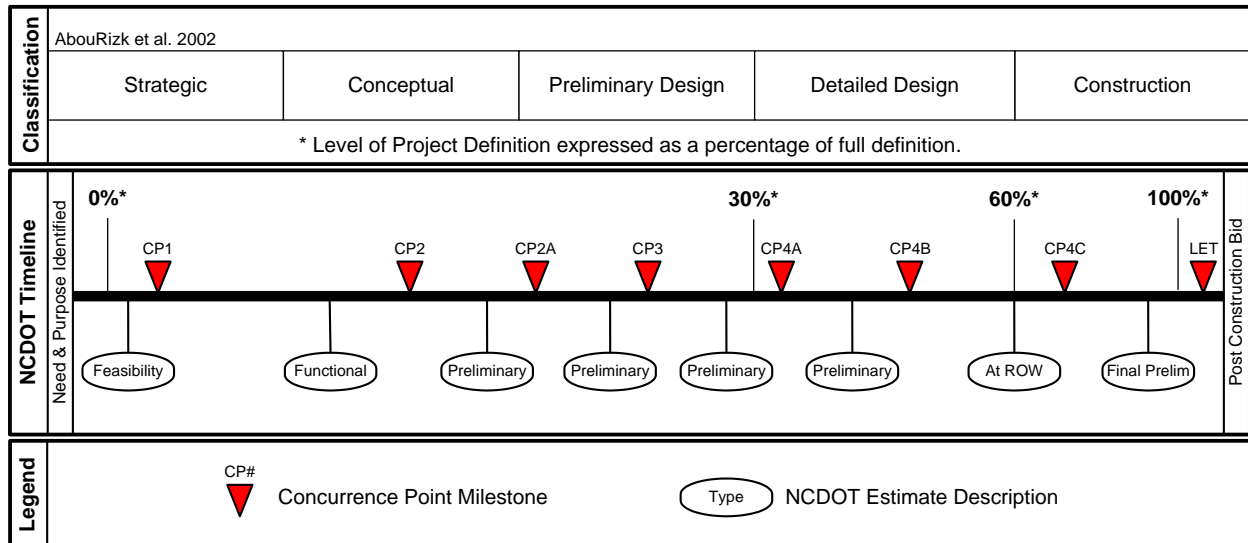


Figure 2.3 NCDOT Estimating Timeline

Table 2.4 NCDOT Estimate Types and Associated Contingencies [Lane et al. 2008]

Construction Estimate Description	Contingency Applied to Roadway Portion	Contingency Applied to Structure Portion
Feasibility	+55%	+15%
Functional	+45%	+15%
Preliminary	+35%	+10%
At ROW	+25%	+10%
Final Preliminary	+15%	+5%

2.4 Applicable Statistical Analysis Techniques

Literature related to factor selection techniques and applications of multiple linear regression is reviewed in the following section.

2.4.1 Factor Analysis by Principal Component Analysis (PCA)

Factor analysis techniques are used to discover any underlying “factor” or hypothetical variable that explains the interrelationships and variability in observed variables. Principal component analysis (PCA) is a factor analysis technique used to transform a set of observed variables into a new set of ranked “factors”. The first factor explains the greatest amount of the data variance. Each additional factor explains less of the variance in descending order. Applying PCA typically reduces the number of variables needed to construct a model. When a large number of variables exist and interrelationships between variables are suspected, PCA groups variables into factors that are independent and thus easier to incorporate in regression modeling [Kim and Mueller 1978].

Lam et al. (2008) utilized PCA to reduce the variables influencing design-build project success from 42 to 12. The original 42 variables were identified through a survey of 92 Hong Kong design-build professionals. Employing PCA, twelve factors were identified accounting for approximately 80 percent of the data variability. When the resulting twelve factors were included in regression modeling to predict design-build project success, only three were found statistically significant. The final regression model, using three factors, yielded an adjusted R^2 value equal to 0.549. Thus, Lam et al. proposed that design-build project success could be forecast using three factors. To validate this proposition, five projects’ success was determined by expert evaluation and by the regression model. Comparing these results using a paired-sample t test confirmed that the regression results did not differ from the experts’ evaluation [Lam et al. 2008]. This research illustrates how PCA can assist in reducing variable quantities for efficient regression modeling.

Interpreting the causal structure of PCA identified factors can be difficult since each factor is a grouping of original, observed variables. To aid in simplifying interpretation of causal effects, PCA results are “rotated” so that the observed variables are loaded onto only one factor [Kim and Mueller 1978]. This simplified loading allows the researcher to interpret and characterize the factor based on the loaded variables.

Akintoye (2000) provides a thorough discussion of using PCA to first identify factors and then apply a rotation technique to allow interpretation of factors. With data from 84 survey responses, Akintoye sought to identify factors influencing a contractor’s cost estimating method. Survey respondents included construction firms of varying sizes (very small to large) who performed both building and civil construction. Respondents rated the influence level of twenty-four variables on their company’s estimating practices. Through PCA with a varimax rotation technique, the twenty-four variables were reduced to seven factors explaining 70.4 percent of the data variance. The resulting seven factors identified by Akintoye are shown below:

1. Project complexity
2. Technological requirements
3. Project information
4. Project team requirement
5. Contractual arrangement
6. Project duration
7. Market requirements

PCA is a statistical technique. Interpretation of factor causal structure is the researchers’ responsibility. Akintoye’s thorough discussion of each factor emphasizes this responsibility [Akintoye 2000].

Trost (1998) collaborated on the Construction Industry Institute's efforts to improve the accuracy of early cost estimates. By analyzing 67 projects in the process industry, Trost developed an estimate scoring tool that would predict the accuracy of early cost estimates. The scoring tool was based on 45 project variables related to estimate preparation. These variables were then grouped to address the "who," the "how," and the "what" related to cost estimation. A fourth category, "other", captured any variable that did not fit into the who, how, or what categories. The 45 variables exhibited multicollinearity. To overcome this, PCA was employed to regroup the variables into orthogonal factors. PCA resulted in the 45 variables being regrouped into 11 factors. Regression of percent cost overrun on the eleven factors resulted in five significant factors explaining 51 percent of the data variation. These five factors were weighted to account for approximately 76 percent of a project's estimate score. The remaining six factors contributed 24 percent to the estimate score. Using the assigned weighting, a new project is scored by evaluating all 45 variables. The lower the score, the higher predicted the estimate quality. A higher quality estimate would require a lower contingency amount to ensure no cost overruns at a chosen confidence level. The estimate scoring tool allows users to determine the contingencies associated with various confidence intervals [Trost 1998; Oberlender and Trost 2001].

2.4.2 Multiple Regression Techniques and Models

The research efforts of Lowe et al. (2006) utilized linear regression techniques to predict the construction costs of buildings using data from 286 buildings constructed in the United Kingdom. A predictive tool was desired that could be used during the early stages of construction cost estimation before the detailed design has been completed. Lowe et al. identified forty-one input variables for use in the regression modeling. The input variables were categorized as either strategic (5 variables), site related (4 variables), or design (32 variables). Departing from previous regression modeling efforts, the researchers rejected using costs at completion of construction (termed "raw costs") as the response (dependent) variable, since cost variance would not be constant across all project sizes. Lowe et al. instead used regression to predict the $\log(\text{cost})$, the cost per unit area ($\$/\text{m}^2$), and $\log(\text{cost per unit area})$. Both forward and backward model selection techniques were used to identify the variables included in each regression model.

From the initial 41 variables considered, 14 variables were included in the best performing model. Of these 14 variables, only two were not associated with specific design parameters: project duration (strategic) and site access (site related). Thus, Lowe et al. concluded that the key linear drivers of cost were predominately design specific. R^2 and mean absolute percentage error (MAPE) was used to judge model performance. Additionally, the researchers recommend reviewing each model's error spread by analyzing the response (dependent) variable versus error illustrated on scatter plots. A review of the model's error distribution for normality is also recommended. Lowe et al. reported that their backward $\log(\text{cost})$ model yielded a R^2 of 0.928 and a MAPE of 19.3% for predicting the cost of building construction. However, the error review highlighted underestimation of very expensive projects and overestimation of very inexpensive projects [Lowe et. al. 2006].

Odeck (2003) also used regression to identify project factors associated with construction cost overruns of 620 Norwegian road projects. Odeck proposed using a quadratic regression model to determine if the impact on cost overrun depended on the magnitude of certain variables, specifically project cost, project delay, and project duration. These variables were included in the regression equation in first and second order form (i.e. cost and cost^2). Initially twenty project parameters were considered as candidate variables. Using a stepwise selection technique, four unique variables were included. These variables were cost, project duration, completion year, and geographic region. Cost and project duration were also included as second order variables (cost^2 and $\text{project duration}^2$). Odeck's regression model only explained about 20% of the variation in cost overruns (adjusted R^2 value of 0.21). Other project factors, not identified in the regression model, influenced the variation in cost overruns. From his model's partial regression coefficients, Odeck concluded that estimated cost overruns decreased with project costs, increased with project duration up to a point and then decreased, and varied with geographic region.

Specifically, Odeck found that cost overruns were more predominate among smaller road projects in Norway [Odeck 2004].

2.4.3 Multilevel Hierarchical Regression Modeling

Multilevel models contain more than one level within which the parameters vary. Multilevel data are those structures that consist of multiple units of analysis one nested within the other [Steenbergen and Jones 2002]. Multilevel modeling consists of regressing by layers. Categorical variables are arranged as separate layers in the model and the dependent variable is regressed against the independent variables in each layer with the previous layer being held constant for every successive layer in the model. The models capture this layered structure of the data and determine how each layer interacts and impacts the dependent variable of interest [Steenbergen and Jones 2002]. It is desirable to have the largest possible number of units in the first level of this multilevel hierarchy since the power of the model is largely unaffected by the number of units in the lower or lowest level [Snijders 2005].

Steenberg and Jones (2002) state that the goal of multilevel analysis to be to account for variance in a dependent variable that is measured in the lower level of analysis by considering the information from all levels of analysis. It allows researchers to combine multiple levels into a single comprehensive model. Multiple levels of analysis allow the model to be more specific than a single level model. Multilevel models also allow checking of cross level interactions and make it possible to determine whether causal variables are singular or vary within the levels. Multilevel models allow for better comparative studies dealing specifically with different time periods or variables of different time periods. As mentioned in previous studies carried out by Geddes (1990) and King et. al. (1994), case selection problems that are a bane to comparative research can be overcome by the use of multilevel models since the causal heterogeneity can be determined.

2.5 Summary

Table 2.5 provides a summary of the research described in this section. Each researcher's cost focus, and techniques for regression modeling and/or factor selection, is indicated (if applicable). Findings relevant to the proposed research are noted.

Table 2.5 Summary Table of Relevant Research

Researcher (Sample Size) [Industry]	Cost Focus			Model		Factor Selection				Findings Significant to Proposed Research
	PE	Right of Way	Construction	Linear	Non-Linear	Forward	Backward	Stepwise	Principal Components	
AbouRizk et al. 2002 (n=213) [Infrastructure]			✓							Accuracy of estimates was determined at 4 stages of projects' life cycle.
Akintoye 2000 (n=84) [Building & Civil]			✓						✓	24 variables reduced to 7 factors. Interpretation of factors emphasized.
Gransberg et al. 2007 (n=31) [Transportation]	✓		✓	✓						Design costs inversely related to construction cost growth. Average construction cost growth = 9.65%. Design costs as percent of construction cost: Roads-2%; Bridges-7.6%
Kyte et al. 2004a, 2004b (n _{roads} =135; n _{bridges} = 23) [Transportation]	✓	✓	✓							PE range 8-20% of construction costs for roads. PE costs inversely related to construction costs. Consultant design costs = 1.5(In-house design costs).
Lam et al. 2008 (n=92) [Design-Build]	✓		✓	✓					✓	47 variables reduced to 12 factors.
Lowe et al. 2006 (n=286) [Buildings]			✓	✓		✓	✓			Log transformation of cost used. Key drivers of cost are design specific.
Nassar et al. 2005 (n=59) [Transportation]	✓			✓				✓		Log transformation of cost used.
Odeck 2003 (n=620) [Transportation]			✓					✓		Cost overruns are inversely related to project cost. Overruns are more predominate in smaller projects.
Trost 1998; Oberlender & Trost 2001. (n=67) [Process Industry]			✓	✓					✓	45 variables reduced to 11 factors. Estimate scoring indicates quality level. Low score = higher quality.
Turochy et al. 2001 (n=9) [Transportation]	✓	✓	✓							PE range 5-20% of construction costs
Wilmont et al. 1999 (n=37) [Transportation]	✓									In-house design costs = 80% (consultant costs).
WSDOT 2002 (n=25) [Transportation]	✓	✓	✓							PE range 4-20% of construction costs. Average = 10.3%. Variability of environmental compliance = 0-10%.

3.0 BRIDGE PROJECTS: PE COST ANALYSES

This section describes the regression modeling strategies we investigated to predict the PE cost ratio for NCDOT bridge projects.

3.1 Database Compilation

The research team obtained project descriptive data, cost estimates, and actual cost expenditures for bridge projects let for construction from NCDOT. The project identification number established in the State Transportation Improvement Plan (STIP) served as the key field linking all data sources and identifying all projects. Preconstruction project data are housed in several independent databases maintained by NCDOT units.

The ten data sources used to populate the bridge project database are listed below.

1. NCDOT Online Bid Tabulations & Annual Bid Averages Summary
2. NCDOT Pre-2002 Project Management Data System (obsolete mainframe system)
3. NCDOT Post-2002 Project Management Data System (SAP based)
4. NCDOT 12-Month Projected Letting List
5. NCDOT National Bridge Inventory System Data (NBIS)
6. NCDOT State Transportation Improvement Plan (STIP)
7. NCDOT Trns·port© Program Modification - Project Type Coding
8. NCDOT Online Construction Plans
9. NCDOT Board of Transportation Minutes and Funding Authorizations
10. North Carolina State Publications Clearinghouse

We accessed NCDOT's online Bid Tabulations and Annual Bid Averages Summary to tabulate data on bridge projects let for construction during January 1, 2001 through June 30, 2009. We queried NCDOT's project management systems to acquire actual PE costs for the bridge projects during this letting period. Projects having complete letting data and PE cost data were considered candidate projects for the bridge database. We used the additional seven data sources to populate the database for each candidate project. NCDOT's data for the National Bridge Inventory System (NBIS) provided values for fourteen independent variables. If projects were missing NBIS data, we removed those projects from the candidate pool.

Table 3.1 lists the 461 bridge projects used in regression analyses organized by calendar year of construction letting. These 461 bridge projects were all included in the NCDOT STIP with a "B" prefix project identifier. NCDOT let additional bridge projects during this same timeframe under bridge purchase order contracts (BPOC) or division design and let (DDL) designations. These projects were not included in the analyses and are not among the 461 projects referenced in Table 3.1.

We selected the ratio of PE costs to estimated STIP construction costs as the preferred response (dependent) variable for cost regression analyses. Using a cost ratio rather than actual cost values allowed modeling across all levels of construction costs and eliminated conversion of cost values to a common base year to account for inflation. Each project's PE costs and estimated STIP construction costs were assumed to be from a similar time period. The ratio of actual PE cost to the estimated STIP

construction cost was tabulated for all 461 bridge projects. This ratio is referred to as the project’s PE cost ratio.

Table 3.1 Bridge Projects Database

Calendar Year of Letting	Number of Bridge Projects	Mean PE Cost Ratio
2001	44	25.6%
2002	62	27.7%
2003	50	25.6%
2004	69	20.9%
2005	48	20.0%
2006	31	9.2%
2007	43	47.1%
2008	98	31.4%
2009 (Jan – Jun)	16	40.3%
Total	461	27.8%

The right most column of Table 3.1 displays the mean PE cost ratio for bridge projects let each calendar year. The total mean PE cost ratio for the 461 projects was 27.8%. The 95th percent confidence interval (CI) for total mean PE cost ratio was 26.0% to 29.6%. The PE cost ratio among the 461 bridge projects was varied, ranging from a minimum of 0.8% to a maximum of 152% of estimated construction costs.

3.2 Validation Sampling

Before regression modeling began, we randomly selected 70 of the 461 bridge projects (15%) to serve as a validation set. The remaining 391 bridge projects comprised the modeling set. We developed regression models using 391 bridge projects and evaluated those models by applying them to the 70 bridge projects comprising the validation set.

3.3 Response Variable for PE Cost Analyses

A project’s PE cost ratio is the ratio of actual PE cost to the estimated STIP construction cost. We calculated this ratio for all 461 bridge projects and investigated the distribution of PE cost ratio values. As evident if Figure 3.1, the PE cost ratio distribution for the 461 bridge projects is left-skewed, exhibiting a non-normal shape. The horizontal axis reflects the range of PE cost ratio values – minimum of 0.008 (0.8%) to a maximum of 1.522 (152%). The vertical axis indicates the number of projects (as a percentage of the total 461 projects) exhibiting a PE cost ratio within a 0.12 range along the x-axis.

To improve normality of the response variable distribution, the project team applied power transformations PE cost ratio values. Normality of the response variable is sought to satisfy multiple linear regression assumptions. We raised the response variable to an exponential power resulting in a transformed variable. By applying the Box-Cox statistical procedure [Sakia 1992] to the non-normal response distribution, the optimal normalized distribution was identified as the cubed root of PE cost ratio, $(PE \text{ cost ratio})^{1/3}$. Figure 3.2 shows the distribution for the transformed response variable, cubed root of PE cost ratio. The distribution for cubed root of PE cost ratio is normal. Subsequent regression analyses used the cubed root of PE cost ratio as the response (dependent) variable.

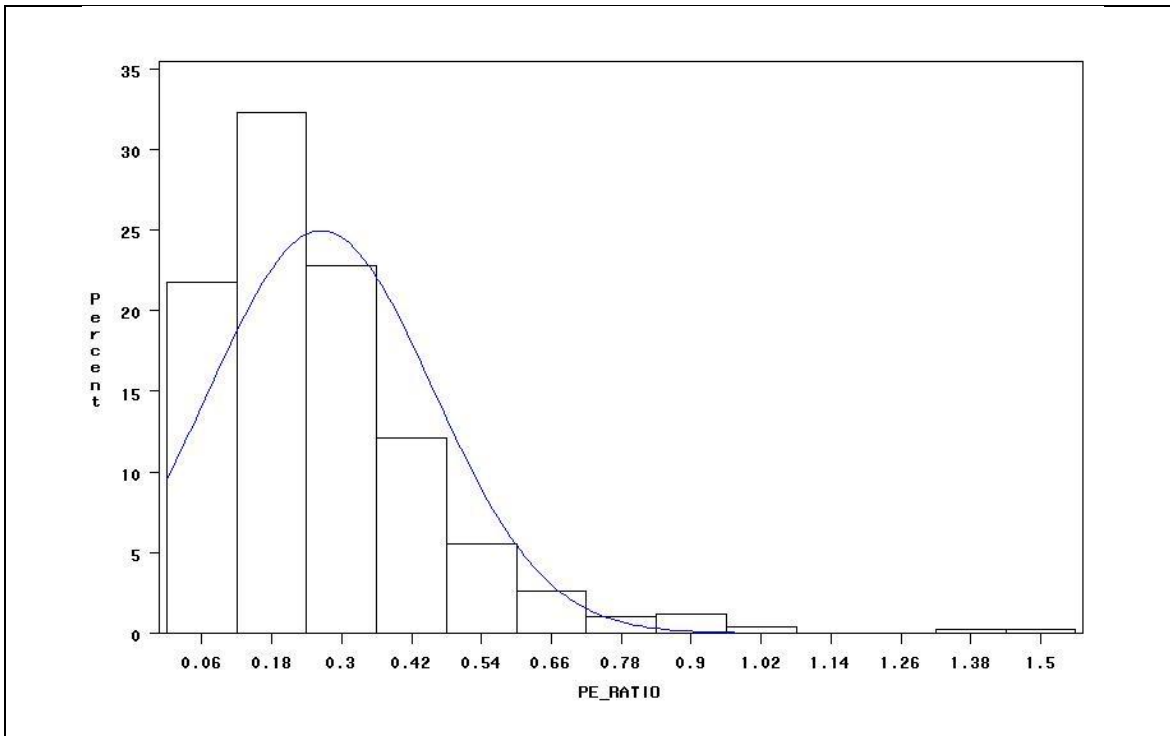


Figure 3.1 Distribution of PE Cost Ratio for Bridge Projects

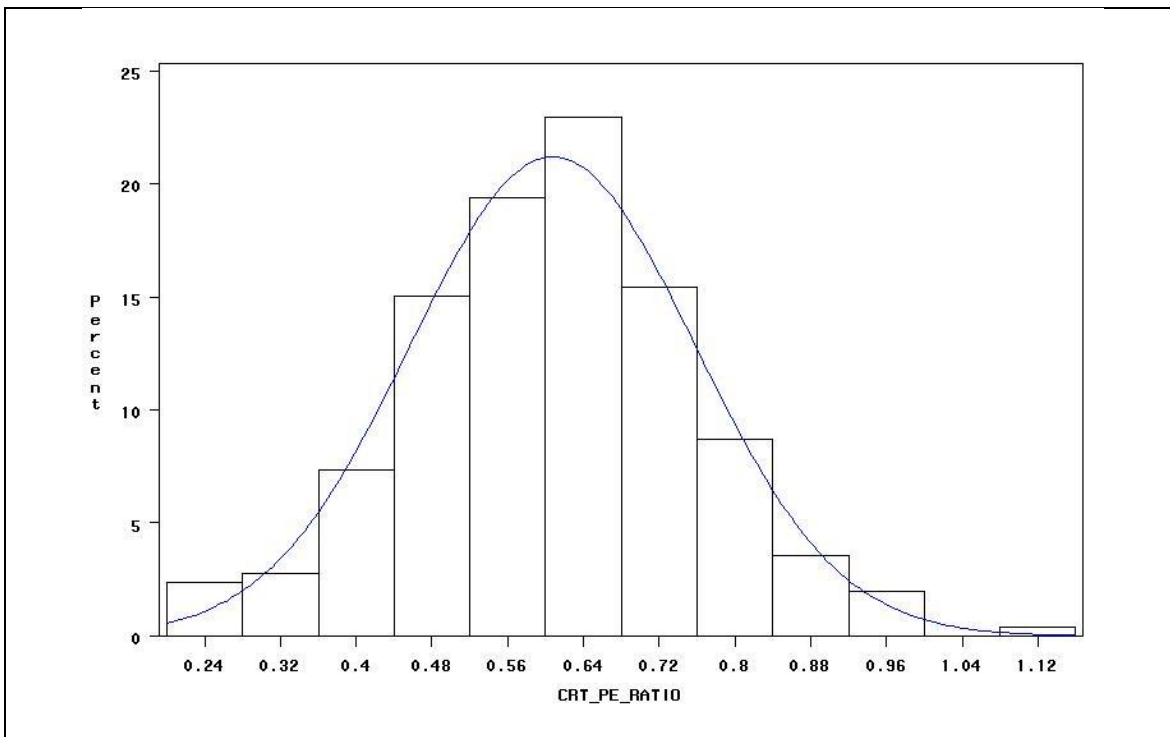


Figure 3.2 Distribution of Cubed Root of PE Cost Ratio for Bridge Projects

Since a power transformation was applied to normalize the response variable (cubed root of PE cost ratio), a back transformation is necessary to report prediction results in terms of the original response variable (PE cost ratio). The back transformation computation for the cubed root transformation is described below [Taylor 1986].

$$\text{Estimated Median Response} = (\text{Predicted Cubed Root of Response})^3$$

$$\text{Transformation Correction Factor} = 1 + \frac{((\text{variance})(1 - \frac{1}{3}))}{2(\text{Predicted Cubed Root of Response})^2}$$

$$\text{Estimated Mean Response} = (\text{Est. Median Response})(\text{Transformation Corr. Factor})$$

Applying this back transformation requires the variance of the predicted response variable. Table 3.2 provides the variance value for converting back to PE cost ratio from the predicted cubed root of this response variable.

Table 3.2 Variance Values for Back Transformation of Response Variables

Response Variable	Modeling Variance
Cubed Root of PE Cost Ratio	0.0229

3.4 Independent Variables for Prediction

We grouped the acquired data for all bridge projects by data function: classification, cost, date, design, dimensional, environmental, and geographical. Correlation and sensitivity analyses followed to statistically assess each candidate variable, resulting in 28 independent variables being identified. These 28 variables describe project-specific parameters. Table 3.3 lists the 28 variables used in model development. Twelve of the 28 variables are numerical. The remaining 16 variables are categorical.

As noted in section 3.1, we relied heavily on information contained in the National Bridge Inventory System (NBIS) database maintained by the Structure Inventory and Assessment Unit to populate the values of fourteen independent variables. All design function variables and five of seven of the dimensional variables were acquired from the NBIS data. This one data source was crucial for obtaining values half of our independent variables.

3.4.1 Variable Sensitivity

For each of the 16 categorical variables, we performed a one-way ANOVA analysis to determine if differences between levels were significant. ANOVA also provided the correlation of determination (R^2) explaining the proportion of the variation in the response variable explained by changes in the independent, categorical variable. This is comparable to simple linear regression between two variables when the levels of all other independent variables are held constant. The R^2 reported by ANOVA was limited to simple effects. Interactions between independent variables were ignored.

Table 3.4 displays the ANOVA analysis results for the 16 categorical variables including R^2 , F-value, and p-values. The p-values (shown in parentheses within the ANOVA results column) were used to identify the categorical variables having statistically significant differences in levels. The seven variables with significant differences in levels are indicated by bullets in the right-most column. Among these seven, the R^2 values were reviewed to determine the level of influence each variable may have on the cubed root of PE cost ratio when all other variables are held constant. The R^2 values range from 0.03 to 0.30. It was surprising that the two variables having the largest R^2 values were date related: year of letting and year of environmental document approval. The other five variables exhibited considerably lower R^2 values.

Table 3.3 Independent Variables for Bridge Projects

Data Function	Independent Variable	Variable Levels or Values	Numerical	Categorical
Classification	Project Construction Scope	Replacement, New Location		■
	Type of Service on Bridge	Highway, Railroad, Pedestrian		■
	Route Signing Prefix	Interstate, US Highway, State Highway		■
	Road System	Arterial, Collector, Local		■
	Structure Type	Bridge or Culvert		■
Cost	ROW Cost to STIP Estimated Construction Cost	Cost Ratio	■	
	Roadway Percentage of Construction Cost	Cost Ratio	■	
	STIP Estimated Construction Cost	Cost in Dollars (\$)	■	
Date	Year of Letting	Calendar Year		■
	Year of Environmental Document Approval	Calendar Year		■
	PE Duration After Environmental Document Approved	Days	■	
Design	Deck Structure Type	Concrete, Steel, Aluminum, Wood		■
	Design Live Load	M9, M13.5, MS13.5, M18		■
	Capacity Rating of Live Load	Metric Tons	■	
	Main Span Structure Type	Concrete, Steel, Wood, Masonry		■
	Design Type	Slab, Girder, Box Beam, Truss		■
Dimensional	Project Length	Miles	■	
	Bypass Detour Length	Kilometers	■	
	Number of Lanes on Bridge	Numerical Count	■	
	Number of Spans in Main Unit	Numerical Count	■	
	Horizontal Clearance for Loads	Meters	■	
	Length of Structure	Meters	■	
	Water Depth	Feet	■	
Environmental	NEPA Document Classification	EIS, EA, CE, PCE, Minimum Criteria		■
	Planning Document Responsible Party	NCDOT or PEF		■
Geographical	NCDOT Division	DIV 01 through DIV 14		■
	Geographical Area of State	Coast, Piedmont, Mountains, Very Mountainous		■
	Classification of Route	Rural or Urban		■

Table 3.4 Categorical Variables: Sensitivity on Cubed Root of PE Cost Ratio

Categorical Independent Variable	ANOVA Simple Effects		
	R ²	F-Value (p-value)	Statistically Significant
Project Construction Scope	0.0322	6.45 (0.0017)	■
Type of Service on Bridge	0.0203	1.13 (0.3420)	
Route Signing Prefix	0.0155	2.03 (0.1090)	
Road System	0.0443	8.80 (0.0002)	■
Structure Type	0.0002	0.06 (0.8054)	
Year of Letting	0.3037	20.83 (<.0001)	■
Year of Environmental Document Approval	0.1220	3.47 (<.0001)	■
Deck Structure Type	0.0225	1.47 (0.1856)	
Design Live Load	0.0302	3.00 (0.0185)	■
Main Span Structure Type	0.0120	0.78 (0.5862)	
Design Type	0.0311	1.22 (0.2767)	
NEPA Document Classification	0.0043	1.66 (0.1982)	
Planning Document Responsible Party	0.0011	0.42 (0.5170)	
Division	0.0728	2.28 (0.0068)	■
Geographical Area of State	0.0361	4.84 (0.0026)	■
Classification of Route	0.0000	0.00 (0.9995)	

The research team anticipated that the year of letting would aid in project identification only. From an investigation of cost trends between 2001 and 2008, we discovered that both STIP estimated construction costs and PE costs exhibited a positively sloped trend line. Figure 3.3 displays this finding; STIP estimated construction costs are graphed on the upper line and PE costs are graphed on the lower. The characteristic of both costs are similar except for the year 2006. PE costs continue to decrease when construction costs begin increasing; then PE costs increase sharply in 2007. The comparative change in PE costs between 2006 and 2007 no longer matches the comparative change in STIP estimated construction costs for the same time period. Unfortunately, discussions with NCDOT personnel did not aid in discovering the cause of this anomaly. No changes in design standards, environmental regulations, or administrative processes could be linked to the evidence. We hypothesized that any longitudinal trend in PE costs would mirror the longitudinal trend in construction costs. Under this assumption, date-based variables with historical levels would not be effective predictor variables during future time periods. There would be no meaningful way to assign past dates (such as year of letting) to represent a future trend. Therefore, the three variables designated in Table 3.3 as date related were rejected as predictor variables.

3.4.2 Variable Correlations

For numerical variables, correlation coefficients indicate the degree of linear association between variables. Correlation coefficients for the 12 numerical independent variables and the response variable are provided in Table 3.5. Correlation coefficients can range from -1 to +1. Larger coefficient values, either positive or negative, indicate a stronger linear association between variables. Coefficient values of zero indicate no linear association between variables. Positive coefficients indicate that the independent variable and response variable move together (positive sloped line); negative coefficients indicate the independent variable and response variable move in opposite directions (negative sloped line).

Table 3.5 reports the Pearson correlation coefficient and accompanying p-value for each numerical variable. Eight statistically significant variables (based on p-values) are indicated by bullet entries in the right-most column. For these eight variables, the coefficient values range from 0.10 to 0.30 indicating a weak linear association with the response variable. The three variables with highest coefficients are project length, STIP estimated construction cost, and right of way cost to STIP estimated construction cost with values of -0.33, -0.31, and +0.31 respectively.

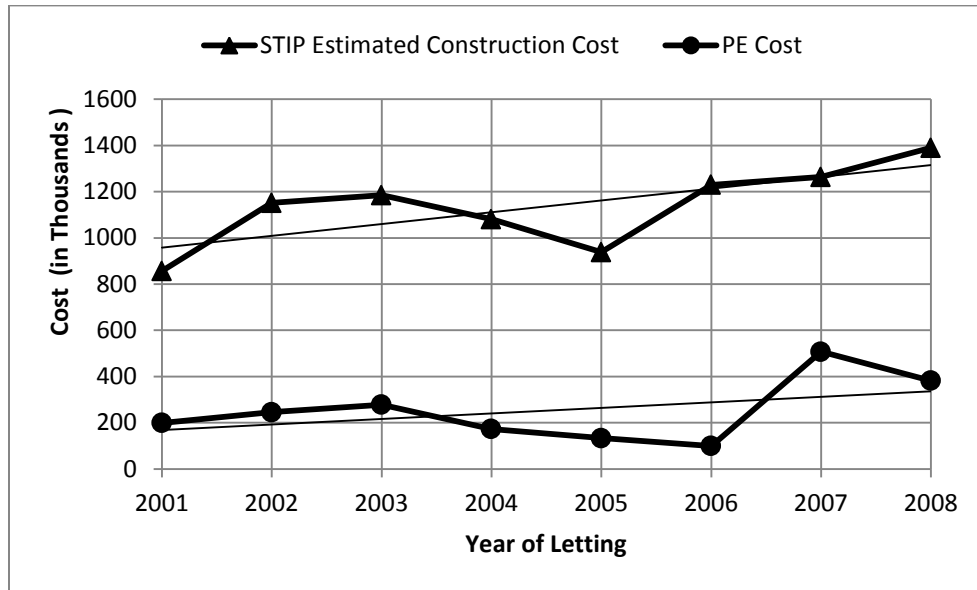


Figure 3.3 Comparison of Cost Trends for Bridge Projects

Table 3.5 Numerical Variables: Correlation with Cubed Root of PE Cost Ratio

Numerical Independent Variable	Pearson Correlation Coefficient		
	Coefficient	(p-value)	Statistically Significant
Lanes on Structure	-0.0545	(0.2431)	
Capacity Rating of Live Load	+0.0486	(0.2974)	
Right of Way Cost to STIP Estimated Construction Cost	+0.3089	(< .0001)	■
Roadway Percentage of Construction Cost	-0.1849	(< .0001)	■
STIP Estimated Construction Cost	-0.3130	(< .0001)	■
PE Duration After Environmental Document Approval	-0.1053	(0.0237)	■
Project Length	-0.3263	(< .0001)	■
Bypass Detour Length	-0.0438	(0.3479)	
Spans in Main Unit	-0.1766	(< .0001)	■
Horizontal Clearance for Loads	-0.1592	(0.0006)	■
Structure Length	-0.1944	(< .0001)	■
Water Depth	-0.0722	(0.1216)	

The project team also determined the correlations between independent numerical variables. This analysis identified the independent variables that were linearly related to each other and would perform the same function in the regression model. For example, lanes on structure and horizontal clearance for loads have a Pearson correlation coefficient of 0.597. Only one of the two variables should be used in a regression model, since there is a moderately high linear association between the two variables. Table 3.5 shows the linear relationship between lanes on structure and the response variable to be statistically insignificant. The correlation coefficient for horizontal clearance for loads and the response variable is significant. Horizontal clearance for loads is the better independent variable for model building.

3.5 Multiple Linear Regression (MLR) Modeling

Selecting the “best” model using multiple linear regression (MLR) can be difficult if there are a large number of independent variables. Common variable selection techniques involve forward-, backward-,

and stepwise-selection methods. To assist in model selection, we utilized the GLMSELECT procedure within the SAS statistical software package. In addition to forward-, backward-, and stepwise-selection, GLMSELECT provides two additional variable selection methods: least angle regression (LAR) and least absolute shrinkage and selector operator (LASSO). The GLMSELECT procedure provides an efficient starting point for model selection. Model refinement can then follow using intuitive insights gained from data familiarity. [Cohen 2006]

Recall from section 3.2 that the bridge database was divided into a modeling set (containing 391 projects) and a validation set (containing 70 projects). Only the modeling set was used for constructing regression models.

We considered both numerical and categorical variables when utilizing the GLMSELECT procedure for variable selection. Initially, all seven categorical variables identified as statistically significant in Table 3.4 were considered. Based on input from NCDOT staff, two environmental variables continued to hold interest and were included even though Table 3.4 reports both as statistically insignificant: NEPA document classification and planning document responsible party.

Similarly, all numerical variables with statistically significant Pearson correlation coefficients (listed in Table 3.5) were included. We compared the adjusted R^2 values associated with the GLMSELECT iterations to assess model fit. The best MLR model achieved an adjusted R^2 of 0.698 utilizing five categorical and three numerical variables with first level interactions. Table 12.1 of the appendices reports the complete regression parameters for the intercept, each significant variable, and each significant interaction for the full model. However, this full model MLR contained the year of letting as a predictor variable. Since we rejected all date-related variables are predictors (discussed in section 3.4.1), the GLMSELECT procedure was repeated with year of letting omitted as a candidate variable. The reduced MLR model selected achieved an Adjusted R^2 of 0.2745 utilizing eight variables: four numerical and four categorical with first level interactions between variables. The selected variables are listed below:

- Right of Way Cost to STIP Estimated Construction Cost
- Roadway Percentage of Construction Cost
- STIP Estimated Construction Cost
- Bypass Detour Length
- Project Construction Scope
- NCDOT Division
- Geographical Area of State
- Planning Document Responsible Party

Table 12.2 in the appendices lists all the regression parameters for the reduced MLR model.

3.6 Hierarchical Linear Model (HLM)

The MLR models described in section 3.5 did select categorical variables for best model fit. A hierarchical modeling approach was investigated to further capitalize on the influence of categorical variables. The HLM utilizes MLR on a selection of projects based on a hierarchy created by shared categorical variable values. The HLM process then is to select the numerical variables for each hierarchical grouping that provides the best model fit. The hierarchical modeling technique used for predicting the cubed root of PE cost ratio for bridge projects is described in this section.

The four categorical variables included in the reduced MLR model are listed below:

- Project Construction Scope
Levels: (3) Replacement (off-site detour), Replacement (on-site detour), New Location

- NCDOT Division
Levels: (14) Division 01 through 14
- Geographical Area of State
Levels: (4) Coast, Piedmont, Mountain, Very Mountainous
- Planning Document Responsible Party
Levels: (2) DOT, Private Engineering Firm (PEF)

Our HLM approach seeks to find the best hierarchy which subdivides the bridge projects into groups based on values for these categorical variables. To quantify the advantage of one scheme over another, we calculated the information gain for competing hierarchies. Information gain is the value of model improvement when adding explanatory variables expressed in universal information units [Shtatland and Barton 1997.] Information gain is calculated by equation below.

$$Information\ Gain = \sum \left(\frac{N_{cell}}{N_{total}} \right) (-\ln(1 - R^2))$$

N_{cell} = number of projects in each subgroup cell

N_{total} = number of total projects

R² = Coefficient of Determination from subgroup MLR modeling

The research team repeated MLR regression modeling for all hierarchical combinations of project construction scope, geographic area of the state, and planning document responsible party. The categorical variable NCDOT division was rejected as a valid hierarchical scheme because it contains 14 levels. The levels for the other three categorical variables numbered between two and four. Additionally, the geographical area of the state is similar to a combined grouping of NCDOT division levels. We determined the information gain at each hierarchical level for the various hierarchies. The organization scheme that maximizes information gain adds greater value to the regression modeling effort. The resulting information gains are reported in Table 3.6.

Table 3.6 Information Gain for Hierarchical Subgroup Combinations

SUBGROUP COMBINATION	INITIAL INFO GAIN	TIER 1 INFO GAIN	TIER 2 INFO GAIN
No Subgroups	0.309		
Geographical Area of State		0.393 CONTROLS	
Geographical Area of State and Project Construction Scope			0.782 CONTROLS
Geographical Area of State and Planning Document Responsible Party			0.560
Project Construction Scope		0.381	
Planning Document Responsible Party		0.347	

We achieved an information gain of 0.393 when using the geographical area of the state as the tier 1 variable. This information gain exceeded the other possible tier 1 combinations [project construction scope (0.381) and planning document responsible party (0.347)]. Information gain values for tier 2 subgroups indicated that project construction scope (0.782) provided greater benefit over planning document responsible party (0.560). We used the remaining categorical variable, planning document responsible party, as the tier 3 subgroup. Our information gain analysis supported using a three tier hierarchy; that hierarchy is presented in Table 3.7.

Table 3.7 Hierarchical Organization of Bridge Projects

Good Fit (13 Cells)
 Poor Fit (6 Cells)
 Not Applicable

Tier 1		Tier 2		Tier 3	
Geographical Area of State	N _{cell}	Project Construction Scope	N _{cell}	Planning Document Responsible Party	N _{cell}
Coast [0.2318] 8NQ1A	77	Off-Site Detour [0.3097] 8NQ1C	58	DOT [0.3786] 8NQ1A	33
				PEF [0.6536] 8NQ1C	25
		On-Site Detour [0.9169] 8NIC	14	DOT [0.8511] various	10
				PEF	4
		New Location [0.7485] all	5	DOT	2
				PEF	3
Mountain [0.3023] 8NQ1D	55	Off-Site Detour [0.2383] 8NIA	23	DOT [0.7293] 8NQ1B	11
				PEF [0.9492] 8NIE	12
		On-Site Detour [0.8841] 5N	9	DOT [0.7431] various	7
				PEF	2
		New Location [0.5103] 8NII	23	DOT [0.9649] 8N	10
				PEF [0.9276] 8NIB	13
Piedmont [0.3148] 8NQ1B	202	Off-Site Detour [0.2498] 8NQ1E	119	DOT [0.2827] 8NIB	84
				PEF [0.4107] 8NIC	35
		On-Site Detour [0.2473] 8NIE	38	DOT [0.4121] 8NIG	17
				PEF [0.3868] 8NQ1A	21
		New Location [0.4802] 8NQ1E	45	DOT [0.7146] 8NQ1E	29
				PEF [0.6999] 8NQ1B	16
Very Mountainous [0.3156] 8NIC	57	Off-Site Detour [0.7391] 8NQ1E	15	DOT [0.7925] various	6
				PEF [0.6264] 5N, 6N, 7N	9
		On-Site Detour [0.9312] 8NQ1B	11	DOT	3
				PEF [0.9335] 5N	8
		New Location [0.2836] 8NQ1B	31	DOT [0.4196] 8NQ1A	16
				PEF [0.6589] 8NII	15
Total Number of Projects	391		391		391

[Adjusted R² values shown in brackets with model identification]

When fitting a MLR model to the hierarchy of Table 3.7, only numerical variables were considered as candidate variables since categorical variables formed the hierarchical tiers. Earlier MLR efforts described in section 3.5 selected four numerical variables for the reduced MLR model. Those four numerical variables are repeated below:

- Right of Way Cost to STIP Estimated Construction Cost
- Roadway Percentage of Construction Cost
- STIP Estimated Construction Cost
- Bypass Detour Length

Three of these four variables had shown significant linear correlation with the response variable, cubed root of PE cost ratio as tabulated in Table 3.5. Only the correlation coefficient of bypass detour length had been insignificant. Though not selected for the reduced MLR model, project length remained of

interest since it is a key project characteristic. The four numerical variables selected for the reduced MLR model, plus project length, were used as candidate variables for MLR modeling applied to the hierarchy represented in Table 3.7. We expanded the pool of candidate variables by considering first level interactions (variable*variable) and the quadratic form (variable²) of the five variables.

The research team applied regression modeling to each subgroup of projects comprising a cell within the hierarchy depicted in Table 3.7. Sixteen models were evaluated and the best fit model determined by comparing adjusted R² values. The simplest tier structure yielding an Adjusted R² value exceeding 0.60 was desired. The Adjusted R² value for each cell is shown in brackets.

- The thirteen cells that achieved an acceptable Adjusted R² value are darkly shaded in Table 3.7. (Adjusted R² values ranged from 0.654 to 0.965)
- Light shading identifies the six cells that did not achieve an acceptable Adjusted R² value. (Adjusted R² values ranged from 0.283 to 0.419)
- Tier 3 cells with no shading were aggregated to the previous tier within the hierarchy.

Table 12.3 of the appendices provides the complete regression parameters for the MLR models fit to the 19 subgroups reflected in Table 3.7 (cells darkly shaded and lightly shaded).

We remodeled the projects contained in each of the six poor fit cell (lightly shaded in Table 3.7) using the regression equations associated with the thirteen cells achieving a good fit. Each project therefore had a pool of thirteen models from which an improved fit was sought. We determined the difference between actual and predicted cubed root of PE ratio (error) for each project using a candidate model, then found the sum of squared errors (SS) for all projects. SS was used to rank candidate models. We selected the model yielding the minimum SS to substitute for the initial model established. The tier 2 model originally fit to the coastal geographical area for new projects minimized SS across all the projects contained in the poor fit cells and therefore was used as a surrogate model for the 6 poor fit cells.

For the 361 projects of the modeling set, the mean cubed root of PE cost ratio was 0.2772. The modeling set mean was used as a second surrogate estimator for all projects within the 6 poor fit cells.

In summary, HLM analyses yielded three modeling strategies:

1. HLM with 19 regression equations (one equation unique for each cell of the hierarchy).
2. HLM with 13 regression equations (one equation for each good fit cell, and one of the thirteen equations reused as a surrogate equation for all poor fit cells).
3. HLM with 13 regression equations (one equation for each good fit cell and the modeling mean used as a surrogate estimate for all poor fit cells).

In section 8, we report each modeling strategy's predictive performance based on our validation process.

4.0 BRIDGE PROJECTS: PE DURATION ANALYSES

The duration of PE plays a vital role in maintaining the costs and schedules of any state infrastructure project. With added pressure owing to a weakened economy, decreased state infrastructure budgets and large spending caps, it is essential that the amount of time and money being spent on the PE phase is accurately estimated prior to authorization. It has been observed that most state infrastructure projects allow for PE to be sanctioned years in advance before actual construction efforts are undertaken. In many cases the PE phase extends much longer than the actual construction duration of the project itself. This leads to difficulties in scheduling construction, allocating funds, and informing the public when the new roadway asset will open.

A majority of the studies carried out in the field of estimation and scheduling have dealt with construction duration. Little to almost no research has been carried out on the estimation of PE duration. In many cases, the duration of PE is not actually planned or estimated prior to the sanctioning of the project. Thus, a detailed study on PE duration will fill a current information gap and provide a great benefit to the NCDOT and its stakeholders.

This study covered 416 NCDOT bridge projects in North Carolina. The period of PE was considered from the date of authorization (initial sanction of funds) until the date of project letting. Right of way acquisition and construction phase durations were not included in this effort.

Four hundred and sixteen projects within the desired range of letting year and with complete data fields were analyzed. The mean PE duration observed was 66.1 months with a 95th percent confidence interval of 64.4 months to 67.8 months. The range of PE duration was extremely large. The minimum duration was 13 months and the largest duration was 164 months.

4.1 Background on PE Duration Estimation

A large amount of research on schedule estimates has involved the construction phase costs and duration. In comparison, the PE phase of infrastructure projects has not been researched. Though there have been selective studies conducted on the PE phase cost aspects, there have been no significant studies carried out on the accurate estimation of PE duration. There have also been many studies on how to reduce or streamline the environmental review process for highway and other infrastructure projects, but those studies have not provided estimation methods and do not cover the whole PE phase.

Extensive research was carried out to find previous literature relevant to PE duration estimation. The research team reviewed several common literature indices and a total of 383 issues from 4 journals in the transportation and construction fields. No literature on or referencing PE duration estimation was found. This literature search highlighted the need for a study on PE duration estimation.

4.2 Database Compilation

The bridge database used to analyze PE duration was the same as for PE cost analyses. Section 3.1 of this report detailed ten data sources for project parameters and descriptive data pertaining to NCDOT bridge projects.

To predict the duration of preliminary engineering (PE) for future projects it was essential that the projects included in the analysis were historically accurate and also reflected the current working procedures and methods of NCDOT. This was achieved by utilizing data acquired from the NCDOT project authorization binders. These binders contain a record of the authorization dates and amounts of preliminary engineering, environmental study, right of way, and utility expenditures. The date of letting was acquired from the bid averages summary database available online through the NCDOT access page. The duration (in months) between authorization and letting, including the design and planning phases, was considered as the period of preliminary engineering and was the focus duration in this research.

Projects were identified on the basis of the State Transformation Improvement Plan (STIP) numbers unique to each project.

The original database consisted of 511 bridge projects. This database consisted of projects let from the year 1999 to 2009. Thirty-four descriptive variables were included in this data base. The duration of PE was determined from the difference (in months) between authorization date and let date. These dates were acquired from the milestone database and were associated with milestone numbers M0005 and M0435 respectively.

From the available variables, four categorical and seven numerical variables were found during initial testing to be statistically significant in predicting PE duration. These variables were retained in the database. Projects that had incomplete data fields were eliminated from the database. For example, projects that returned missing values for the bridge project construction scope were eliminated. A majority of the projects with incomplete data fields were those let in the years 1999 and 2000. For this reason, the projects in the final dataset were those projects that were let between the years 2001 to 2009 and which contained all the data for every variable being analyzed. After removal of these incomplete projects there were 416 projects in the final dataset.

Initial analysis of the data showed that the mean duration of the projects, graphed on the basis of letting year, represented a progressively increasing trend (Figure 4.1). From 2001 to 2007 the trend was a consistently increasing duration. In 2008 the mean PE duration dropped but then it increased again in 2009. There is no way to know, based on these data, whether the 2008 and 2009 data represent a leveling of the mean PE duration or whether 2008 was a temporary interruption of a longer term trend that will resume in 2010 and afterward. For the analysis, Figure 4.1 leads to questions on whether there is an important difference in the nature of the projects in those two ranges (2001 to 2006 vs. 2007 to 2009). This also led to the question of the selection of the appropriate dataset for analysis and ultimately for the prediction of PE duration in the future. That is, will future bridge projects behave more like projects from 2001 to 2006 or from 2007 to 2009?

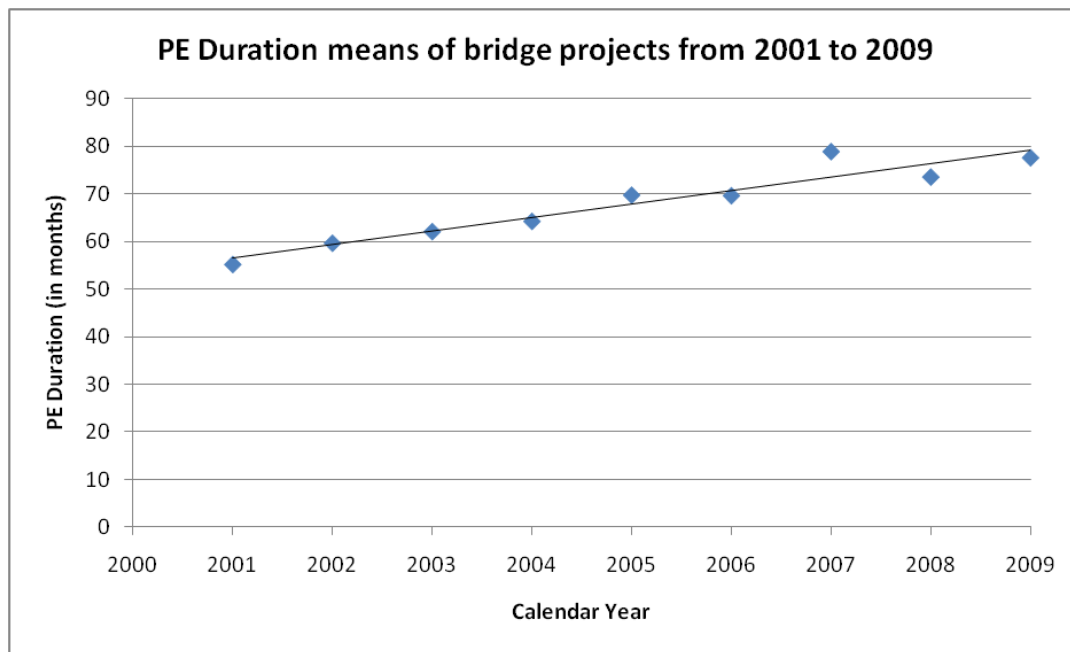


Figure 4.1 Mean PE Duration versus Year of Letting

In the end, the research team decided to select all the projects from 2001 to 2009 as the model dataset. This decision was made on the basis that the inclusion of all the projects would provide a more thorough

range of projects and would encompass both historical data as well as more recent data that better reflected the present working methods of the NCDOT. This selection also allowed for the largest possible sample size to be made available for use in multilevel modeling. An alternative to selecting the entire range of projects would have been to analyze each range individually, determine the dissimilarities amongst the projects of each range, and then arrive at conclusive results for each case. Using this alternative, the accuracy of the results using the adopted procedure would be reduced. The usefulness of the end result would also be compromised since it would involve an analyst having to select a prediction equation to predict PE duration of future projects.

4.3 Validation Sampling

To assess the effectiveness of the regression model we needed to create a separate validation data set. This validation set was created by randomly selecting a portion of projects from the overall dataset of 416 projects. Sixty projects (approximately 15%) were selected and separated from the remainder of the data set. These 60 projects were not involved in the estimation of the prediction equation. The remaining projects (N=356) were used as the modeling set. In practice, at most half of the data (and usually less) are so reserved, and estimates based on splitting have comparatively high variability (Picard & Cook 1984).

4.4 Independent Variations for Prediction

Tables 4.1 and 4.2 show the scope and range of the independent variables used to calibrate the prediction equation (356 projects altogether). These tables provide a description of the PE duration range for projects based on location, the agency responsible for planning documents, the construction type of the projects, and the classification of the environmental documentation.

Table 4.1 Scope of Categorical Independent Variables used in PE Duration Estimation

Categorical Variable	Presence in Dataset	Range of Preliminary Engineering Duration		
		Minimum Duration (months)	Mean Duration (months)	Maximum Duration (months)
Geographical Area of State	N=356			
• Coast	n=65	30	66	131
• Mountains	n=52	41	67	130
• Very mountainous	n=53	49	70	131
• Piedmont	n=186	13	66	164
NEPA Document Classification	N=356			
• Categorical Exclusion(CE)	n=263	13	69	164
• Programmable CE	n=93	42	61	130
Planning Document Responsible Party	N=356			
• DOT	n=203	13	64	130
• PEF	n=153	28	71	164
Project Construction Scope	N=356			
• Replacement with on-site detour (B_EX_ON)	n=69	28	67	164
• Replacement with off-site detour (B_EX_OFF)	n=192	13	65	131
• New location(B_NEW))	n=95	47	69	131

Table 4.2 Scope of Numerical Independent Variables used in PE Duration Estimation

Independent Variables	Percentage of total projects with variable	Variable Value Range		
		Minimum Value	Mean Value	Maximum Value
Right of way ratio (cost ratio)	97%	0.00004	0.096	0.8504
TIP Cost (\$)	100%	225,000	1,170,000	14,500,000
Project Length (miles)	100%	0.038	0.212	0.663
Number of spans in main unit (count)	100%	1	2.7	31
Roadway percentage of construction cost	100%	0.127	0.373	0.728

4.4.1 Variable Sensitivity

Analysis of variance (ANOVA) was carried out on categorical variables prior to performing a multilevel regression analysis to determine if the differences in the levels of the variables were statistically significant. ANOVA also provided the value of the coefficient of determination (R^2). The R^2 values provided an explanation of the proportion of variation in the dependent variable due to a change in the independent variable. This analysis was restricted to simple effect changes and does not include complex interaction or quadratic variable variation. Based on the results of this analysis it was observed that eight of the fifteen categorical variables displayed a statistically significant effect on the prediction of PE duration. Table 4.3 shows the results of this ANOVA.

Table 4.3 One-Way ANOVA Results

Categorical Independent Variable	ANOVA Simple Effects			
	R^2	F-Value (p-value)	Statistically Significant	Statistically significant (individual level assessment)
Project Construction Scope	0.0064	1.10 (0.3331)		■
Type of Service on Bridge	0.0064	0.31 (0.9485)		
Route Signing Prefix	0.0075	0.86 (0.4625)		
Road System	0.0036	0.63 (0.5317)		
Year of Letting	0.1496	7.39 (<.0001)	■	■
Year of Environmental Document Approval	0.7409	4.22 (<.0001)	■	■
Deck Structure Type	0.0515	3.68 (0.0029)	■	■
Design Live Load	0.0732	6.72 (<.0001)	■	■
Main Span Structure Type	0.0162	1.12 (0.3495)		
Design Type	0.0886	3.25 (0.0005)	■	
NEPA Document Classification	0.0418	14.97 (0.0001)	■	■
Planning Document Responsible Party	0.0374	13.35 (0.0003)	■	■
Division	0.0647	1.76 (0.0476)	■	■
Geographical Area of State	0.0069	0.79 (0.4994)		■
Classification of Route	0.0042	1.45 (0.2292)		

A second iteration of ANOVA was carried out on the variables with the levels treated individually. For example, ANOVA was carried out on the variable “geographic area of state” with each of the four levels assessed independently. It was observed that 3 of the 4 levels proved to be significant when tested individually as compared to when they were analyzed as a whole. This procedure was adopted on the variables project construction scope, division, NEPA document classification, and planning document responsible party. Two variables, geographical area of state and project construction scope, were now found to show statistically significant differences in levels.

The year of letting was observed to show the highest value R^2 with a statistically significant difference in levels but was excluded from consideration as a variable in the prediction equation. Including letting year

as a prediction variable would compromise the validity of the prediction equation as a tool to predict the PE duration of future projects. Of the remaining significant categorical variables, four variables were considered to have a major impact on the duration of PE for a project. These variables are listed below:

- Geographical Area of State
- Planning Document Responsible Party
- NEPA Document Classification
- Project Construction Scope

These four categorical variables were later used as the four main tiers of the multilevel model.

4.4.2 Variable Correlations

A correlation coefficient between an independent variable and a response variable indicates the presence of a linear association between the two. Correlation coefficients, ranging from +1 to -1, depict the nature of the linear relation between the variables. If the value of the coefficient is negative, it is indicative that a positive change in one variable will result in a negative change in the other. If the value of the coefficient is positive, it indicates that both variables move in the same direction. Based on this correlation coefficient, the slope of the regression line will be positive or negative. The absence of a linear relation is represented by a correlation coefficient near zero.

Table 4.4 shows the values of the correlation coefficients between the numerical independent variables and the response variable PE duration. Four of the 11 numerical variables show the strongest correlations to PE duration. The strongest four variables are Roadway Percentage of Construction Cost, STIP Estimated Construction Cost, Bypass Detour Length, and Water Depth.

The correlation amongst the independent variables was analyzed to determine the degree of collinearity between the predictor variables. Collinearity is generally agreed to be present if there is an approximate linear relationship among some predictor variables in the data [Mason and Perreault Jr. 1991]. Since variables that were linearly related to each other would essentially perform the same function in the regression model, the variables that exhibited a strong relationship with another predictor variable were weighed on the basis of their relationship with the response variable. The predictor variable with the weaker relationship with the response was then eliminated from selection in the prediction equation. The independent variables Roadway Portion of Project Cost and Capacity Rating of Live Load were found to have a correlation coefficient of 0.591. The variable for the Roadway Portion of the Project Cost had a stronger linear relationship with the response variable and was considered as the better independent variable for building the model.

Table 4.4 Correlation Coefficient Values

Numerical Independent Variable	Pearson Correlation Coefficient
Lanes on Structure	-0.0731
Capacity Rating of Live Load	-0.0989
Right of Way Cost to STIP Estimated Construction Cost	0.0280
Roadway Percentage of Construction Cost	-0.1327
STIP Estimated Construction Cost	0.2574
Project Length	0.0271
Bypass Detour Length	0.1059
Spans in Main Unit	0.0161
Horizontal Clearance for Loads	0.0111
Structure Length	0.0736
Water Depth	0.1120

4.5 Hierarchical Linear Model (HLM)

From the 34 descriptive bridge project variables, 24 independent variables and one response variable were identified for regression modeling. Seven of these 24 variables were categorical independent variables. Attempts at multiple linear regression did not produce credible results from these data. To facilitate the creation of a nested design model, four of the categorical independent variables were used to create a multilevel hierarchical structure. These four variables were not included in the regression model as categorical variables but were used to group projects on the basis of those specific characteristics.

Multilevel hierarchical regression allows analysts to study the effect of one characteristic on another characteristic in another tier of the structure or hierarchy. The use of hierarchical regression techniques allows the creation of customized equations for each subgroup of the overall model.

Project subgroups were created in a hierarchical tier scheme using one of four categorical variables for each tier. The tier structure along with the levels for each was:

- Tier 1 Geographical Area of State (GEO_AREA)
Levels: (4) Coast, Mountain, Piedmont, Very Mountainous
- Tier 2 NEPA Document Classification (NEPA_DOC)
Levels: (2) CE, PCE
- Tier 3 Bridge Project Construction Scope (B_SCOPE)
Levels: (3) Replacement (on-site detour), Replacement (off-site detour), New Location
- Tier 4 Planning Document Responsible Party (PLAN_RESP)
Levels: (2) DOT, Private Engineering Firm (PEF)

Table 4.5 shows that regression modeling was applied to each cell in this hierarchical structure.

4.5.1 Tier Selection using Information Gain Theory

Four categorical variables were grouped into four tiers. The initial tier grouping order was geographic area, project scope, environmental document type, and planning responsibility. This grouping was chosen by estimation of the value of each parameter in predicting duration. To verify the statistical merit of this selection, information gain theory was used to quantify the information gain achieved by using a categorical variable at each tier level. The variable grouping that maximized information gain was selected. Information gain was computed for each multilevel cell using Equation 1. The cumulative information gain is calculated by summing across all cells used in the multilevel structure.

$$\sum \left[\frac{N_{cell}}{N_{total}} \right] * [-LN(1 - R^2)] \quad \text{Equation 1}$$

Where:

- N_{cell} = total number of projects in the cell
- N_{total} = total number of projects in the entire model
- $-LN$ = negative natural log

Table 4.6 shows the result of the information gain procedure performed on the four categorical variables. The R^2 values were used to determine the value of information gain. The table shows GEO_AREA as the variable showing the greatest information gain, greater than the value when there is no tier order as well greater than the information gain number of the other three variables. GEO_AREA is therefore the ideal candidate for tier 1. B_SCOPE and NEPA_DOC were the third and second highest values after GEO_AREA and were investigated for the tier 2 position. Since the combination of GEO_AREA & NEPA_DOC showed a higher information gain we accepted NEPA_DOC as the tier 2 variable, with B_SCOPE as the tier 3 variable and PLAN_RESP in tier 4.

Table 4.5 Hierarchical Level Table for PE Duration

TIER 1		TIER 2		TIER3		TIER4	
GEO_AREA	N	NEPA_DOC	N	B_SCOPE	N	PLAN_RESP	N
COAST (0.6092)	64	CE (0.3108)	45	B_EX_OFF (0.5403)	30	DOT (0.8323)	6
						PEF (0.6116)	24
				B_EX_ON (0.8770)	12	DOT (0.9328)	8
						PEF (0.8447)	4
				B_NEW (0.9986)	3	DOT (N/A)	2
				PEF (N/A)	1		
		PCE (0.1274)	19	B_EX_OFF (0.4259)	17	DOT (0.4259)	17
						PEF (N/A)	--
				B_EX_ON (N/A)	02	DOT (N/A)	02
						PEF (N/A)	--
B_NEW (N/A)	--			DOT (N/A)	--		
		PEF (N/A)	--				
MTN (0.3957)	52	CE (0.2338)	43	B_EX_OFF (0.7492)	12	DOT (N/A)	1
						PEF (0.8728)	11
				B_EX_ON (0.8741)	9	DOT (0.8233)	7
						PEF	2
		B_NEW (0.5233)	22	DOT (0.8635)	10		
				PEF (0.7338)	12		
		PCE (0.8812)	9	B_EX_OFF (0.8637)	9	DOT (0.8812)	9
						PEF (N/A)	--
				B_EX_ON (N/A)	-	DOT (N/A)	--
				PEF (N/A)	--		
B_NEW (N/A)	-			DOT (N/A)	--		
		PEF (N/A)	--				
PDMT (0.2361)	182	CE (0.1912)	121	B_EX_OFF (0.1008)	49	DOT (0.2791)	19
						PEF (0.1480)	30
				B_EX_ON (0.6125)	36	DOT (0.1927)	16
						PEF (0.8082)	20
		B_NEW (0.4560)	36	DOT (0.6166)	22		
				PEF (0.6712)	14		
		PCE (0.1313)	61	B_EX_OFF (0.1150)	57	DOT (0.1150)	57
						PEF (N/A)	--
B_EX_ON (N/A)	1			DOT (N/A)	1		
				PEF (N/A)	--		
B_NEW (0.9942)	3	DOT (0.6110)	2				
		PEF (N/A)	1				
V MTN (0.4861)	47	CE (0.2610)	44	B_EX_OFF (0.8695)	11	DOT (0.9973)	4
						PEF (0.3769)	7
				B_EX_ON (0.9062)	9	DOT (N/A)	1
						PEF (0.4510)	8
				B_NEW (0.3819)	24	DOT (0.8557)	13
				PEF (0.3475)	11		
		PCE (0.8291)	3	B_EX_OFF (0.9383)	3	DOT (N/A)	2
						PEF (N/A)	1
				B_EX_ON (N/A)	--	DOT (N/A)	--
						PEF (N/A)	--
B_NEW (N/A)	--			DOT (N/A)	--		
		PEF (N/A)	--				
Total Projects	345		345		345		345

Table 4.6 Information Gain Results for Tier Classification

SUBGROUP COMBINATION	INITIAL INFO GAIN	TIER 1 INFO GAIN	TIER 2 INFO GAIN
NONE	0.310		
GEO_AREA		0.358 CONTROLS	
GEO_AREA & B_SCOPE			0.90
GEO_AREA & NEPA_DOC			0.97 CONTROLS
B_SCOPE		0.232	
NEPA_DOC		0.240	
PLAN_RESP		0.218	

Multiple linear regression was applied to each cell (or sub group) in Table 4.5 and yielded a corresponding adjusted R^2 . This is indicated by the values in parenthesis in Table 4.5. These R^2 values, along with the values for absolute error, were used to determine the best fit model amongst all the models. A base R^2 value of 0.6 was used to determine the good fit cells and the poor fit cells. All projects that achieved or surpassed the base value were classified as good fit cells. All cells that failed to achieve the base value were classified as poor fit cells.

Table 4.5 shows the different classification of each cell. The 12 good fit cells are highlighted in blue, the eight poor fit cells are highlighted in yellow, and all the cells with hatchings are considered “not applicable.”

PE durations of the 356 Bridge projects from the modeling set were used to calibrate the prediction equation during the multilevel modeling effort described above. Regression procedures carried out in the multilevel modeling procedure followed all the basic assumptions of regression and involved the use of four categorical variables, seven numerical variables, three categorical variables which were assigned numerical values, and ten quadratic and interaction numerical variables. The adjusted R^2 values of the regression equations created for each combination of projects were used to rank the predictive ability of each. The Mallows Cp value was used as a determinant of fit and equations with missing Cp values were rejected. Simulations performed using this prediction equation returned an average PE duration of 65 months. The 95th percent confidence interval for this mean PE duration was 63.5 months to 66.5 months.

From the eight poor fit cells, the two cells populated by the largest number of projects were remodeled using the equations from the good fit cells. Since there were 12 cells that exhibited a good fit, there were 12 possible equations for each of the poor fit cells. The error in duration was calculated by subtracting the predicted from the actual PE duration. The sum of squared error (SS) was calculated and on the basis of this SS the models were ranked. The model equation with the lowest SS value was considered as the best substitute equation. The equation from the cell for COAST was observed to have the lowest SS error in each poor fit cell and was thus identified as the best substitute equation.

Table 4.7 shows the 12 prediction equations for the good fit cells of the HLM as shown in Table 4.5. As noted earlier, the equation for COAST applies for the “bad fit” cells in Table 4.5.

The prediction equations were then tested on a validation sample of 60 projects separated from the modeling database. Results of the validation process are reported in section 8 of this report.

Table 4.7 Regression Equations from HLM Procedure

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 4.5					
		Coast	MTN PCE	V MTN PCE	Mountain CE B_EX_ON	Mountain CE B_EX_OFF	Piedmont CE B_EX_ON
Intercept		52.67	173.92	76.8	627.14	91.78	82.70
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO	-35.63	-34.911		194.40	-304.012	
STIP Estimated Construction Cost	TIP_COST	0.0000165	-5.1E-05	-1.7E-05	-0.00071	-1.5E-05	
Project Length	LEN	-56.918			1706.2	-156.54	
Quadratic Form – Project Length	len22		-1078.83				
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN	4.0033	5.19		50.74		
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M						
Design Type	n43b		-1.66		-2.08		
Roadway Percentage of Construction Cost	RW		135.35		-1224.93		-59.04
Quadratic form - STIP Estimated Construction Cost	tip22	-9.5E-13			1.8E-10		
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22						1243.14
Interaction – Right of Way with STIP estimated construction cost	row2tip					0.000276	

Table 4.7 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 4.5					
		Piedmont PCE B_NEW	V Mountain CE B_EX_OFF	V Mountain CE B_EX_ON	Piedmont CE B_NEW DOT	Piedmont CE B_NEW PEF	V Mountain CE B_NEW DOT
Intercept		52.06	138.71	95.63	60.18	123.83	-5.45
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO			-79.34			-25.42
STIP Estimated Construction Cost	TIP_COST	9.71E-06	-7.2E-05	-4.6E-05	-1.2E-05		9.5E-05
Project Length	LEN		-187.58	-6047.54		-309.25	
Quadratic Form – Project Length	len2					384.68	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN				-3.51		-3.74
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M			14.65	0.27		1.33
Design Type	n43b			40.90	-0.36	-0.7306	1.83
Roadway Percentage of Construction Cost	RW		-240.74	-591.12			
Interaction - STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw		0.00038		4.22E-05		
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw			9567.2			
Interaction - STIP Estimated Construction Cost with Project Length	tip2len						-0.00021

5.0 ROADWAY PROJECTS: PE COST AND PE DURATION ANALYSES

Roadway projects were analyzed using similar processes and procedures to the bridge projects previously described in Sections 3 and 4 of this report. The following section describes the specifics of the roadway analyses for both PE costs and PE duration.

5.1 Database Compilation

Previously, we used ten data sources to populate the bridge projects database. Those sources are listed below. The fifth source listed, NCDOT National Bridge Inventory System Data (NBIS), applies specifically to bridge projects. We utilized the remaining nine sources to populate the roadway projects database.

11. NCDOT Online Bid Tabulations & Annual Bid Averages Summary
12. NCDOT Pre-2002 Project Management Data System (obsolete mainframe system)
13. NCDOT Post-2002 Project Management Data System (SAP based)
14. NCDOT 12-Month Projected Letting List
15. **NCDOT National Bridge Inventory System Data (NBIS) [not used for roadway projects]**
16. NCDOT State Transportation Improvement Plan (STIP)
17. NCDOT Trns·port© Program Modification - Project Type Coding
18. NCDOT Online Construction Plans
19. NCDOT Board of Transportation Minutes and Funding Authorizations
20. North Carolina State Publications Clearinghouse

For roadway projects let between January 1, 1999 and June 30, 2009, we acquired actual PE costs for 188 projects. Table 5.1 summarizes these projects by STIP prefix type (Interstate, Rural, and Urban) and provides a breakdown by let year. These 188 projects form our roadway database and were used in the PE cost analyses for roadway projects.

Table 5.1 Roadway Projects Database

Let Year	Dataset Projects by STIP Prefix Type			Total Projects
	Interstate (I)	Rural (R)	Urban (U)	
1999	4	3	13	20
2000	4	2	8	14
2001	4	5	4	13
2002	4	13	7	24
2003	2	12	8	22
2004	7	7	10	24
2005	2	7	4	13
2006	5	2	4	11
2007	2	7	10	19
2008	1	6	10	17
2009 (Jan – Jun)	1	4	6	11
Total by Type	36	68	84	188

Unfortunately, we were unable to acquire the PE authorization date for all 188 roadway projects. Each project’s PE authorization date is required to determine an actual PE duration. Therefore, we analyzed a reduced dataset, consisting of 113 of the 188 roadway projects, for which a duration was known, when studying PE duration. Table 5.2 compares the datasets used for PE cost analyses and PE duration analyses by project type.

Table 5.2 Roadway Dataset Differences for PE Cost and PE Duration Analyses

Goal of Analysis	Dataset Projects by STIP Prefix Type			Total Projects
	Interstate (I)	Rural (R)	Urban (U)	
PE Cost Prediction	36	68	84	188
PE Duration Prediction	13	54	46	113

5.2 Validation Sampling

We reserved a portion of the roadway project dataset to use for validation purposes. We randomly selected of 23 projects comprising 20 percent of the 113 projects in the PE duration dataset. These 23 projects were validation projects for both PE cost and PE duration analyses. Since the PE cost dataset included an additional 75 roadway projects (that were not part of the PE duration dataset), 15 of these additional 75 projects were randomly selected for use in validating the PE cost analyses.

PE cost analyses used 150 roadway projects for modeling and 38 projects for validation purposes.

PE duration analyses used 90 (subsample of the same 150) roadway projects for modeling and 23 (subsample of the same 38) projects for validation purposes.

5.3 Response Variables: PE Cost Ratio and PE Duration

Regression modeling seeks to define a relationship between roadway project parameters and project PE cost ratio and PE duration. Regression analyses for PE cost ratio and PE duration were performed independently. Handling of the response variables for each analysis is described in this section.

For future prediction capability, neither response variable was considered as a valid regressor for the other. However, our analyses did find a significant correlation between PE cost ratio and PE duration. When the cubed root of PE cost ratio was regressed on PE duration, the resulting R^2 value is 0.1604 (F-value 16.24; p-value 0.0001). The Pearson correlation coefficient was +0.4160 (p-value < .0001).

5.3.1 PE Cost Ratio

Consistent with the bridge project analyses described in Section 3, we continued to use the ratio of PE costs to estimated STIP construction costs as the preferred response (dependent) variable. The ratio of actual PE cost to the estimated STIP construction cost was tabulated for all 188 roadway projects. This ratio is referred to as the project’s PE cost ratio.

Our review of total 188 NCDOT roadway projects identified a mean PE cost ratio of 11.7%. The 95th percent confidence interval (CI) for mean PE cost ratio was 9.7% to 13.7%. The PE cost ratio among roadway projects ranged from a minimum of 0.01% to a maximum of 129% of estimated construction costs. The distribution of PE cost ratio values for the roadway database is left-skewed, exhibiting a non-normal shape.

To improve normality of the response variable distribution, the project team applied power transformations using the Box-Cox statistical procedure [Sakia 1992] to the roadway projects' PE ratio cost values. We found the optimal, normalized distribution to be the cubed root of PE cost ratio, $(PE \text{ cost ratio})^{1/3}$, which was consistent with our findings for bridge projects. Figure 5.1 shows the distribution of the transformed response variable - cubed root of PE cost ratio. The distribution of cubed root of PE cost ratio for roadway projects is normal. Subsequent regression analyses use the cubed root of PE ratio as the response (dependent) variable.

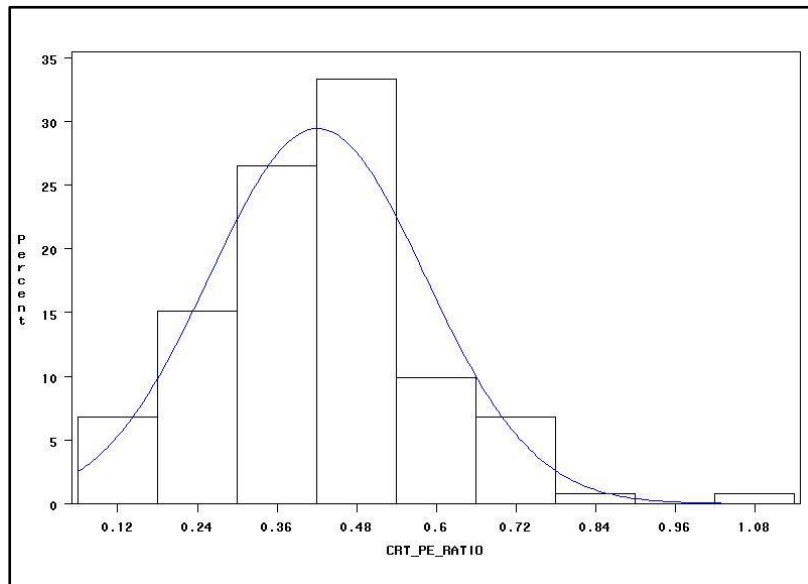


Figure 5.1 Distribution of Transformed Response Variable for PE Cost Analyses

5.3.2 PE Duration

Our review of the 113 roadway projects (for which duration data was available) identified the mean PE duration as 55.7 months with a 95th percent CI of 48.3 months to 63.1 months. The range of PE duration was extremely large; the minimum PE duration was 1 month, and the maximum duration was 163 months.

We identified the distribution of PE duration values as left-skewed and non-normal in shape. Through application of the Box-Cox procedure, we found the optimal transformation for normalizing PE duration to be the cubed root of PE duration $(PE \text{ duration})^{1/3}$. Similar to PE cost analyses; subsequent regression efforts use the cubed root of PE duration as the response variable.

5.3.3 Back Transformation of Response Variables

Since a power transformation was applied to normalize both response variables (PE cost ratio and PE duration), a back transformation is necessary to report prediction results in terms of the original response variable. The back transformation computation for the cubed root transformation is described below [Taylor 1986].

$$\text{Estimated Median Response} = (\text{Predicted Cubed Root of Response})^3$$

$$\text{Transformation Correction Factor} = 1 + \frac{\left((\text{variance}) \left(1 - \frac{1}{3} \right) \right)}{2(\text{Predicted Cubed Root of Response})^2}$$

$$\text{Estimated Mean Response} = (\text{Est. Median Response})(\text{Transformation Corr. Factor})$$

Applying this back transformation requires the variance of the predicted response variable. Table 5.3 provides the variance values for converting back to PE cost ratio or PE duration from the predicted cubed root of these response variables.

Table 5.3 Variance Values for Back Transformation of Response Variables

Response Variable	Modeling Variance
Cubed Root of PE Cost Ratio	0.0265
Cubed Root of PE Duration	1.1077

5.4 Independent Variables for Prediction

The roadway project data were grouped by function: classification, cost, dimensional, environmental, and geographical. Through correlation and sensitivity analyses, 16 project-specific parameters were identified as candidate independent variables for regression modeling. Table 5.4 lists the 16 candidate independent variables for roadway analyses. The candidate variables consist of ten numerical variables and six categorical variables.

When compared with bridge project data, one functional area was unavailable for roadway projects – design data. Design specific parameters could not be easily acquired for all 188 roadway projects from the electronic data sources investigated. Design specific parameters for bridge projects were acquired from the National Bridge Inventory System (NBIS) data gathered as a federal requirement for bridge inspection. No comparable inventory for roadway projects was available.

Table 5.4 Independent Variables for Roadway Projects

Category	Independent Variable	Variable Levels or Values	Numerical	Categorical
Classification	Project Construction Scope	New Location, Widening, Rehabilitation & Resurfacing, Interchange		■
	STIP Prefix	Interstate (I); Rural (R); Urban (U)		■
	Federal Funding Utilized	Yes (1) or No (0)	■	
Cost	Estimated Right of Way Cost	Cost in Dollars (\$)	■	
	Right of Way Cost to STIP Estimated Construction Cost	Cost Ratio	■	
	Roadway Percentage of Construction Cost	Cost Ratio	■	
	Structure Percentage of Construction Cost	Cost Ratio	■	
	STIP Estimated Construction Cost	Cost in Dollars (\$)	■	
Dimensional	Project Length	Miles	■	
	Number of Lanes	Numerical Count	■	
	Length of Structures within Project	Miles	■	
Environmental	NEPA Document Classification	EIS, EA, CE, PCE, State Min Criteria		■
	Planning Document Responsible Party	NCDOT or PEF		■
Geographical	NCDOT Division	DIV 01 through DIV 14		■
	Geographical Area of State	Coast, Piedmont, Mountains, Very Mountainous		■
	Metropolitan Area Designation	Metropolitan (1) or Nonmetropolitan (0)	■	

5.4.1 Variable Sensitivity and Correlations with PE Cost Ratio Response

For each of the categorical variable, we used one-way ANOVA analysis to assess if differences between levels were significant. Table 5.5 summarizes the ANOVA results for the categorical variables. Of the six categorical variables identified for roadway projects, two exhibited statistically significant differences in levels. R^2 values quantify the proportion of variation in the response variable explained by changes in each categorical variable. Using the simple effect R^2 values as rankings, the most influential categorical variables on cubed root of PE cost ratio are project construction scope and STIP prefix.

Table 5.5 Categorical Variables: Sensitivity on Cubed Root of PE Cost Ratio

Categorical Independent Variable	ANOVA Simple Effects		
	R^2	F-Value (p-value)	Statistically Significant
Project Construction Scope	0.1971	7.22 (<.0001)	■
STIP Prefix	0.0903	7.45 (0.0008)	■
NEPA Document Classification	0.0433	1.67 (0.1593)	
Planning Document Responsible Party	0.0002	0.03 (0.8660)	
NCDOT Division	0.1041	1.24 (0.2563)	
Geographical Area of State	0.0245	1.25 (0.2947)	

Table 5.6 provides the correlation coefficients for each of the ten numerical independent variables and the cubed root of PE cost ratio (response variable). A larger coefficient value indicates a stronger linear association between the independent and response variable. Five variables exhibited a statistically significant coefficient. These five variables are identified by bullet indicators in the far right column of Table 5.6. Although the correlations are statistically significant, coefficient values ranging from 0.1666 to 0.3959 (ignoring \pm signs) indicated only a weak linear association with the response variable. Linear association strength increases as coefficient values approach +1 or -1. These correlations were confirmed by reviewing scatter plots. The scatter plots did not visually exhibit any linear associations.

Table 5.6 Numerical Variables: Correlation with Cubed Root of PE Cost Ratio

Numerical Independent Variable	Pearson Correlation Coefficient		
	Coefficient	(p-value)	Statistically Significant
Federal Funding Utilized	+0.2706	(0.0002)	■
Estimated Right of Way Cost	+0.1213	(0.0991)	
Right of Way Cost to STIP Estimated Construction Cost	+0.3432	(< .0001)	■
Roadway Percentage of Construction Cost	-0.2375	(0.0010)	■
Structure Percentage of Construction Cost	+0.0735	(0.3160)	
STIP Estimated Construction Cost	-0.1090	(0.1365)	
Project Length	-0.3959	(< .0001)	■
Number of Lanes	-0.1666	(0.0445)	■
Length of Structures within Project	-0.0653	(0.3732)	
Metropolitan Area Designation	-0.0261	(0.7219)	

5.5 Multiple Linear Regression (MLR) Modeling

We used the GLMSELECT procedure within SAS as a starting point for model selection. The procedure uses forward, backward, stepwise, LAR, and LASSO variable selection techniques. MLR models were

developed for predicting both response variables of interest – the cubed root of PE cost ratio and the cubed root of PE duration.

5.5.1 MLR Baseline for PE Cost Ratio Prediction

We considered all 16 independent variables for model selection using the GLMSELECT procedure in SAS. The categorical variables were analyzed as split variables so that only the significant level(s) of each categorical variable would be considered in the regression result. Table 5.7 identifies the best variable selection achieving a model fit having an adjusted R² value of 0.5210. Coefficients for each selected variable are included in Table 5.7.

Table 5.7 Baseline MLR Model for Predicted Cubed Root of PE Cost Ratio

Selected MLR Variables		Coefficient	
Intercept		β_0	0.7308
Project Construction Scope (split)			
x_1	Rehabilitation and Resurfacing	β_1	0.2398
x_2	Interchange	β_2	-0.1573
x_3	Right of Way Cost to STIP Estimated Construction Cost	β_3	0.2501
x_4	Roadway Percentage of Construction Cost	β_4	-0.3154
x_5	Number of Lanes	β_5	-0.0240
NCDOT Division (split)			
x_6	Division 08	β_6	0.1011

The predicted response can be found from the regression equation below utilizing the Table 5.7 coefficients.

$$\text{Predicted cubed root of PE cost ratio} = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_3(x_3) + \beta_4(x_4) + \beta_5(x_5) + \beta_6(x_6)$$

5.5.2 MLR Baseline for PE Duration Prediction

We performed a similar GLMSELECT analysis in SAS using all 16 variables to determine the best fit model for predicting the cubed root of PE duration. Categorical variables were again split. Table 5.8 presents the variables selected. The model achieved an adjusted R² value of 0.7281.

Table 5.8 Baseline MLR Model for Predicted Cubed Root of PE Duration

Selected MLR Variables		Coefficient	
Intercept		β_0	3.2872
Project Construction Scope (split)			
x_1	Rehabilitation and Resurfacing	β_1	-1.0134
STIP Prefix (split)			
x_2	Rural (R)	β_2	-0.5012
x_3	Right of Way Cost to STIP Estimated Construction Cost	β_3	1.4387
x_4	STIP Estimated Construction Cost	β_4	5.8132 E-8
x_5	Project Length	β_5	-0.0234
NCDOT Division (split)			
x_6	Division 04	β_6	0.4520

The predicted response can be found from the regression equation below utilizing the Table 5.8 coefficients.

$$\text{Predicted cubed root of PE duration} = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_3(x_3) + \beta_4(x_4) + \beta_5(x_5) + \beta_6(x_6)$$

5.6 Hierarchical Linear Model (HLM)

Recall from Table 5.4 that the roadway database contains six project parameters that are categorical. The baseline MLR models did select categorical variables for best model fit. A hierarchical modeling approach was investigated to further capitalize on the influence of categorical variables. The HLM utilizes MLR on a selection of projects based on a hierarchy created by shared categorical variable values. The HLM process then is to select the numerical variables for each hierarchical grouping that provides the best model fit. Previously, in the bridge project analyses, the HLM process supported three hierarchical levels. However, the smaller number of roadway projects limits hierarchical levels to only two for roadways. The hierarchical modeling techniques used for predicting the two response variables – cubed root of PE cost ratio and cubed root of PE duration – are described in this section.

We employed information gain theory to assess which hierarchical organization scheme was most beneficial for model fitting. The equation for calculating information gain was presented in Section 3 and is repeated here. Information gain is calculated by equation below.

$$\text{Information Gain} = \sum \left(\frac{N_{\text{cell}}}{N_{\text{total}}} \right) (-\ln(1 - R^2))$$

N_{cell} = number of projects in each subgroup cell

N_{total} = number of total projects

R² = Coefficient of Determination from subgroup MLR modeling

By utilizing candidate categorical variables as an organizing scheme, the roadways projects were divided into subgroups. Each subgroup was used to fit a MLR model. The objective of this HLM approach is to more closely fit a predictive model based on commonalities among projects within the same subgroup and delineate the differences between subgroups as evidence by differences in the resulting MLR equations. Different organizing schemes were applied for each response variable. Both are described in the following sections.

5.6.1 HLM for PE Cost Ratio

Table 5.9 displays the comparative information gain values achieved using different organizing schemes applied to the 150 roadway modeling projects. The organization scheme that maximizes information gain adds greater value to the regression modeling effort. We modified values for geographical area of the state to combine mountainous projects with very mountainous projects. Too few projects with the very mountainous level assignment existed to support regression analysis. We employed ANOVA analysis to test for differences among the variable's levels. If differences between levels were not supported statistically, we combined levels to increase the number of projects within a subgroup. These changes were made to combine levels within the project construction scope variable. Interchange scope projects were combined with new location scope projects. Projects classified as "other" were combined with rehabilitation and resurfacing projects, and projects with no level assigned were combined with widening scope projects.

Table 5.9 displays the initial information gain of 0.397. An increase in information gain was evident by applying a tier 1 sub grouping by either STIP prefix, project construction scope, or geographical area of the state. This gain was maximized when the geographical area of the state was used as the sub grouping scheme (0.634 compared with 0.494 or 0.419). Therefore the geographical area of the state was used for tier 1 of the hierarchy. Similarly, the tier 2 categorical variable was chosen by comparing the increase in information gain. Project construction scope used as the sub grouping scheme for tier 2 maximized the information gain (1.120 compared with 1.008). This scheme is reflected in the hierarchy presented in Table 5.10.

Table 5.9 Information Gain Comparison for PE Cost Ratio Hierarchical Schemes

SUBGROUP COMBINATION	INITIAL INFO GAIN	TIER 1 INFO GAIN	TIER 2 INFO GAIN
No Subgroups	0.397		
Geographical Area of State (modified) <ul style="list-style-type: none"> Mountainous combined with Very Mountainous 		0.634 CONTROLS	
Geographical Area of State (modified) and Project Construction Scope (modified) <ul style="list-style-type: none"> New Location combined with Interchanges Rehab & Resurfacing combined with Other Widening combined with Unspecified 			1.120 CONTROLS
Geographical Area of State (modified) With STIP Prefix			1.008
Project Construction Scope		0.494	
STIP Prefix		0.419	

When fitting a MLR model to the hierarchy of Table 5.10, only numerical variables were considered as candidate variables. Prior efforts (using GLMSELECT) had selected three numerical variables for predicting cubed root of PE cost ratio. These same three were used in the MLR applied to the hierarchical organization. Because of the correlation between PE cost ratio and PE duration, the numerical variables selected from duration modeling were also considered. In addition to the five numerical variables, first level interactions (variable*variable) and the quadratic form (variable²) of the five variables comprise the full set of candidate variables used in model fitting. In total, 16 candidate models were evaluated for model fit using adjusted R² as the fit criteria.

Using the hierarchical organization of Table 5.10, a MLR model was fit to each subgroup of projects. Five subgroups achieved an adjusted R² value exceeding 0.60; an improvement in fit over the baseline MLR model with no grouping by categorical variables (adjusted R² of 0.5210). These five subgroups are darkly shaded in Table 5.10 and identified as the “Good Fit” cells. Their adjusted R² values ranged from 0.6139 to 0.9440.

Model fitting for two subgroups depicted in Table 5.10 did not achieve a better fit than the baseline MLR model. These two subgroups are lightly shaded in Table 5.10, are identified as “Poor Fit” cells, and produced adjusted R² values of 0.2022 and 0.4389.

Table 5.11 provides the complete regression parameters for the MLR models fit to the seven subgroups reflected in Table 5.10 (cells darkly shaded and lightly shaded).

In section 8, we compare models further through our validation process to assess predictive performance.

Table 5.10 Hierarchical Organization for Predicting PE Cost Ratio

Good Fit (n=69) 5 cells
 Poor Fit (n=81) 2 cells
 Tier Not Applicable

Tier 1		Tier 2	
Geographical Area of State	N	Project Construction Scope	N
Coast [0.4686, Q1B]	28	New Location or Interchanges [0.6139, 5N]	5
		Rehabilitation and Resurfacing or Other [0.9440, Q1A]	8
		Widening or Unspecified [0.7950, IG]	15
Mountainous and Very Mountainous [0.7105, Q1A]	26	New Location or Interchanges [0.9902, 5N]	3
		Rehabilitation and Resurfacing or Other [0.9385, 5N]	10
		Widening or Unspecified [0.8441, 5N]	13
Piedmont [0.3754, Q1D]	96	New Location or Interchanges [0.6936, IG]	15
		Rehabilitation and Resurfacing or Other [0.2022, Q1C]	40
		Widening or Unspecified [0.4389, Q1D]	41
Total Projects	150		150

[Adjusted R² values shown in brackets with model identification]

Table 5.11 HLM Model for Predicted Cubed Root of PE Cost Ratio

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 5.10						
		Coast	Coast	Coast	Mountain	Piedmont	Piedmont	Piedmont
		New Location	Rehab & Resurface	Widening		New Location	Rehab & Resurface	Widening
Intercept		0.9070	-1.4874	0.6789	0.6335	1.3281	0.4650	0.6259
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO		0.2224	0.0873	0.0597			0.3364
STIP Estimated Construction Cost	TIP_COST				-4.5E-09	-1.7E-08	-1.6E-08	
Project Length	LEN		0.0180	-0.0342	-0.0137			-0.0069
Length of Structures within Project	ST_LEN		-93.5090	-69.0615		-5.0112		-1.8068
Roadway Percentage of Construction Cost	RW	-0.5202	5.5868	-0.2212		-0.8875	-0.1104	-0.2029
Quadratic form - Roadway Percentage of Construction Cost	rw22		-4.0127		-0.1691			
Interaction - Project Length with Length of Structures within Project	len2st			18.1779		1.8816		
Quadratic form - STIP Estimated Construction Cost	tip22						4.74E-16	
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22							-0.1945

5.6.2 HLM for PE Duration

A hierarchical organization was also investigated to improve model fit for predicting PE duration. Table 5.12 displays the information gain values for two possible hierarchical schemes.

Table 5.12 Information Gain Comparison for PE Duration Hierarchical Schemes

SUBGROUP COMBINATION	TIER 2 INFO GAIN
Geographical Area of State and Project Construction Scope	1.805
Geographical Area of State With STIP Prefix	2.270 CONTROLS

A two tiered hierarchy by geographical area of the state (Tier 1) with STIP prefix (Tier 2) maximized the information gain (2.270 compared with 1.805). Table 5.13 shows the resulting hierarchy. MLR analyses were then used to select the numerical variables providing best model fit.

Table 5.13 Hierarchical Organization for Predicting PE Duration

Good Fit
 Poor Fit
 Tier Not Applicable

Tier 1		Tier 2	
Geographic Area of State	N	STIP Prefix	N
Coast [0.6140]	13	INTERSTATE (I) (NA)	
		RURAL (R) [0.8450]	8
		URBAN (U) [0.9127]	5
Mountain [0.7118]	8	INTERSTATE (I) (NA)	
		RURAL (R) [0.9666]	6
		URBAN (U) (NA)	2
Piedmont [0.6303]	42	INTERSTATE (I) [0.9883]	4
		RURAL (R) [0.5653]	20
		URBAN (U) [0.7035]	18
Very Mountainous [0.9991]	4	INTERSTATE (I) (NA)	
		RURAL (R) [0.9999]	3
		URBAN (U) (NA)	1
Total Projects	67		67

[Adjusted R² values shown in brackets]

The modeling set for PE duration analysis contains 90 projects. However, 23 of these 90 projects were not classified into subgroups within the hierarchy of Table 5.13 because of missing data. Thus, the hierarchical scheme reflects grouping for only 67 projects. The first tier of the hierarchy achieves a model fit (assessed by adjusted R² values) higher than 0.60 for three of the four cells. The cell depicting

the very mountainous level of geographical area of the state does not contain a sufficient number of projects for valid regression results. Similarly, most of the cells within Tier 2 have small numbers of observations causing regression analyses to be invalid. Therefore, the second tier hierarchy depicted in Table 5.13 provides limited additional benefit.

Table 5.14 provides the complete regression parameters for the MLR models fit to the three subgroups reflected in Table 5.13 (cells darkly shaded).

Table 5.14 HLM Model for Predicted Cubed Root of PE Duration

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 5.13		
		Coast	Mountain	Piedmont
Intercept		2.1508	3.4690	4.0196
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO	2.1666	0.8906	1.3652
STIP Estimated Construction Cost	TIP_COST	1.09E-07	6.06E-08	7.33E-08
Project Length	LEN		-0.1110	-0.0302
Roadway Percentage of Construction Cost	RW			-1.1865

In section 8, we compare models further through our validation process to assess predictive performance.

6.0 ESTIMATING PE COST RATIO USING MULTILEVEL DIRICHLET PROCESS LINEAR MODELING

6.1 Introduction

Estimating PE cost is challenging because its characteristic is not fully understood. Prediction is a primary reason of modeling because it provides an estimate of interest under uncertainty from historical or observable data. In response to increasing complexity of data, modeling methods have become sophisticated and complicated. However, the current art of prediction modeling for complicated data sets is human-added or a brute-force search for the covariate-response relationship (linear or nonlinear), the form of variance component (homoscedasticity or heteroscedasticity), the degree of heterogeneity (the number of statistically different components), the choice of covariate (fixed or random effects), the existence of interactions between covariate variables or component variances, etc. Such uncertainty factors have significant effects on both prediction performance and computational efforts. Unfortunately, there is no model or systematic way of modeling that can deal with such uncertainties simultaneously. The goal of this study is to propose a way of modeling complex data, such as PE cost, in order to create a new prediction model. In contrast to existing models where a model-maker resolves uncertainty through numerous statistical tests, this proposed model is built in a layer-by-layer fashion with minimal assumptions on uncertainty. The modeling procedure decides the layer of prediction to avoid overfitting, and the whole procedure can be performed without human efforts in result. We provide the definition of heterogeneity and explain the Dirichlet process linear model (DPLM) in Section 6.2, and develop the multilevel Dirichlet process linear model (MDPLM) and the modeling procedure in Section 6.3. In Section 6.4, we construct the bridge model for PE cost ratio estimation using the developed MDPLM and analyze the performance of resulting model. Additionally, Section 6.5 provides our initial efforts to develop a roadway model for PE cost ratio estimation with DPLM. The final modeling results are reported in Section 8 along with a discussion of the benefits and deficiencies of MDPLM.

6.2 Background

We provide the definition of heterogeneity and explain the Dirichlet process linear model (DPLM) in this section.

6.2.1 Heterogeneous Population

In general heterogeneity means the lack of uniformity in the concept of interest. With respect to data, two possible interpretations exist about heterogeneity. A data set collected from different sources may have different data types and scales. For instance, the heights of a group are measured in feet, but the heights of another group are measured in two levels such as ‘tall’ and ‘short’. The other interpretation of data heterogeneity is statistical difference. A data set is assumed to be sampled from a heterogeneous population if the sample can be partitioned into sets with each having statistically significant different parameters. Such a non-overlapping set in the partition is called a class or cluster and such a partition can be obtained by minimizing a loss measure or maximizing a likelihood measure.

This paper assumes that a complicated data set consists of heterogeneous population data filled with uncertainty about the covariate-response relationship, the form of variance component, the degree of heterogeneity, the existence of interactions between covariates or component variances.

6.2.2 Dirichlet Process Linear Mixture Model (DPLM)

A Dirichlet process linear model (DPLM) is a nonparametric regression technique which can handle heterogeneous population data. The prediction of DPLMs is robust because a Bayesian model averaging estimate over competing models accommodates the variation of prediction and avoids overfitting. A DPLM is also flexible in that the degree of heterogeneity (the number of statistically different mixture

components) does not need to be decided before modeling. Therefore it provides a convenient way of reducing human efforts and errors incurred by statistical tests and decision of arbitrary thresholds.

For covariate x and response y , the distribution of $y|x$ can be represented in a nonparametric way with latent parameter $\theta = (\theta_x, \theta_y)$ as like

$$\begin{aligned} F(y|x) &= \int F(y|x, \theta) dG(\theta) \\ &= \int F_y(y|x, \theta_y) F_x(x|\theta_x) dG(\theta) \end{aligned} \quad (1)$$

where θ_x and θ_y are two independent parameters specific to x and y , respectively, and F, G, F_x and F_y denote the distribution functions of their own arguments.

Applying Dirichlet process (DP) prior to θ as follows

$$\begin{aligned} \theta &\sim G \\ G &\sim DP(G_0, \alpha), \end{aligned} \quad (2)$$

the *stick-breaking* construction makes the parameter θ have a discrete distribution with probability 1 as follows

$$G(\theta) \sim \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}(\theta). \quad (3)$$

where $\delta_{\theta_j}(\theta)$ is the Dirac delta function which is 1 when θ is equal to θ_j or 0 otherwise.

Due to the DP prior, the posterior distribution of θ also follows DP distribution as

$$G|\theta_{1:K} \sim DP\left(\frac{1}{\alpha + K} \sum_{j=1}^K \delta_{\theta_j} + \frac{\alpha}{\alpha + K} G_0, \alpha + n\right). \quad (4)$$

By the *Polya urn scheme*, the marginal distribution of θ is

$$G(\theta_i|\theta_{-i}) \sim \frac{1}{\alpha + K - 1} \sum_{j=1}^K \delta_{\theta_j} + \frac{\alpha}{\alpha + K - 1} G_0 \quad (5)$$

and Eq. 1 has a mixture representation

$$F(y|x) = \frac{1}{\alpha + K - 1} \sum_{j=1}^K F_y(y|x, \theta_{y,j}) F_x(x|\theta_{x,j}) + \frac{\alpha}{\alpha + K - 1} \int F_y(y|x, \theta_y) F_x(x|\theta_x) dG_0(\theta). \quad (6)$$

where $\theta_{x,j}$ and $\theta_{y,j}$ are parameters related to the covariate and the response of component j , respectively.

Given a value of covariate x , the estimate of response y is

$$\begin{aligned} E(y|x) &= \int f_y(x, \theta) F(y|x, \theta) dG(\theta) \\ &= \frac{1}{b} \sum_{j=1}^K y_j F_y(y_j|x, \theta_{y,j}) F_x(x|\theta_{x,j}) + \frac{\alpha}{b} \int f_y(x, \theta) F_y(y|x, \theta_y) F_x(x|\theta_x) dG_0(\theta) \end{aligned} \quad (7)$$

where f_y is the relationship function of the response, y_j is the response estimate of mixture component j (component-wise estimate, that is, $y_j = f_y(x, \theta_j)$), and b is the probability normalizing constant as follows

$$b = \sum_{j=1}^K F_y(y_j|x, \theta_{y,j}) F_x(x|\theta_{x,j}) + \alpha \int F_y(y|x, \theta_y) F_x(x|\theta_x) dG_0(\theta). \quad (8)$$

In a DPLM, each mixture component is assumed to have a linear covariate-response relation for F_y with hyper-parameters β and s_y and multivariate normal distribution for F_x with hyper-parameter μ and s_x as follows

$$y \sim \mathcal{N}(x\beta, s_y^{-1}) \quad \text{and} \quad x \sim \mathcal{N}(\mu, s_x^{-1}). \quad (9)$$

Hannah et al. (2010) and Hurn et al. (2000) describe detailed structure and sampling methods for other hyper-parameters of DPLM.

6.3 Multilevel Dirichlet Process Linear Model (MDPLM)

Since the DPLM still assumes that random error of mixture components is independent and identically distributed (*i.i.d.*), it has difficulty in modeling the case of data where variance components are of

heteroscedasticity and interaction. As a nonlinear extension of DPLM, the Dirichlet process generalized linear model (DPGLM) [Hannah et al. 2010; Mukhopadhyay and Gelfand 1997] allows nonlinear relationship functions between the linear predictor and the mean of response through the inverse of link function as like generalized linear models [McCullagh and Nelder 1989; Wedderburn 1974]. The DPGLM model also makes possible the change of random error according to covariate by incorporating variance components into the linear predictor. However, the choice of the covariate-response function (or distribution) and the form of variance component requires human efforts throughout possible candidates, and the random errors of mixture components are still assumed to be independent. As another extension of DPLM, we propose the multilevel Dirichlet process linear model (MDPLM) to overcome the deficiencies of DPLM models while keeping the simple assumptions of DPLM.

6.3.1 Model Construction

The MDPLM consists of DPLM layers, each of which has two distinct DPLMs (a generic DPLM and an overfit DPLM) except the last layer as shown in Figure 6.1. While the generic DPLM model, denoted by DPLM, of a layer is equivalent to the DPLM model, the overfit DPLM model, denoted by cDPLM, provides the overfit estimate of target value which will be used to generate a hidden random effect covariate value. Assuming every layer is conditionally independent, the given covariate and random error is also *i.i.d.* The estimate of response y at the layer L given a sequence of covariates is simply the sum of estimates of layers as follows:

$$E(y|\mathbf{X}^{(0)}, \dots, \mathbf{X}^{(L)}) = \hat{y} + \sum_{\ell=0}^{L-1} \hat{e}_{\ell} \quad (10)$$

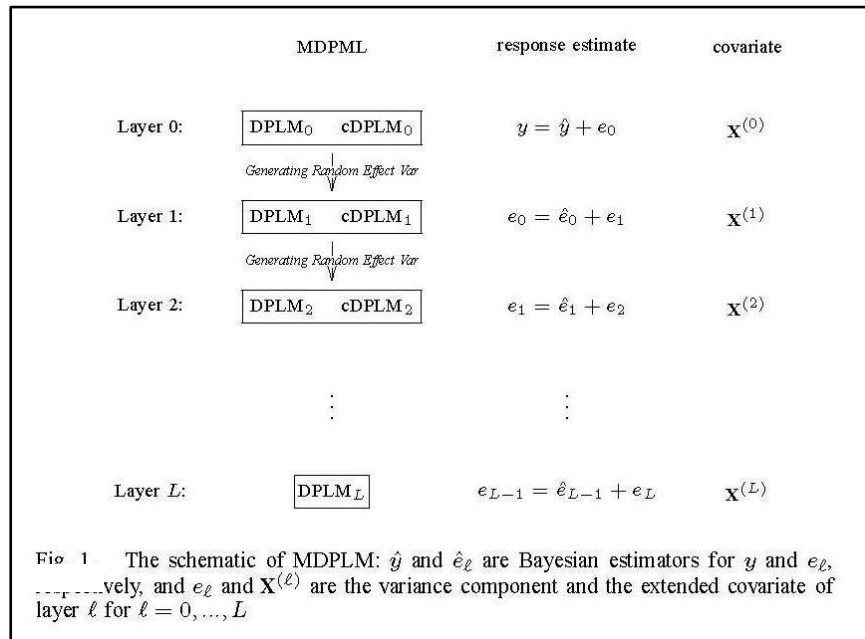


Figure 6.1 MDPLM Schematic

6.3.2 Covariate Extension of Layers

The simple description of MDPLM is a sequential variance estimate model where the variance component of a layer is estimated in the next layer. If every layer used the same covariate, variance would not change and this can be easily shown because the DPLM assumes a variance component is *i.i.d.*

$$\text{var}(e_\ell) = \text{var}(e_{\ell-1}|x) = \text{var}(e_{\ell-1}) \quad \text{for } \ell = 1, \dots, L \quad (11)$$

Hence it is necessary to extend the covariate for reducing the variance component of layers. Suppose we extend the covariate of layers in the following way:

$$\begin{aligned} \mathbf{X}^{(\ell)} &= [\mathbf{X}^{(0)} : \mathbf{x}_e^{(\ell-1)}] \quad \text{for } \ell = 1, \dots, L \\ \mathbf{X}^{(0)} &= \mathbf{X} \\ \mathbf{x}_e^{(\ell)} &= \frac{c_2}{\pi} \{ \arctan(\mathbf{d}^{(\ell)} - c_1) + \arctan(\mathbf{d}^{(\ell)} + c_1) \} \quad \text{for } \ell = 1, \dots, L \\ \mathbf{d}^{(\ell)} &= [d_1^{(\ell)}, \dots, d_n^{(\ell)}]^T \quad \text{for } \ell = 1, \dots, L \\ d_i^{(\ell)} &= \hat{e}'_{\ell-1,i} - \hat{e}_{\ell-1,i} \quad \text{for } \ell = 1, \dots, L \text{ and } i = 1, \dots, n \end{aligned} \quad (12)$$

where $\hat{e}'_{\ell,i}$ and $\hat{e}_{\ell,i}$ are the estimates of target value $e_{\ell,i}$ by cDPLM and DPLM, respectively, and c_1 and c_2 are constant. Assuming $|e_{\ell,i} - \hat{e}'_{\ell,i}| \leq |e_{\ell,i} - \hat{e}_{\ell,i}|$, that is, $\hat{e}'_{\ell,i}$ is more close to $e_{\ell,i}$ than $\hat{e}_{\ell,i}$, the covariate extension process is equivalent to the generation of a random effect variable categorized into over-and-under fitting of target in the DPLM framework. Define $S^- = \{e_{\ell,i} | d_i^{(\ell)} < 0, i = 1, \dots, n\}$ and $S^+ = \{e_{\ell,i} | d_i^{(\ell)} \geq 0, i = 1, \dots, n\}$. Let $n_n = |S^-|$, $n_p = n - n_n$ and $E(S^+) = a \geq 0$. Since $E(e_\ell) = 0$, $E(S^-) = -a$. Then the variance component of layer ℓ can be decomposed as follows.

$$\begin{aligned} \text{var}(e_\ell) &= \frac{1}{n} \sum_{i=1}^n (e_{\ell,i})^2 \\ &= \frac{1}{n} \left(\sum_{e \in S^+} e^2 + \sum_{e \in S^-} e^2 \right) \\ &= \frac{1}{n} \left\{ \sum_{e \in S^+} (e - a + a)^2 + \sum_{e \in S^-} (e + a - a)^2 \right\} \\ &= \frac{n_p}{n} \text{var}(S^+) + \frac{n_n}{n} \text{var}(S^-) + a^2 \end{aligned} \quad (13)$$

Now let $d^{(\ell)}$ be the random variable representing the distribution of $d_i^{(\ell)}$. The variance component of layer $(\ell+1)$ can be written by

$$\begin{aligned} \text{var}(e_{\ell+1}) &= \text{var}(e_\ell | \mathbf{X}^{(\ell+1)}) \\ &= \text{var}(e_\ell | \mathbf{X}^{(\ell+1)}, d^{(\ell)} \geq 0) p(d^{(\ell)} \geq 0 | \mathbf{X}^{(\ell+1)}) + \text{var}(e_\ell | \mathbf{X}^{(\ell+1)}, d^{(\ell)} < 0) p(d^{(\ell)} < 0 | \mathbf{X}^{(\ell+1)}) \\ &= \frac{n_p}{n} \text{var}(S^+) + \frac{n_n}{n} \text{var}(S^-) \end{aligned} \quad (14)$$

Thus, the covariate extension described by Eq. 12 reduces variance layer by layer, and if error is well distributed, the variance would be reduced by a factor of 2 in the next layer.

6.3.3 Overfit Model of Layers

It is no wonder that we can always find an overfit model given a usual performance measurement. For instance, including more explanatory variables to a regression model produces a model which reduces the mean square error over the data set. In the DPLM framework, an overfit model can be obtained by adding designed noises to the sampling scheme.

In general Eq. 7 cannot be computed analytically because of non-closed-form integrals involved. However, it is computed approximately using the Monte Carlo integration. Suppose $\theta^{(t)}$ be the set of model parameters sampled at iteration t counted after burn-in for $t = 1, \dots, T$. Given the value of x , the response estimate is approximately

$$E(y|x) \approx \frac{1}{T} \sum_{t=1}^T E(y|x, \theta^{(t)}) \quad (15)$$

and the model-wise estimate is

$$E(y|x, \theta^{(t)}) = \frac{\sum_{j=1}^K E(y|x, \theta_j^{(t)}) f_x(x|\theta_j^{(t)}) + \alpha^{(t)} \int E(y|x, \theta) f_x(x|\theta) G_0(d\theta)}{\sum_{j=1}^K f_x(x|\theta_j^{(t)}) + \alpha^{(t)} \int f_x(x|\theta) G_0(d\theta)}. \quad (16)$$

During the parameter sampling at some iteration $s \in [1, T]$, we add a new component $(K + 1)$ such that $E(y|x, \theta_{K+1}^{(s)}) = y_i$ and $f_x(x_i|\theta_{K+1}^{(s)}) \gg f_x(x_i|\theta_j^{(s)})$, $\forall j \in [1, K]$, where (x_i, y_i) is a covariate-response pair in the training data set. Assuming $\alpha^{(s)}$ is neglectable (which is true in most case after burn-in), the model-wise estimate at iteration s is

$$E(y|x_i, \theta^{(s)}) \approx y_i. \quad (17)$$

Suppose that there are two DPLM models, denoted by DPLM and cDPLM separately, and we add such a new component as a designed error during the parameter sampling of cDPLM. Let S be the set of iterations at which such a new component is added for a covariate and response pair (x_i, y_i) . Also, let $\hat{y}_{i,t}$ and $\hat{y}'_{i,t}$ be the model-wise response estimates for y_i at iteration t by DPLM and cDPLM, respectively. Then the mean square error of cDPLM over DPML for (x_i, y_i) is

$$\frac{\sum_{t=1}^T (\hat{y}'_{i,t} - y_i)^2}{\sum_{t=1}^T (\hat{y}_{i,t} - y_i)^2} = \frac{\sum_{t \notin S} (\hat{y}'_{i,t} - y_i)^2}{\sum_{t \notin S} (\hat{y}_{i,t} - y_i)^2 + \sum_{t \in S} (\hat{y}_{i,t} - y_i)^2}. \quad (18)$$

Assuming $\sum_{t \notin S} (\hat{y}'_{i,t} - y_i)^2 \approx \sum_{t \notin S} (\hat{y}_{i,t} - y_i)^2$, which is achievable by setting $|S| \ll |T|$,

$$\sum_{t=1}^T (\hat{y}'_{i,t} - y_i)^2 \leq \sum_{t=1}^T (\hat{y}_{i,t} - y_i)^2. \quad (19)$$

Similarly, for the covariate-response pairs (\mathbf{X}, \mathbf{Y}) , an overfit DPLM model, denoted by cDPML $_{\ell}$, in the layer ℓ of the MDPLM can be also constructed by adding designed errors during parameter sampling.

6.3.4 The Layer of Prediction

The variance reduction of MDPLM was proved in Section 6.3.1, but the reduction of variance in a higher layer does not mean that the prediction at that layer is better than at lower layers because of data overfitting. Since the MDPLM utilizes data overfitting to find unobservable random effects, this may cause the MDPLM to overfit a training data set if the found random effects exist only in the training data set. Thus choosing a proper layer of prediction would result in a robust MDPLM avoiding overfit of the model.

The MDPLM can be considered as a nonlinear system where the data set (\mathbf{X}, \mathbf{Y}) is input, a performance measure is output, and the random effect variables generated in layers are noises. In the stochastic resonance framework where there is an optimal level of noise to minimize (or maximize) a performance measure, we use the layer of prediction in the MDPLM as the level of noise because more layers involve more random effects generated [Benzi et al. 1983; McNamara and Wiesenfeld 1989]. Such system-wise representation of the MDPLM is shown in Figure 6.2.

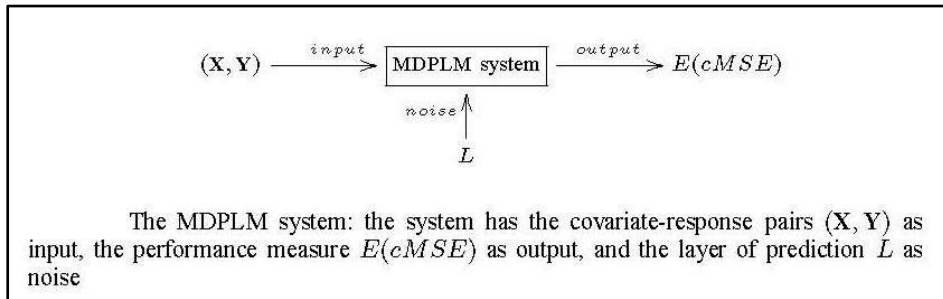


Figure 6.2 System-wide Representation of the MDPLM

The best model balances between training (observed) and validating (unobserved) errors. Since it is common that the size of training data is larger than that of validation, the simple arithmetic mean of two measurements tends to be biased and the training data dominates the performance measure. Therefore we use a different way of merging two mean square errors, denoted by $cMSE$, as the performance measure to resolve uncertainty when two measurements are very different and formulate in Eq. 20.

$$\begin{aligned}
cMSE &= p \cdot MSE_t + (1 - p) \cdot MSE_v & (20) \\
v_1 &= 1 - \frac{MSE_t}{MSE_t + MSE_v} \quad \text{and} \quad v_2 = 1 - v_1 \\
h_1 &= -v_1 \log_2 v_1 \quad \text{and} \quad h_2 = -v_2 \log_2 v_2 \\
p &= \frac{h_1}{h_1 + h_2}
\end{aligned}$$

where MSE_t and MSE_v are the mean square errors of training and validation, respectively.

As mentioned in Section 6.2.1, heterogeneous data are filled with lots of uncertainties and it is difficult to obtain a stratified training validation data set pair. Using the DPLM trained with the whole data $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$, the expected number of mixture components can be computed by $E(K) = \frac{1}{T} \sum_{t=1}^T K^{(t)}$ where $K^{(t)}$ is the number of mixture components at iteration t . To make the training set have approximate half of instances for each mixture component, the size of validation sets is set to $\frac{n}{2E(K)}$. For less biased performance measure, h -fold cross-validation is used to compute $E(cMSE)$ with modification. We sample h validation sets, denoted by \mathbf{Dv}_i for $i = 1, \dots, h$, without replacement and the training sets, denoted by \mathbf{Dt}_i for $i = 1, \dots, h$ such that $\mathbf{Dt}_i = \mathbf{D} \setminus \mathbf{Dv}_i$. Let $cMSE_i$ be the performance measure of $(\mathbf{Dt}_i, \mathbf{Dv}_i)$. Then the comprehensive performance measure is $E(cMSE) = \frac{1}{h} \sum_{i=1}^h cMSE_i$.

6.4 Bridge Model for Predicting PE Cost Ratio

We assume that construction project data of the response PE cost ratio are heterogeneous in that we are ignorant of the degree of heterogeneity, the covariate-response relations, the form of variance components, and the existence of interactions between covariance and component variances. With the MDPLM and the procedure developed in Section 6.3, we construct a prediction model and evaluate its performance.

The bridge construction data contains 505 bridge projects let for construction between January 1999 and June 2008. For constructing a prediction model, 13 independent variables were selected as covariate with the response, the percentage scaled PE cost ratio, as shown in Table 6.1.

From the DPLM trained by 505 bridge project observations, the expected number of mixture components was approximately 10. As described in Section 6.3.4, the size of a validation set was set to 25. We set the number of training-validation data set pairs to 10 for cross-validation. Figure 6.3 shows the predictive performance of the model averaged over 10 training-validation data set pairs by layer transition. Mean square error (MSE), mean absolute error (MAE), and mean absolute relative error ($MARE$) were used as base error measurements in panels (a), (b), and (c), respectively. The prefix “ t ” or “ v ” designates which data set (training or validation) was used to compute the error measurements. The prefix “ c ” is the merged MSE computed from equation 20. Throughout plots in Figure 6.3 a clear trend is observed; for validation data sets, error measures ($E(vMSE)$, $E(vMAE)$ and $E(vMARE)$) decrease at the beginning as the layer of prediction become higher, and start to increase after a certain layer of prediction. $E(cMSE)$ also follows the trend, and has a minimal value at layer 4. The benefit of using MDPLM over DPLM (equivalent to a single layer MDPLM) can be clearly observed in Table 6.2. The validation errors have reduced by a factor of 3.356, 1.839 and 1.523 for mean square error; mean absolute error and mean relative error, respectively, comparing with the results of DPLM.

Table 6.1 MDPLM (Bridge Model) Listing of Variables

Covariate		Description
x1	Structure Type	Structure designation (bridge=1 or culvert=2)
x2	Roadway Percentage of Construction Cost	Proportion of total construction cost that can be attributed to the roadway portion (numerical)
x3	Structure Percentage of Construction Cost	Proportion of total construction cost that can be attributed to the structure components (numerical)
x4	Road System	Designation for arterial, collector, or local route system carried by structure (arterial = 1, collector=2, local route system = 3)
x5	NEPA Document Classification	Complexity of the environmental documentation required for each project (Categorical exclusion = 1 or Programmatic categorical exclusion = 2)
x6	Project Length	Total length of the project in miles (numerical)
x7	Planning Document Responsible Party	Party responsible for ensuring delivery of the NEPA document (Department of Transportation = 1 or Private Engineering Firm =2)
x8	Geographical Area of State	Project location (Coast=1, Mountain=2, Piedmont=3, Very Mountainous=4)
x9	Functional Classification	Classification of route (Rural=1 or Urban=2)
x10	STIP Estimated Construction Cost	Updated construction cost estimate (numerical)
x11	Estimated Right of Way Cost	Costs attributed to right of way (numerical)
x12	Right of Way Cost to STIP Estimated Construction Cost	Ratio of right of way costs over estimated construction costs (numerical)
x13	Number of Spans	Number of spans in main structure (numerical)
Response		Description
y	PE Cost Ratio (%)	Percentage of PE costs over STIP estimated construction costs

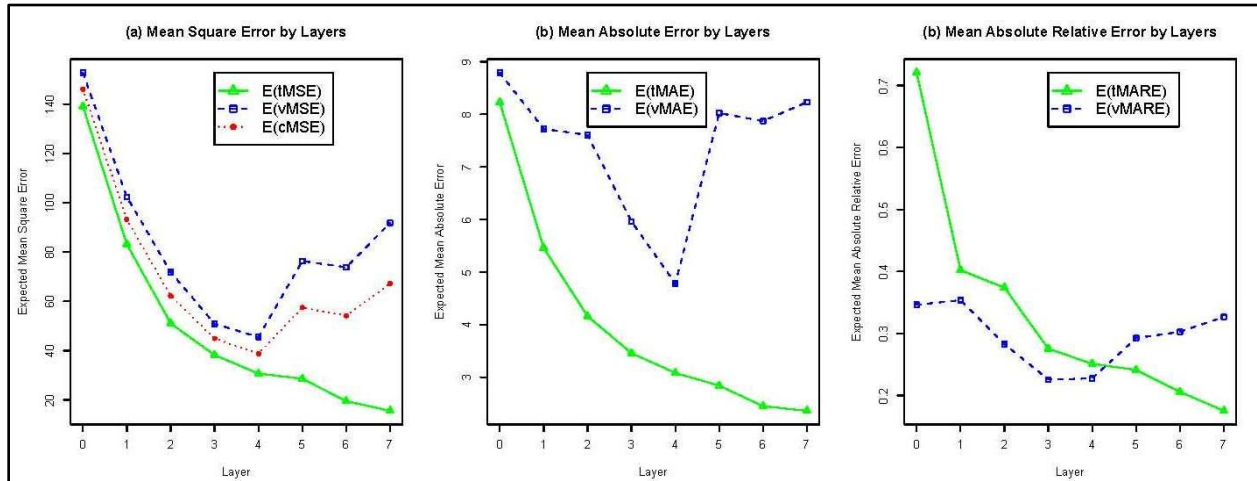


Figure 6.3 MDPLM (Bridge Model) Expected Error Measurements (using 10-Pair Cross Validation)

Table 6.2 MDPLM (Bridge Model) Comparison of Error Measurements (between Layer 0 and Layer 4)

Base Measure of Expected Value	Layer 0		Layer 4	
	Training	Validation	Training	Validation
Mean square error (<i>MSE</i>)	139.090	152.757	30.679	42.521
Mean absolute error (<i>MAE</i>)	8.227	8.792	3.081	4.782
Mean absolute relative error (<i>MARE</i>)	0.721	0.346	0.251	0.224

With the stochastic resonance theory, we have showed that the prediction at layer 4 balances observable (training) and unobservable (validating) data. The final bridge model was trained with all data available

(505 construction projects) and Figure 6.4 shows the plots of percentage scaled PE cost ratio estimates, denoted by \hat{y} , over the value in the data set by layers.

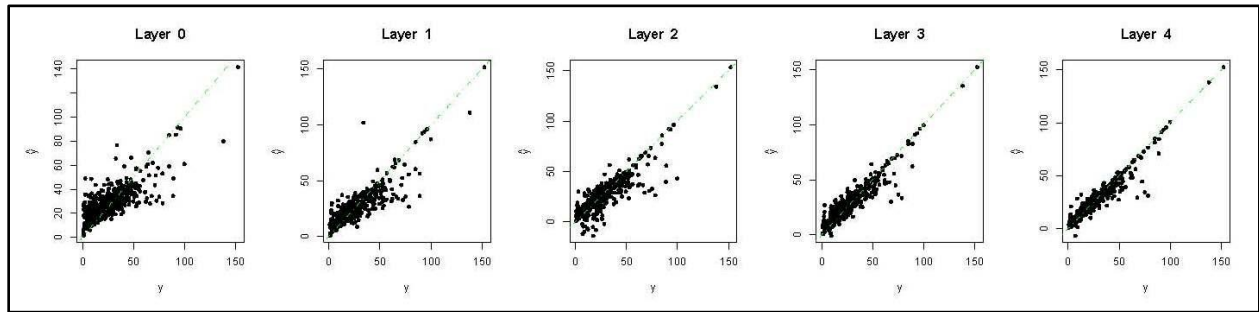


Figure 6.4 MDPLM (Bridge Model) Layer-wise Predictive Results
[y (actual) versus \hat{y} (estimate)]

6.5 Roadway Model for Predicting PE Cost Ratio

We constructed the roadway model for PE cost ratio estimation with the DPLM, which is a single layer MDPLM. Different from the MDPLM, the DPLM does not involve stochastic resonance and the generation of random effect variables.

The roadway data has 181 roadway projects let for construction between January 1999 and June 2008. For model construction, 11 independent variables shown in Table 6.4 were selected.

To measure the performance of the DPLM modeling approach applied to the roadway projects, we used 10-fold cross-validation with 3 different error measurements (mean square error, mean absolute error and mean absolute relative error). Table 6.5 shows the error measurements of 10 pairs of training-validation data sets and the averaged results over the pairs are listed in Table 6.6. Large difference in mean square error between training and validation implies that the pairs of data sets are not well stratified.

Table 6.3 Roadway Model Listing of Variables

Covariate		Description
x1	STIP Prefix	Designation of roadway project type in STIP (Interstate=1, Rural=2, or Urban=3)
x2	Roadway Percentage of Construction Cost	Proportion of total construction cost that can be attributed to the roadway portion (numerical)
x3	NEPA Document Classification	Complexity of the environmental documentation required for each project (Categorical exclusion = 1, Environmental assessment = 2, Environmental impact statement = 3, Programmatic categorical exclusion = 4, or State minimum requirements =5)
x4	Federal Funding Utilized	Indicates if federal funding is used (No=1 or Yes=2)
x5	Project Length	Total length of the project in miles (numerical)
x6	Planning Document Responsible Party	Party responsible for ensuring delivery of the NEPA document (Department of Transportation = 1 or Private Engineering Firm = 2)
x7	Geographical Area of State	Project location (Coast =1, Mountain=2, Piedmont=3, or Very Mountainous=4)
x8	Metropolitan Area Designation	Indicates if project is located in a metropolitan region (No=1 or Yes=2)
x9	STIP Estimated Construction Cost	Updated construction cost estimate (numerical)
x10	Estimated Right of Way Cost	Costs attributed to right of way (numerical)
x11	Right of Way Cost to STIP Estimated Construction Cost	Ratio of right of way costs over STIP estimated construction costs (numerical)
Response		Description
y	PE Cost Ratio (%)	Percentage of PE costs over STIP estimated construction costs

Table 6.4 DPLM (Roadway Model) Error Measurements
(of 10-Fold Cross Validation)

Error Measurements		Fold									
		1	2	3	4	5	6	7	8	9	10
Training	tMSE	87.367	88.754	71.534	51.437	41.269	82.383	48.612	69.489	92.944	177.097
	tMAE	5.899	5.573	4.315	3.937	4.095	4.135	4.533	4.714	3.887	5.896
	tMARE	2.017	4.751	3.540	3.652	4.176	1.553	5.204	2.317	2.611	4.961
Validation	vMSE	80.949	77.842	72.950	1167.050	834.057	92.396	93.193	100.277	303.531	1213.790
	vMAE	7.157	4.922	5.667	17.981	16.221	6.983	7.782	6.340	9.486	14.292
	vMARE	4.882	1.588	2.111	2.073	1.715	17.864	14.296	1.369	2.582	7.474

Table 6.5 DPLM (Roadway Model) Averaged Error Measurements
(over 10-Fold Cross Validation Data)

Averaged Error Measurements	Training	Validation
Mean square error (<i>MSE</i>)	81.089	403.604
Mean absolute error (<i>MAE</i>)	4.698	9.683
Mean absolute relative error (<i>MARE</i>)	3.478	5.595

7.0 USER INTERFACE APPLICATION

The research team has developed a graphical user interface application to facilitate ease of use of the predictive PE cost models developed. The interface is described in this section.

The interface application does not support predictive modeling of PE duration.

7.1 Interface Programming

The research team's programmer used Microsoft Visual C++ with MFC libraries to develop the user interface application. The interface acts as a hub to collect user data, direct data to the selected modeling library, report estimation results, and create a text file archive of the estimation results.

The interface software has dependencies on other runtime libraries provided by Microsoft. The required runtime libraries need to be installed on the user's computer before executing the interface software.

- a. Microsoft Visual Studio 2005 Redistributable package
 - Visual C++ 2005
 - <http://www.microsoft.com/downloads/en/confirmation.aspx?FamilyID=32bc1bee-a3f9-4c13-9c99-220b62a191ee&displaylang=en>
 - Visual C++ 2005 Service Pack 1
 - <http://www.microsoft.com/downloads/en/confirmation.aspx?FamilyID=200b2fd9-ae1a-4a14-984d-389c36f85647&displaylang=en>
- b. Microsoft .NET framework 2.0
 - <http://www.microsoft.com/downloads/en/confirmation.aspx?FamilyID=0856eacb-4362-4b0d-8edd-aab15c5e04f5&displaylang=en>

7.2 Estimate Initialization Inputs

The following figures illustrate the screen images from the user interface. Movement through the interface is controlled by action buttons positioned along the bottom of each screen.

Figure 7.1 shows the initial Login Screen.

Once logged in, the interface will guide the user through a series of input screens to control the modeling processes. The Project Startup screen is shown in Figure 7.2. The user decides initially if a PE cost estimate is desired for a Roadway project (typically having STIP prefixes of I, R, or U) or a Bridge project (with STIP prefix B). Secondly, one of two predictive models (Bayesian or Regression) is selected. As described elsewhere in this report, the Bayesian model generally provides a more accurate prediction of PE cost but the mathematics of the method are complex and more difficult to explain, while the Regression model generally provides a less accurate prediction but is easier to explain. Drop-down menus with suitable options are provided for both selections. Figure 7.2 displays the input screen for these decisions. Screen help is accessible by clicking on the “?” button located in the upper right corner of the screen. Movement is advanced by the “NEXT>>” action button positioned along the bottom of the screen.

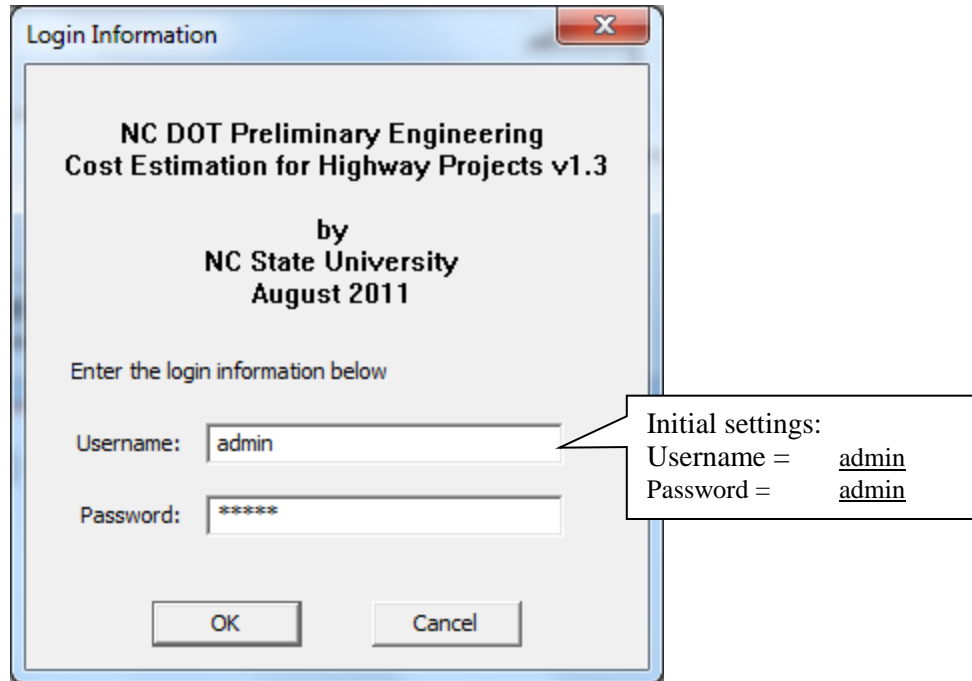


Figure 7.1 User Interface - Login Screen

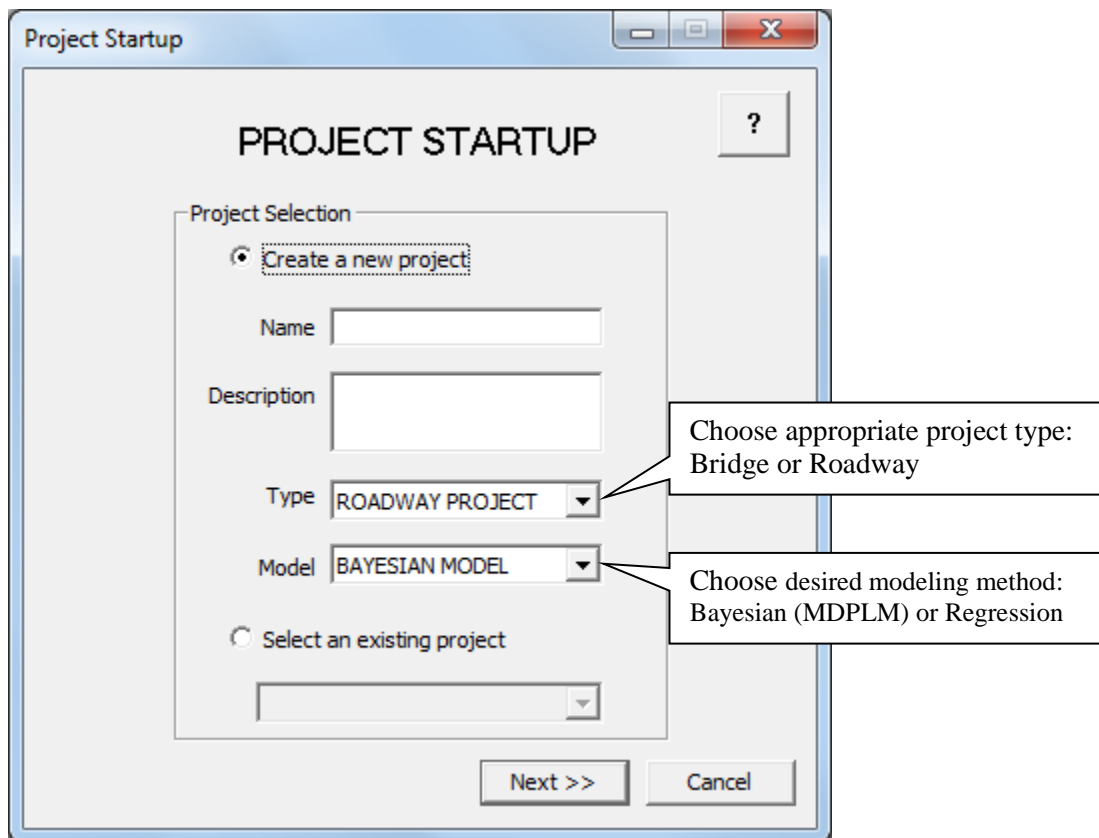


Figure 7.2 User Interface – Project Startup Screen

At the Project Startup screen, the user may also retrieve information from an earlier estimate (“Select an existing project” seen at bottom of Figure 7.2). As PE cost estimates are generated through the interface, project data are stored locally for easy retrieval. Uniform project naming schemes and project description schemes are recommended to make output reporting clear to users.

Following selection of project type and model, the user supplies specific cost information on the project on the third interface screen shown in Figure 7.3. The project’s STIP estimated construction cost and estimated right of way cost are requested. Additionally, the proportion of the STIP estimated construction cost attributed to roadway construction is desired. For bridge projects only, the proportion of STIP estimated construction cost attributed to structures is also desired. The STIP estimated construction cost includes the roadway portion, the structures portion, a mobilization and contingency factor, and administrative overhead. Right of way and utility relocation costs are not included in the STIP estimated construction cost. The current STIP may provide a project’s estimated construction cost and estimated right of way cost (including utilities) if the project has been programmed for STIP inclusion.

Either model predicts PE cost ratio (PE costs over estimated project construction cost) and reports an estimated PE cost (in \$) based on the user’s input of estimated project construction cost.

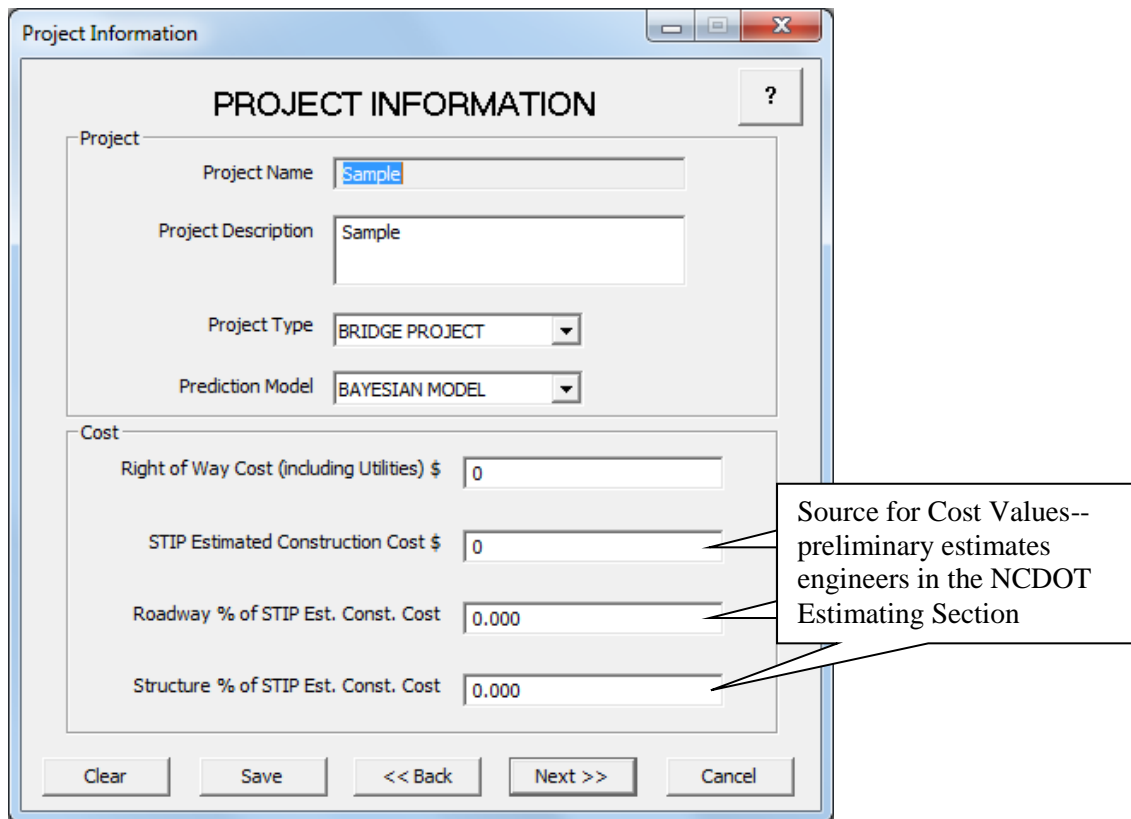


Figure 7.3 User Interface – Project Information Screen

When the predicted PE cost ratio is desired at an early stage of project development, only feasibility or functional cost estimates may be available. Estimates are prepared in the Contract Standards and Development Unit - Estimating Section by the preliminary estimates engineers. Figure 7.4 shows an excerpt from a functional estimate worksheet acquired from a typical NCDOT project. The full worksheet is attached in the Appendices for reference. The estimated construction cost, roadway portion of cost, and structure (with utilities) portion of cost are identified on the estimate worksheet. These three

cost figures, taken from the estimate worksheet, provide the needed data for the Project Information screen of the user interface.

North Carolina Department of Transportation
Preliminary Estimate

TIP No.
Route
From
Typical Section

Prepared By:
Requested By:

Estimate Type

Func County: _____

Date _____
Date _____

Estimated Construction Cost

CONSTR. COST
\$0

Line Item	Des	Sec No.	Description	Quantity	Unit	Price	Amount
			Clearing and Grubbing		Acre		\$ -
			Earthwork		CY		\$ -
			Design Existing Location		Milac		\$ -
			NEW IN. SIGNAL WITH GAUGES		Each		\$ -
			Rubber Railroad Crossing		Each		\$ -
			Upgrade Traffic Signal		Each		\$ -
			Traffic Signal (New)		Each		\$ -
			Traffic Control		Miles		\$ -
			Thermo and Markers		Miles		\$ -
			Structures				
			ML / Creek "Wx" 'L		SF		\$ -
			RC Box Culverts				
			Ex. 3@10x10-50'Extension-3'Fill-90Skew	50	LF		\$ -
			Utility Construction				
			Relocate Existing Water Line		LF		\$ -
			Relocate Existing Sewer Line		LF		\$ -
			Misc. & Mob (15% Strs&Util)				\$ -
			Misc. & Mob (45% Functional)				\$ -
Lgth ___ Miles							
Contract Cost							\$ -
E. & C. 15%							\$ -
Construction Cost							\$ -

Roadway % of STIP Est. Construction Cost

Roadway
\$ -

Structures % of STIP Est. Construction Cost

Strs & Util
\$ -

Figure 7.4 Excerpt from a Preliminary Estimate Worksheet

7.3 Inputs Based on Project Type

Recall that the project type (roadway or bridge) is specified on the Project Start Up screen (Figure 7.2). Based on the selected project type, two additional interface screens direct the user to input project specific information including scope, dimensional, environmental, and geographical parameters. The Help feature on these screens provides the user with parameter definitions and the mean value for each numerical parameter.

Figure 7.5 displays the Roadway Project Details screens and Help menus. Figure 7.6 shows the Bridge Project Details screens and Help menus.

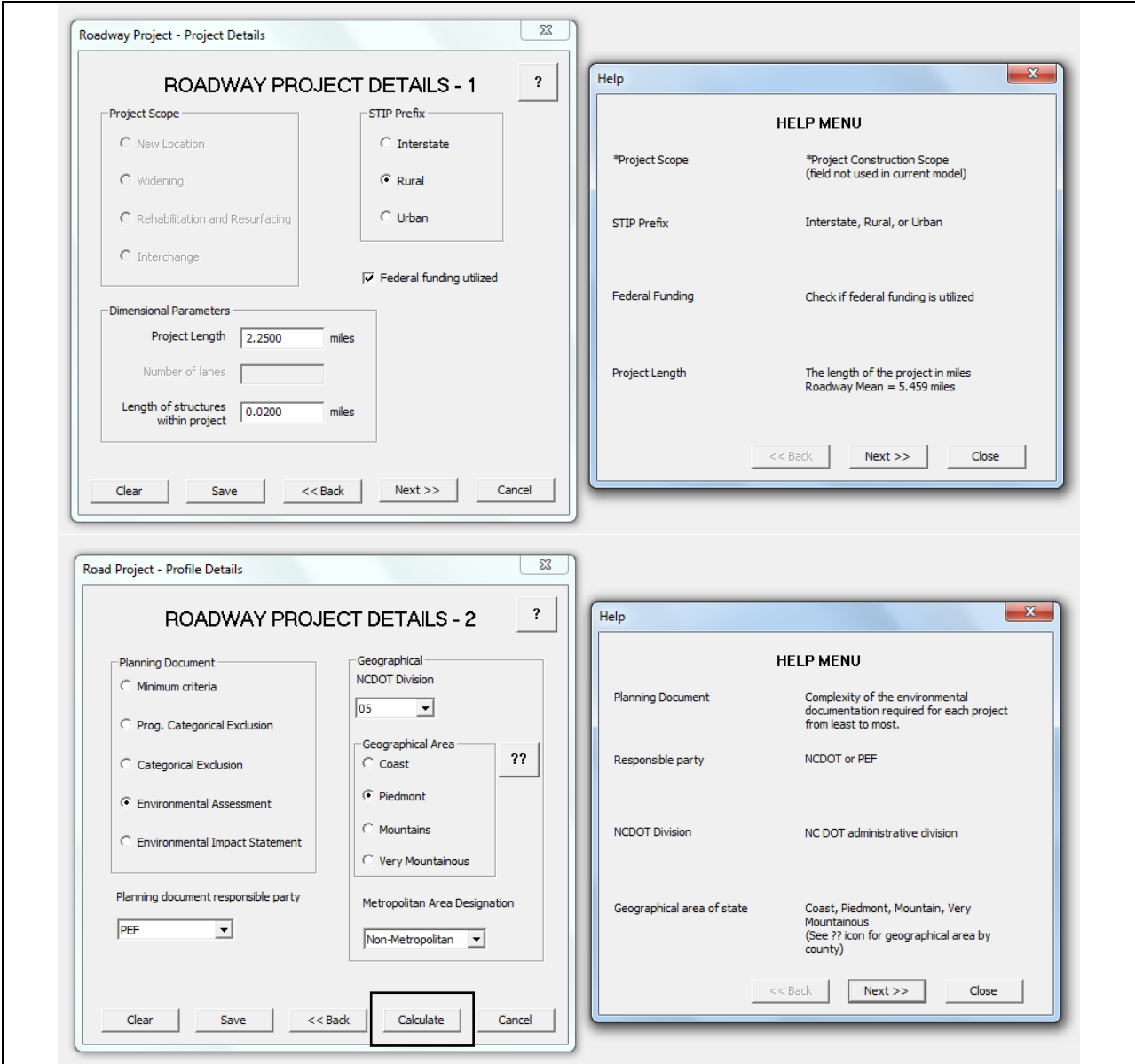


Figure 7.5 User Interface – Roadway Project Details

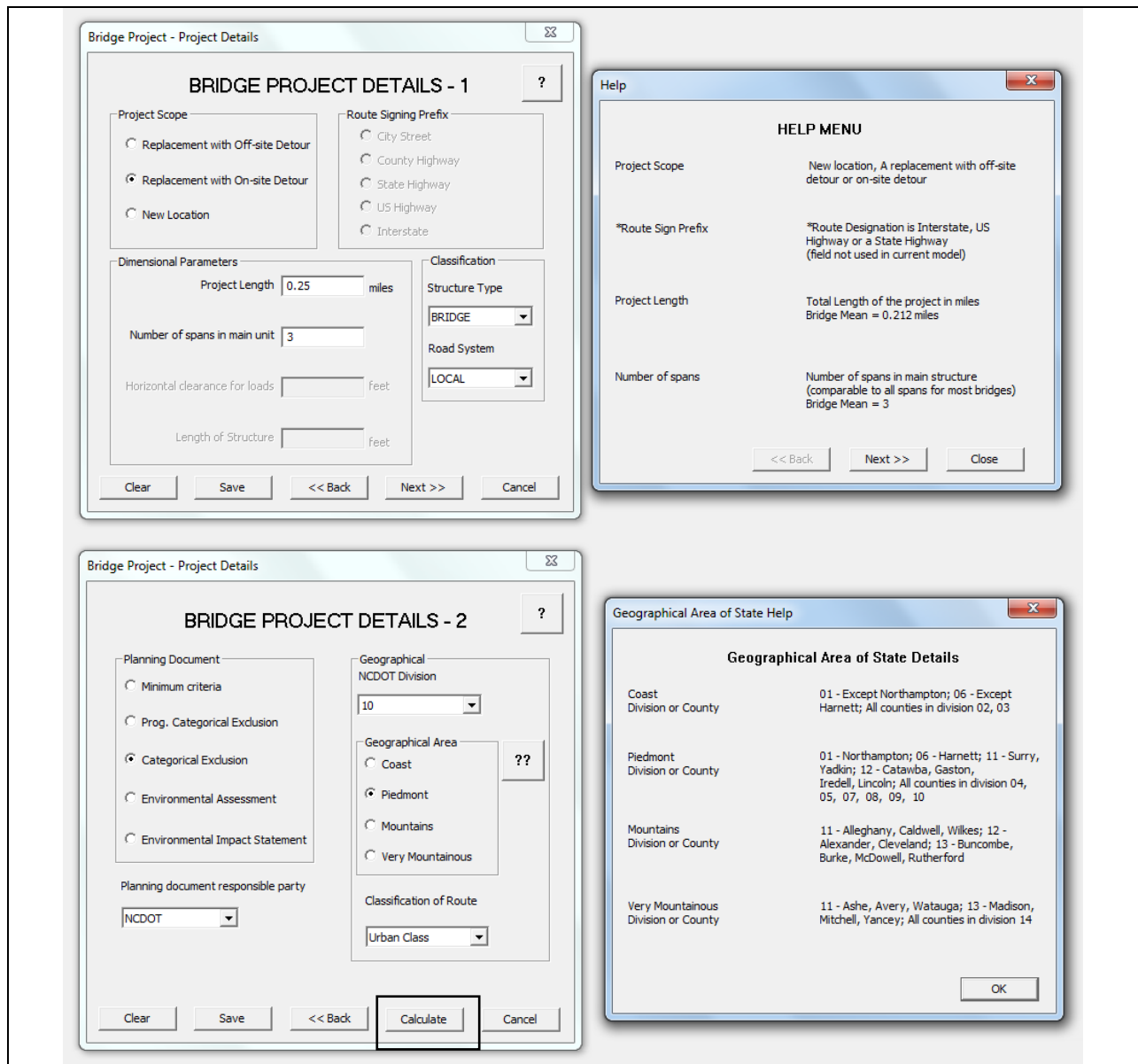


Figure 7.6 User Interface – Bridge Project Details

7.4 Estimate Results and Archived Summary Report

The second of the Project Details screens presents the user with the option to produce a PE cost ratio estimate by pressing the “CALCULATE” action button along the bottom of the screen. This action directs the interface to pass the user’s inputs to the appropriate modeling library. An estimate for a project’s PE cost ratio and PE cost (in \$) is generated. The estimate results are presented in an informational popup screen immediately after model processing. The popup screen shown in Figure 7.7 depicts how the user immediately receives estimate results.

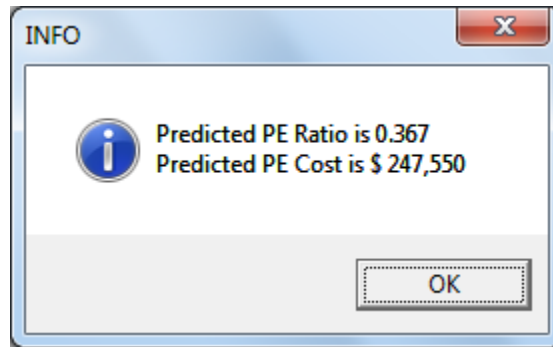


Figure 7.7 User Interface - Results Popup Screen

In addition, an archive text file is generated, date stamped, and saved in the tool’s directory. The archived text file presents a summary of the user inputs and resulting estimate based on project type selection and modeling process selection. Figure 7.8 displays a sample archive text file generated by the interface tool.

PROJECT SUMMARY

1. BASIC INFORMATION

Project Name: B-12345
Project Description: Sample Bridge
Project Type: Bridge Project
Prediction Model: Regression Model

2. COST DETAILS

Right of Way Portion Cost: (\$) 181000
STIP Estimated Construction Cost: (\$) 675000
Roadway % of STIP Est. Const. Cost: 0.314
Structure % of STIP Est. Const. Cost: 0.000

3. CLASSIFICATION

Project Scope: Replacement with Off-Site Detour
Road System: Arterial
Structure Type: Bridge

4. DIMENSIONAL PARAMETERS

Project Length(miles): 0.193 miles
Number of Spans: 3

5. ENVIRONMENTAL PARAMETERS

NEPA Document Classification: Categorical Exclusion
Planning Document Responsible Party: PEF

6. GEOGRAPHICAL PARAMETERS

NCDOT Division: 06
Geographical Area of State: Coast
Classification of Route: Rural Class

7. RESULTS

PE Cost Ratio Regression(proportion): 0.367
PE Cost Regression: \$247,550

Figure 7.8 User Interface - Archived Estimate Summary Report

8.0 FINDINGS AND CONCLUSIONS

Various modeling approaches were utilized in this research. This section summarizes the validation results of each predictive model. Based on our validation results, we incorporated the best predictive models into the user interface as described in section 7.

8.1 Modeling Results: Bridge Projects

8.1.1 MDPLM Results for Bridge PE Cost Ratio Prediction

Table 8.1 shows the model performance measures of the final bridge MDPLM for PE cost ratio. A mean absolute percentage error (MAPE) of 0.208 was achieved.

Table 8.1 MDPLM Performance for Final Bridge Model at Layer 4

Base Measure of Expected Value	Final Bridge Model
Mean square error (<i>MSE</i>)	31.624
Mean absolute error (<i>MAE</i>)	3.013
Mean absolute percentage error (<i>MAPE</i>)	0.208

As previously described in section 6.4, the PE cost ratio estimates generated by the final model (layer 4) are noticeably better than estimates produced in layers 0, 1, 2, or 3. Figure 8.1 presents this comparison graphically. The robustness of the estimated PE cost ratio is guaranteed by stochastic resonance theory.

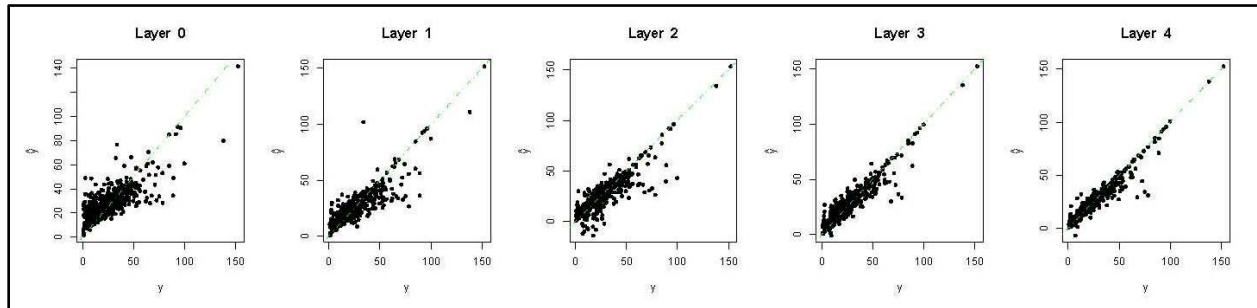


Figure 8.1 MDPLM (Bridge Model) Layer-wise Predictive Results
[y (actual) versus \hat{y} (estimate)]

The MDPLM was the best predictive model for the PE cost ratio of bridge projects. Users can access the final MDPLM model for future predictions through the user interface application. The MDPLM modeling routine is called upon for prediction computations based on the user's selections as described in section 7.2 of the final report.

8.1.2 Regression Results for PE Cost Ratio Prediction

MLR using the SAS GLMSELECT procedure was applied to the 391 bridge projects comprising the modeling dataset. The MLR analysis produced one viable model:

1. Reduced MLR model with year of letting omitted.

HLM analyses on the same group of 391 bridge projects yielded three additional modeling strategies:

2. HLM with 19 regression equations (one unique equation for each cell)
3. HLM with 13 regression equations (one unique equation for each good fit cell, and one of the thirteen equations used as a surrogate equation for all poor fit cells)

4. HLM with 13 regression equations (one unique equation for each good fit cell and the modeling mean used as a surrogate estimate for all poor fit cells)

Additionally, a single point estimate (using the mean predicted value) was compared to the four regression models. This single parameter estimate served to establish a baseline target for prediction capability of the other modeling efforts since NCDOT was accustomed to using a single parameter estimator for PE costs.

5. Single Point Estimate (using mean value)

All five of these modeling approaches were validated using the 70 bridge projects of the validation set. A complete listing of regression parameters for each model is available in section 12.1 of the appendices.

Table 8.2 shows the error analysis (in rank order) for the four regression models and the single point estimate, when applied to the validation set. The reduced MLR model provided the minimal MAPE among the five approaches. Note that MAPE values shown in Table 8.2 serve to rank the regression model approaches. For cost regression analyses, the response variable was the cubed root of PE cost ratio. The error assessment values (AE, AAE, and MAPE) are all being reported in terms of the response, cubed root of PE cost ratio. The MAPE values for the MDPLM model (section 8.1.1) are in terms of PE cost ratio without any transformation. Therefore, MAPE values are not directly comparable across different response variables.

Table 8.2 Performance of PE Cost Ratio Regression Models for Bridges

Model Description [Reference Table for model parameters]	Model Fit Assessment	Validation Error Assessment for Predicted Cubed Root of PE Ratio		
	Adjusted R ²	Average Error (AE)	Average Absolute Error (AAE)	Mean Absolute Percentage Error (MAPE)
Reduced MLR Model (with year of letting omitted) [Table 12.2]	0.2745	0.0112	0.0965	0.1889
Single Point Estimate (using mean value)	- N/A-	-0.0051	0.1043	0.1937
HLM with mean value as surrogate for poor fit cells [Table 12.5]	varies by cell	0.0041	0.1154	0.2157
HLM with surrogate equation for poor fit cells [Table 12.4]	varies by cell	0.0250	0.1247	0.2376
HLM [Table 12.3]	varies by cell	-0.0004	0.1376	0.2425

The best predictive regression model, the reduced MLR model, achieved a MAPE of 0.1889. The range (in MAPE) among the modeling approaches was 0.1889 to 0.2425. The user interface application includes the reduced MLR model as an option for PE cost ratio prediction using regression.

8.1.3 Regression Results for PE Duration Prediction

Three prediction modeling strategies were evaluated through testing on the validation sample of 60 bridge projects separated from the complete database. Table 8.3 shows the results from the validation exercise. The predicted mean PE duration of these projects was 68 months with a 95th percent confidence interval of 65.9 months to 70.1 months. Using the mean PE duration of the 356 bridge projects as the predicted

duration, the MAPE observed for the 60 validation projects was 0.2210. In comparison, the MAPE while using the HLM modeling procedure on the 60 validation projects was 0.2131. The error is reported in terms of MAPE since it is a generalized percentage that is easier to understand. A lower MAPE represents a lower percentage of error thus translating to a better prediction accuracy of the model in question.

Table 8.3 Performance of PE Duration Regression Models for Bridges

Model Description [Reference Table for model parameters]	Model Fit Assessment	Validation Error Assessment for Predicted PE Duration		
	Adjusted R ²	Average Error (AE)	Average Absolute Error (AAE)	Mean Absolute Percentage Error (MAPE)
HLM with surrogate equation for poor fit cells [Table 4.7]	Varies by cell (>0.6000)	-0.9491	15.118	0.2131
Single Point Estimate (using mean value)	- N/A-	-0.00008644	15.209	0.2210
MLR	0.0892	23	30	0.4542

Table 4.7 contains a complete listing of the regression parameters for the HLM model.

8.2 Modeling Results: Roadway Projects

8.2.1 DPLM Results for Roadway PE Cost Ratio Prediction

The error measurements of the final roadway DPLM trained with 181 project observations are shown in Table 8.4 and the plot of Figure 8.2 shows the PE cost ratio estimates over the 181 construction projects.

Table 8.4 DPLM Performance for Final Roadway Model

Base Measure of Expected Value	Final Roadway Model
Mean square error (<i>MSE</i>)	50.831
Mean absolute error (<i>MAE</i>)	4.271
Mean absolute percentage error (<i>MAPE</i>)	2.814

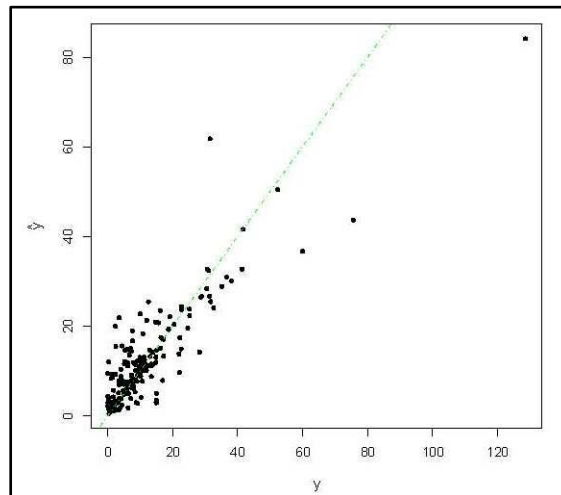


Figure 8.2 DPLM (Roadway Model) Predictive Results
[y (actual) versus \hat{y} (estimate)]

The DPLM model for the predicting the PE cost ratio of roadway projects is accessible through use of the user interface application.

8.2.2 Regression Results for PE Cost Ratio Prediction

Regression analyses on the 150 roadway projects of the modeling dataset, yielded five candidate modeling strategies:

1. MLR without interactions between selected variables
2. MLR with interactions between selected variables
3. HLM with no tier structure
4. HLM with 1 tier grouping
5. HLM with 2 tier grouping

As investigated for the bridge dataset, using the mean predicted value as the estimate was also included as a candidate strategy:

6. Single Point Estimate (using mean value)

We applied each of the six modeling strategies to the 38 roadway projects designated as the validation set. Table 8.5 displays the results of the model validation process. MAPE values from 0.2773 to 0.6474 were observed. The model with the least MAPE during validation was considered as the best model for predicting the PE cost ratio of future projects. The HLM model with no tier grouping provided the minimum MAPE of 0.2773.

Table 8.5 Performance of PE Cost Ratio Regression Models for Roadways

Model Description [Reference Table for model parameters]	Model Fit Assessment	Validation Error Assessment for Predicted Cubed Root of PE Ratio		
	Adjusted R ²	Average Error (AE)	Average Absolute Error (AAE)	Mean Absolute Percentage Error (MAPE)
HLM with no tiers [Table 12.7]	0.3054	-0.0022	0.0877	0.2773
MLR Model (without interactions) [Table 5.7]	0.5210	0.0516	0.1159	0.3457
HLM with 2 tiers [Table 5.11]	varies by cell	0.0301	0.1157	0.3981
Single Point Estimate (using mean value)	N/A	0.0082	0.1166	0.4036
MLR Model (with interactions) [Table 12.6]	0.5858	0.0316	0.1480	0.4513
HLM with 1 tier [Table 12.8]	0.3479 Coast 0.6553 Mtn 0.3407 Pdmt	0.0851	0.1657	0.6474

A complete listing of the regression parameters applicable to each model is available in the referenced table identified in the left-most column of Table 8.5.

8.2.3 Regression Results for PE Duration Prediction

Two regression strategies and a single point estimate (using the mean value) were evaluated to determine prediction performance. Each model was used to estimate the PE duration of the 23 roadway validation projects. Table 8.6 shows the prediction performance of each strategy. The MLR model performed best, minimizing MAPE. The MAPE values ranged from 0.1375 to 0.4841.

Table 8.6 Performance of PE Duration Regression Models for Roadways

Model Description [Reference Table for model parameters]	Model Fit Assessment	Validation Error Assessment for Predicted Cubed Root of PE Duration		
	Adjusted R ²	Average Error (AE)	Average Absolute Error (AAE)	Mean Absolute Percentage Error (MAPE)
MLR [Table 5.8]	0.7281	0.0779	0.4365	0.1375
Single Point Estimate (using mean value)	- N/A-	-0.0491	1.0338	0.3811
HLM [Table 5.14]	0.6140 Coast 0.7118 Mtn 0.6303 Pdmt	-0.1923	1.2114	0.4841

Tables 5.8 and 5.14 lists the complete regression parameters for the MLR model and HLM model respectively.

8.3 Conclusions

The bridge models performed better in prediction capability than the roadway models using the same modeling strategy except for one – the MLR approach applied to PE duration estimation. Table 8.7 lists the MAPE values between bridge models and roadways models for each modeling strategy. The strategies form the columns of Table 8.7 and the project type is arranged by row. For every column, except Duration MLR, the MAPE achieved by the bridge model is lower than the MAPE achieved by the roadway model. The breadth of the bridge database contributes to better modeling performance. Recall that bridge project data were more accessible and the number of bridge projects greatly exceeds the number of roadway projects for the same time period.

Table 8.7 Error Comparisons between Bridge and Roadway Models

Project Type	Comparison of Best MAPE Achieved by Modeling Approach				
	Cost MDPLM	Cost MLR	Cost HLM	Duration MLR	Duration HLM
Bridges	0.208	0.1889	0.2157	0.4542	0.2131
Roadways	2.814	0.3457	0.2773	0.1375	0.4841

MDPLM and DPLM Discussion

Estimating preliminary engineering cost ratio (simply PE cost ratio) is difficult because its characteristic is not fully understood. We assumed that construction project data are heterogeneous such that the relationships between PE cost ratio and other variables are complicated and there are more than one relationship existing. In order to make a predictive model for such heterogeneous data, we extended the Dirichlet process linear model (DPLM) to a multilevel model, called the multilevel Dirichlet process

linear model (MDPLM). Our proposed model can deal with more complex situations than existing regression techniques such as generalized linear mixed models [Breslow and Clayton 1993], multilevel mixed linear or nonlinear models [Goldstein 1986; Goldstein 1991], or Dirichlet process generalized linear models [Hannah et al. 2010; Mukhopadhyay and Gelfand 1997], and produce more robust estimates with less variance.

Unlike usual parametric models which requires the estimation of parameters, the MDPLM samples parameters according to their posterior distributions and generates competing sub-models (the sets of parameters) to be integrated for prediction. The prediction can be performed without estimating parameters by numerically integrating estimates of sub-models independently as like DPLM. Since parameters exist in the form of instances, it is difficult to understand how the covariate affects the response. In Dirichlet process prior mixture models such as DPLM and DPGLM a difficulty of parameter estimation comes from a so-called label switching [Hurn et al. 2000; Stephens 2000]. In a set of parameters sampled, switching mixture component labels between components does not affect the distribution of data and such a switch also occurs during parameter sampling. Since MDPLM has DPLM as the base model, it also suffers from label-switching. Even if parameters were properly estimated, interpretation or sensitivity analysis of model would not be straightforward since the MDPLM has a multilevel structure with nonlinearly generated random effects.

In spite of difficulty in interpretation or sensitivity analysis, the MDPLM is still competitive to other modeling techniques. The MDPLM is a convenient way of modeling in that it reduces human efforts. It is not necessary to describe how data are distributed or structured, nor specify the random effects. The MDPLM can fit complex data with less variance. By finding hidden or unobservable random effects, the reduction of variance in prediction is achieved. The prediction of MDPLM is robust. With help of cross-validation and stochastic resonance theory, the MDPLM avoids overfitting and results in better prediction for unseen data.

PE Cost Ratio Discussion

When budgeting for PE costs of highway projects, the rule of thumb that PE costs are 10% of estimated construction costs is not accurate if applied across all project types. The regression models developed provide an improved estimator over the use of a constant percentage value. Findings support the need for a separate model for bridge projects as compared to roadway projects.

Instances of extremely high or extremely low PE ratio were difficult to account for in regression modeling efforts. Obviously, some projects became problematic, driving PE costs up. Problems may be linked to public resistance, environmental impacts, or difficulty in right-of-way acquisition. However, our quantitative approach could not easily capture such risk factors for PE cost escalation. Similarly, other projects had PE costs that were exceptionally low when compared to other projects of similar size and scope. We were unable to discover the definitive causal factors for extreme PE cost ratios – either high or low.

PE Duration Discussion

Predicting PE duration is an important objective that has not attracted any past research attention. Hierarchical regression was the analysis tool used to derive the prediction model. Regression models within the tiered structure fit the data very well and the model was also easy to use. However, the validation of the hierarchical model showed that the model was only marginally more effective in predicting PE durations for projects in the validation data set than using the mean duration.

The analysis showed that predicting the duration of PE for highway projects is a very difficult problem. Conversations with responsible NCDOT engineers revealed several unquantifiable causes of prolonged

PE duration, including legal issues pertaining to right of way acquisition, unforeseen problems in utility installation, changes to project prioritization resulting in a halt of the ongoing projects, and funding reallocations leading to project freezes.

9.0 RECOMMENDATIONS

NCDOT personnel should utilize the user interface application to access the full range of PE prediction tools developed for bridge and roadway projects. Although the regression analyses are sufficiently documented to apply independently of the interface, the MDPLM analyses are complex and require a programming background to access directly. Therefore, the interface application is critical for successful implementation of the developed models.

9.1 Future Needs

Current NCDOT project data collection practices limit the precision of the PE cost ratio predictive models. Reasons for this limitation include:

- Record keeping procedures are not uniform across all management units.
- Reporting of direct charges for PE activities varies among personnel due to different motivations and lack of clarity in procedures.
- Strict institutional procedures were not in place during the project review period (1999-2008) and affected how PE costs were charged to specific projects. PE costs were often treated as an overhead burden that could not be accurately assigned to individual projects.

As administrative processes related to PE accounting practices become more consistent, improved PE databases are anticipated. Expanding the current databases is recommended. The modeling strategies should be updated periodically as the databases grow. The accompanying user interface application will also need periodic updating to utilize updated models.

9.2 Future Opportunities

- As more PE duration data becomes available, NCDOT may benefit from further analyses which would support adding a PE duration module to the user interface application.
- Our investigation of the detailed financial information acquired on a selection of STIP projects was limited under the current project scope. There may be additional value from a more detailed analysis of such information. A summary of our investigation is included in section 12.4 of the appendices.

10.0 IMPLEMENTATION AND TECHNOLOGY TRANSFER PLAN

We will facilitate transfer of our research efforts to NCDOT through development and delivery of the user interface application described in Section 7. We anticipate the primary NCDOT audience will be members of the Project Management Unit within the Program Development Branch. A training session involving the research team and intended NCDOT users (who have the interface application installed on their own computer) is recommended. Additionally, we would like to present the user interface application at appropriate events and meetings as recommended by the steering and implementation committee.

10.1 Research Products

This research effort has yielded the following research products:

- Predictive regression model for PE cost ratio for bridge projects
- Predictive regression model for PE duration for bridge projects
- Predictive MDPLM model for PE cost ratio for bridge projects
- Predictive regression model for PE cost ratio for roadway projects
- Predictive regression model for PE duration for roadway projects
- Predictive DPLM model for PE cost ratio for roadway projects
- User interface application for executing PE cost ratio models
- One conference presentation and publication in conference proceedings

Hollar, D. A., Arocho, I., Hummer, J., Liu, M., & Rasdorf, W. (2010). "Development of a Regression Model to Predict Preliminary Engineering Costs." Institute of Transportation Engineers 2010 Technical Conference and Exhibit, Savannah, Georgia. March 14 – 17, 2010.

- Two future conference presentations (with publication in conference proceedings) expected
"Predicting Preliminary Engineering Costs for Bridge Projects." 3rd International/9th Construction Specialty Conference, Ottawa, Ontario. June 14 – 17, 2011. (Accepted)
"Duration of Preliminary Engineering Activities for Bridge Projects." (Future conference yet to be determined.)
- Three future peer-reviewed journal papers expected
"Predictive Modeling for Preliminary Engineering Costs of Bridge Projects." (2011) Anticipated submission to the ASCE Journal of Construction Engineering and Management.
"Regression Modeling of Roadway Projects' Preliminary Engineering Cost and Duration." (2011) Anticipated submission to the ASCE Journal of Construction Engineering and Management.
"Estimating Preliminary Engineering Cost Ratio using Multilevel Dirichlet Process Linear Model." (2011) Anticipated submission to the IEEE Journal.

11.0 REFERENCES

- AACE International. (2003). "Cost Estimate Classification System" AACE International Recommended Practices, Morgantown, WV. 17R-97.
- 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(5), 653-661.
- Akintoye, A. (2000). "Analysis of Factors Influencing Project Cost Estimating Practice." *Construction Management and Economics*, 18(1), 77-89.
- Alavi, S. and Tavares, M. P. (2009). "Highway Project Cost Estimating and Management for the Montana Department of Transportation." FHWA/MT-08-007/8189
- Anderson, S., Molenaar, K., and Schexnayder, C. (2006). "Final Report for NCHRP Report 574: Guidance for Cost Estimation and Management for Highway Projects during Planning, Programming, and Preconstruction." NCHRP Web-Only Document 98, Transportation Research Board, <http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w98.pdf> (July 7, 2008)
- Anderson, S., Molenaar, K., and Schexnayder, C. (2007). *Guidance for Cost Estimation and Management for Highway Projects during Planning, Programming, and Preconstruction, NCHRP Report 574*, Transportation Research Board, Washington, D. C.
- Benzi, R., Parisi, G., Sutera, A., and Vulpiani, A. (1983). "A theory of stochastic resonance in climatic change." *SIAM Journal on Applied Mathematics*, 43(3), 565–578.
- Breslow, N. E. and Clayton, D. G. (1993). "Approximate inference in generalized linear mixed models." *Journal of the American Statistical Association*, 88(421), 9–25.
- Byrnes, J. E. (2002). *Best Practices for Highway Project Cost Estimating*. Master of Science Thesis, Arizona State University, December 2002.
- Chan, D. W. M., and Kumaraswamy, M. M. (2002). "Compressing Construction Durations: Lessons learned from Hong Kong building projects." *International Journal of Project Management*, 20, 23-35.
- Chen, C., and Li, K. (1998). "Can SIR be as Popular as Multiple Linear Regression?" *Statistica Sinica*, 8(1998), 289-316.
- Cohen, J. (1968). "Multiple Regression as a General Data-Analytic System." *Psychological Bulletin*, 70(6), 426-443.
- Cohen, R. A. (2006). "Introducing the GLMSELECT PROCEDURE for Model Selection." Proceedings of the Thirty-first Annual SAS Users Group International Conference, SAS Institute Inc., Cary, NC. Paper 207-31. <<http://www2.sas.com/proceedings/sugi31/207-31.pdf>>
- Cox, D. and Carroll, J. (2010). "Planning Level Cost Estimation – Georgia DOT's Innovative Processes and Procedures." Proceedings, TRB 89th Annual Meeting, Transportation Research Board, Washington, D.C., Paper 10-1109.
- FHWA (2003). "FHWA Division Responses to Cost Estimating Questions." Federal Highway Administration, Washington, D.C. <<http://construction.colorado.edu/NCHRP8-49/Modules/DetailDoc.aspx?DocID=39&mid=10>> (7/7/2008)
- Geddes, B. (1990). "How The Cases You Choose Affect The Answers You Get: Selection Bias In Comparative Politics." *Political Analysis*, 2, 131 – 150.

- Goldstein, H. (1986). "Multilevel mixed linear model analysis using iterative generalized least squares." *Biometrika*, 73(1), 43–56.
- Goldstein, H. (1991). "Nonlinear multilevel models, with an application to discrete response data." *Biometrika*, 78(1), 45–51.
- Gransberg, D. D., Lopez Del Puerto, Carla, Humphrey, D. (2007). "Relating Cost Growth from the Initial Estimate to Design Fee for Transportation Projects." *Journal of Construction Engineering and Management*, American Society of Civil Engineers, Reston, VA. 133(6), 404-408.
- Gross, K. (2008). "ST 512 Course Materials (Fall 2008)." North Carolina State University, <http://courses.ncsu.edu/st512/lec/001/#Course_notes> (August 20, 2008).
- Hamby, D. M. (1994). "A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models." *Environmental Monitoring and Assessment*, 32(2), 135-154.
- Hannah, L. A., Blei, D. M., and Powell, W. B. (2010). "Dirichlet process mixtures of generalized linear models."
- Hecht, H., and Niemeier, D. (2002). "Too Cautious? Avoiding Risk in Transportation Project Development." *Journal of Infrastructure Systems*, American Society of Civil Engineers, Reston, VA. 8(1), 20-28.
- Hendren, P., and Niemeier, D. A. (2006). "Evaluating the Effectiveness of State Department of Transportation Investment Decisions Linking Performance Measures to Resource Allocation." *Journal of Infrastructure Systems*, American Society of Civil Engineers, Reston, VA. 12(4), 216-229.
- Hoffman, G.J. et. al. (2007). "Estimating Performance Time for Construction Projects." *Journal of Management in Engineering*, American Society of Civil Engineers, Reston, VA. 23(4), 193-199.
- Hurn, M., Justel, A., Robert, C. P., and Roberty, C. P. (2000). "Estimating mixtures of regressions."
- Isaacs, B., Saito, M., and McKnight, C. E. (1998). "Notice Ratings as a Management Tool to Reduce Truck Rollovers on Ramps." *Journal of Infrastructure Systems*, 4(4), 156 – 163.
- Kim, J., and Mueller, C. (1978). *Factor Analysis: Statistical Methods and Practical Issues*. Sage Publications, Inc., Beverly Hills, CA.
- King, G., Keohane, R.O., and Verba, S. (1994). "Designing Social Inquiry: Scientific Inference in Qualitative Research." *Princeton*. Princeton University Press.
- Knight, K. and Fayek, A. R. (2002). "Use of Fuzzy Logic for Predicting Design Cost Overruns on Building Projects." *Journal of Construction Engineering and Management*, American Society of Civil Engineers, Reston, VA. 128(6), 503-512.
- Kyte, C. A., Perfater, M. A., Haynes, S., and Lee, H. W. (2004a). "Developing and Validating a Highway Construction Project Cost Estimation Tool." Virginia Transportation Research Council, Report 05-R1, <<http://vtrc.virginia.gov/PubDetails.aspx?PubNo=05-R1>> (August 20, 2008).
- Kyte, C. A., Perfater, M. A., Haynes, S., and Lee, H. W. (2004b). "Developing and Validating a Tool to Estimate Highway Construction Project Costs." *Transportation Research Record*, Transportation Research Board of the National Academies, Washington, DC. (1885), 35-41.
- Lam, E. W. M., Chan, A. P. C., Chan, D. W. M. (2008). "Determinants of Successful Design-Build Projects." *Journal of Construction Engineering and Management*, 134(5), 333-341.
- Lane, D., Davenport, R., and Garris, R. (2008). Meeting with NCDOT Project Services Group held August 14, 2008. Raleigh, NC.

- Lowe, D. J., Emsley, M. W., Harding, A. (2006). "Predicting Construction Cost using Multiple Regression Techniques." *Journal of Construction Engineering and Management*, American Society of Civil Engineers, Reston, VA. 132(7), 750-758.
- Maas, C. J. M., and Hox, J. J. (2005). "Sufficient Sample Sizes for Multilevel Modeling." Utrecht University, The Netherlands, *Methodology*, 1(3), 86-92.
- Mason, C. H., and Perrault Jr., W.D. (1991). "Collinearity, Power, and Interpretation of Multiple Regression Analysis." *Journal of Marketing Research*, American Marketing Association, 28(3), 268-280.
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. 2nd Edition, Chapman and Hall, London.
- Mckinsey & Company (2007). "Laying the Foundation for a Successful Transformation." *North Carolina Department of Transportation*, 1-22.
- McNamara, B. and Wiesenfeld, K. (1989). "Theory of stochastic resonance." *Phys. Rev. A*, 39(May), 4854-4869.
- Merritt, Leslie W. Jr. (2008). "Performance Audit. Department of Transportation - Highway Project Schedules and Costs (Audit No. PER-2007-7229)." Office of the State Auditor, State of North Carolina. <<http://www.ncauditor.net/EPSSWeb/Reports/Performance/PER-2007-7229.pdf>> (August 5, 2008).
- Mukhopadhyay, S. and Gelfand, A. E. (1997). "Dirichlet process mixed generalized linear models." *Journal of the American Statistical Association*, 92(438), 633-639.
- Nassar, K. M., Hegab, M. Y., Jack, N. W. (2005). "Design Cost Analysis of Transportation Projects." *Proceedings, Construction Research Congress 2005: Broadening Perspectives - Proceedings of the Congress*, American Society of Civil Engineers, Reston, VA, United States, 949-956.
- NCDOT. (2008a). *North Carolina Department of Transportation State Transportation Improvement Program 2009-2015*, North Carolina Department of Transportation, <<http://www.ncdot.org/planning/development/TIP/TIP/>> (August 8, 2008).
- NCDOT. (2008b). "Webpage: How a Road Gets Built." North Carolina Department of Transportation, <<http://www.ncdot.org/projects/roadbuilt/road.html>> (September 23, 2008).
- NCDOT. (2008c). "Webpage: NCDOT: Dashboard – Delivery Rate." North Carolina Department of Transportation, <<https://apps.dot.state.nc.us/dot/dashboard/DeliveryRate.aspx>> (November 21, 2008).
- NCDOT. (2008d). "Webpage: NCDOT: Finance and Budget." North Carolina Department of Transportation, <<http://www.ncdot.gov/about/finance/>> (November 21, 2008).
- NCDOT. (2008e). "Webpage: NCDOT: Organizational Performance Dashboard." North Carolina Department of Transportation, <<http://www.ncdot.org/programs/dashboard/>> (November 21, 2008).
- NCDOT. (2008f). "Webpage: NCDOT: PDEA Merger 01 Process." North Carolina Department of Transportation, <<http://www.ncdot.org/doh/preconstruct/pe/MERGER01/>> (September 17, 2008).
- NCDOT. (2008g). "Webpage: Project Letting Page." North Carolina Department of Transportation, <<http://www.ncdot.org/doh/preconstruct/ps/contracts/letting.html>> (September 15, 2008).
- NCDOT. (2008h). "Webpage: Statistics for Money Spent on Construction." North Carolina Department of Transportation, <<http://www.ncdot.org/programs/dashboard/content/download/AwardsSummary.pdf>> (August 8, 2008).

- Oberlender, G. D., and Trost, S. M. (2001). "Predicting Accuracy of Early Cost Estimates Based on Estimate Quality." *Journal of Construction Engineering and Management*, 127(3), 173-182.
- Odeck, J. (2003). "Cost Overruns in Road Construction – What are Their Sizes and Determinants?" *Transport Policy*, World Conference on Transport Research Society, Lyon, France. 11(2004), 43-53.
- PDEA. (2008). "Meeting with PDEA Eastern Branch Members Hanson, R., Cox, C., McInnis, J., and Yamamoto, B." North Carolina Department of Transportation, Highway Building, Raleigh, NC. September 19, 2008.
- Picard, R. R. and Cook, D. R. (1984). "Cross Validation of Regression Models." *Journal of the American Statistical Association*, 79(387), 575-583.
- Robertson, H. D., Hummer, J. E., and Nelson, D. C. (2000). *Manual of Transportation Engineering Studies*, Institute of Transportation Engineers, Washington D.C.
- Sakia, R. M. (1992). "The Box-Cox Transformation Technique: A Review." *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(2), 169-178.
- Schexnayder, C. J., Weber, S. L., and Fiori C. (2003). "Project Cost Estimating – A Synthesis of Highway Practice." NCHRP Project 20-7, Task 152, Transportation Research Board, Washington, D.C.
- Sheiner, L. B. and Beal, S. L. (1981). "Some Suggestions for Measuring Predictive Performance." *Journal of Pharmacokinetics and Biopharmaceutics*, Springer Healthcare Communications Ltd, New York, NY. 9(4), 503-512.
- Shtatland, E. S. and Barton, M. B. (1997). "Information as a Unifying Measure of Fit in SAS Statistical Modeling Procedures." *Northeast SAS Users Group NESUG '97 Proceedings*, Baltimore, MD. 875-880.
- SLNC. (2008). "Webpage: State Library Catalogs." State Library of North Carolina. <<http://statelibrary.dcr.state.nc.us/catalog.html>> (October 7, 2008).
- Snijders, T. A. B. (2005). "Power and Sample Size in Multilevel Linear Models." *Encyclopedia of Statistics in Behavioral Science*, 3, 1570–1573.
- Steenbergen, M. R. and Jones, B. S. (2002). "Modeling Multilevel Data Structures." *American Journal of Political Science*, 46(1), 218-237.
- Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society, Ser. B*, 62, 795–809.
- Taylor, J. M. G. (1986). "The Retransformed Mean after a Fitted Power Transformation." *Journal of the American Statistical Association*, 81(393), 114-118.
- Trost, S. M. (1998). *A Quantitative Model for Predicting the Accuracy of Early Cost Estimates for Construction Projects in the Process Industry*, PhD Dissertation, Oklahoma State University, Stillwater, OK.
- Tu, J. (1996). "Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes." *Journal of Clinical Epidemiology*, 49(11), 1225-1231.
- Turochy, R. E., Hoel, L. A., Doty, R. S. (2001). *Highway Project Cost Estimating Methods used in the Planning Stage of Project Development*, Virginia Transportation Research Council, Technical Assistance Report 02-TAR3, Charlottesville, VA.
- Wedderburn, R. W. M. (1974). "Quasi-likelihood functions, generalized linear models, and the gauss-newton method." *Biometrika*, 61(3), 439–447.

Wilmot, C. G., Deis, D. R., Schneider, H., Coates, C. H., Jr. (1999). "In-House Versus Consultant Design Costs in State Departments of Transportation." *Transportation Research Record*, (1654), 153-160.

WSDOT. (2002). "Highway Construction Cost Comparison Survey." Washington State Department of Transportation, April 2002. <http://www.wsdot.wa.gov/biz/construction/pdf/I-C_Const_Cost.pdf> (July 7, 2008)

12.0 APPENDICES

Contents

- 12.1 Regression Results for Bridge PE Cost Models
- 12.2 Regression Results for Roadway PE Cost Models
- 12.3 Sample Functional Cost Estimate Worksheet
- 12.4 Analyses of Financial Details

12.1 Regression Results for Bridge PE Cost Ratio Models

Tables contained in this section provide details on the regression parameters generated through MLR and HLM modeling efforts for bridge PE cost ratio models. The cubed root of PE cost ratio is the response variable for all models.

Table 12.1 Full Bridge MLR Model Definition

Parameter	Estimate	StdErr	tValue	Probt
Intercept	0.621706	0.010111	61.48901	2.4E-195
CY_2006	-0.23113	0.027425	-8.42761	8.12E-16
CY_2007	0.180376	0.016156	11.16447	4.09E-25
CY_2001*DIV_D06	-0.15398	0.038023	-4.04954	6.26E-05
CY_2002*DIV_D11	-0.20854	0.041561	-5.01759	8.17E-07
CY_2005*DIV_D13	-0.29765	0.040262	-7.39272	9.83E-13
CY_2006*DIV_D13	-0.22904	0.056269	-4.07045	5.75E-05
CY_2007*DIV_D12	-0.099	0.040645	-2.43566	0.01534
GEO_AREA_V MTN	0.08364	0.012929	6.468947	3.16E-10
DIV_D05*B_SCOPE_B_EX_OFF	0.063914	0.016409	3.895147	0.000117
DIV_D12*B_SCOPE_B_EX_ON	-0.17209	0.040694	-4.22886	2.97E-05
CY_2002*PLAN_RESP_DOT	0.065328	0.016939	3.85678	0.000136
CY_2006*PLAN_RESP_PEF	0.125381	0.036911	3.396813	0.000756
DIV_D01*PLAN_RESP_PEF	0.132886	0.033936	3.915816	0.000107
B_SCOPE_B_NEW*PLAN_RESP_DOT	0.033752	0.012608	2.677151	0.007758
ROW_RATIO	0.271869	0.038866	6.995092	1.26E-11
ROW_RATIO*CY_2004	-0.27134	0.089006	-3.04853	0.002466
ROW_RATIO*DIV_D04	0.27367	0.10883	2.514646	0.012342
RW*CY_2003	0.210023	0.039685	5.292187	2.08E-07
RW*CY_2008	0.305638	0.032636	9.364925	7.86E-19
RW*CY_2009	0.357045	0.069688	5.123514	4.86E-07
RW*DIV_D13	0.198851	0.042501	4.67874	4.07E-06
TIP_COST*DIV_D12	5.84E-08	7.27E-09	8.033824	1.3E-14
RW*TIP_COST	-2.4E-07	1.53E-08	-15.5058	4.61E-42

Table 12.2 Reduced Bridge MLR Model with Year of Letting Omitted

Parameter	Estimate	StdErr	tValue	Probt
Intercept	0.674139	0.01277	52.78925	7.5E-178
DIV_D12*B_SCOPE_B_EX_ON	-0.16571	0.05978	-2.77201	0.005843
DIV_D06*PLAN_RESP_DOT	-0.10866	0.036463	-2.97998	0.003067
GEO_AREA_V MTN*PLAN_RESP_DOT	0.070104	0.026785	2.617241	0.009216
ROW_RATIO	0.290892	0.054251	5.361938	1.43E-07
TIP_COST*DIV_D12	4.45E-08	1.09E-08	4.088008	5.3E-05
RW*TIP_COST	-1.9E-07	2.21E-08	-8.53243	3.38E-16
N19_DET_LEN_KM*DIV_D07	-0.01588	0.005455	-2.91034	0.003821

Table 12.3 Bridge HLM Model Definition

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7			
		Coast Off-Site Det DOT	Coast Off-Site Det PEF	Coast On-Site Det	Coast New Loc
Intercept		0.4176	0.4658	1.1235	0.7583
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO	-2.0239	-0.2845	-3.5158	
STIP Estimated Construction Cost	TIP_COST		-7.18E-08	-2.48E-07	
Project Length	LEN	-0.7072	-3.8328	0.4128	-0.4786
Bypass Detour Length	N19_DET_LEN_KM			-0.0098	
Roadway Percentage of Construction Cost	RW		0.4055	-1.0825	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN		0.0284		
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M	0.0111	0.0134	0.0062	
Design Type	n43b		0.0258	0.0221	
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22	6.7934			
Quadratic form - STIP Estimated Construction Cost	tip22				
Quadratic Form – Project Length	len22		6.0923		
Quadratic Form – Roadway Percentage of Construction Cost	rw22				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det			0.3360	
Interaction – STIP Estimated Construction Cost with Project Length	tip2len				
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw				
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw				

Table 12.3 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7				
		Mountain Off-Site Det DOT	Mountain Off-Site Det PEF	Mountain On-Site Det	Mountain New Loc DOT	Mountain New Loc PEF
Intercept		-8.2468	-1.7809	0.9595	-1.4172	1.2102
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO	41.2536	1.3802	-0.1658		5.1669
STIP Estimated Construction Cost	TIP_COST	1.30E-05	1.09E-06	-1.52E-07	7.97E-08	
Project Length	LEN	5.1245	8.0597		-11.8960	2.3055
Bypass Detour Length	N19_DET_LEN_KM	-0.2066	0.0365	0.0150	0.0070	-0.0096
Roadway Percentage of Construction Cost	RW	6.2607	0.4555	-0.3829	10.4362	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN	-1.1218	0.0767		0.1694	-0.1660
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M	-0.0282	0.0397		0.0150	-0.0256
Design Type	n43b	0.1067			-0.0865	-0.0084
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22					
Quadratic form - STIP Estimated Construction Cost	tip22	-4.98E-12				
Quadratic Form – Project Length	len22					
Quadratic Form – Roadway Percentage of Construction Cost	rw22					
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len					-17.8440
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det					
Interaction – STIP Estimated Construction Cost with Project Length	tip2len		-1.15E-05			
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw					
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw					

Table 12.3 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7					
		Piedmont Off-Site DOT	Piedmont Off-Site PEF	Piedmont On-Site DOT	Piedmont On-Site PEF	Piedmont New Loc DOT	Piedmont New Loc PEF
Intercept		0.7488	0.7704	-0.5328	0.6206	0.6933	0.7952
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO						
STIP Estimated Construction Cost	TIP_COST	-1.99E-08	-2.24E-07	6.96E-07		-5.95E-08	
Project Length	LEN	-0.6944		0.3034	-0.3274		-0.8226
Bypass Detour Length	N19_DET_LEN_KM			-0.0020			-0.0156
Roadway Percentage of Construction Cost	RW	-0.2168	-0.5540	2.5953		1.6378	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN	0.0192	-0.0421		-0.0155	-0.0519	-0.1045
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M		0.0083	0.0054			0.0082
Design Type	n43b		0.0049	-0.0133			
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22				2.3367		
Quadratic form - STIP Estimated Construction Cost	tip22						4.55E-14
Quadratic Form – Project Length	len22						
Quadratic Form – Roadway Percentage of Construction Cost	rw22					-2.7718	
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len	4.2371					
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det		0.2231				
Interaction – STIP Estimated Construction Cost with Project Length	tip2len						
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw			-2.15E-06			
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw						

Table 12.3 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7			
		Very Mtn Off-Site Det	Very Mtn On-Site Det	Very Mtn New Loc DOT	Very Mtn New Loc PEF
Intercept		1.5996	1.5348	0.7018	-0.6644
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO		0.9645	-4.1488	0.6287
STIP Estimated Construction Cost	TIP_COST		-5.64E-07	-1.12E-07	1.17E-07
Project Length	LEN	-0.7495	3.3434		6.8212
Bypass Detour Length	N19_DET_LEN_KM		-0.0109	-0.0214	0.0124
Roadway Percentage of Construction Cost	RW	-3.6020	-0.4270		3.3410
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN	-0.0298		0.0415	
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M	-0.0062	-0.0284	0.0091	
Design Type	n43b	-0.0086		0.0103	0.0423
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22			16.0689	
Quadratic form - STIP Estimated Construction Cost	tip22		7.26E-14		
Quadratic Form – Project Length	len22				
Quadratic Form – Roadway Percentage of Construction Cost	rw22	6.0211			
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det				
Interaction – STIP Estimated Construction Cost with Project Length	tip2len				
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw				
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw				-21.1948

Table 12.4 Bridge HLM (with Surrogate) Model Definition

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7			
		Coast Off-Site Det DOT	Coast Off-Site Det PEF	Coast On-Site Det	Coast New Loc
Intercept		0.7583	0.4658	1.1235	0.7583
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO		-0.2845	-3.5158	
STIP Estimated Construction Cost	TIP_COST		-7.18E-08	-2.48E-07	
Project Length	LEN	-0.4786	-3.8328	0.4128	-0.4786
Bypass Detour Length	N19_DET_LEN_KM			-0.0098	
Roadway Percentage of Construction Cost	RW		0.4055	-1.0825	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN		0.0284		
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M		0.0134	0.0062	
Design Type	n43b		0.0258	0.0221	
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22				
Quadratic form - STIP Estimated Construction Cost	tip22				
Quadratic Form – Project Length	len22		6.0923		
Quadratic Form – Roadway Percentage of Construction Cost	rw22				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det			0.3360	
Interaction – STIP Estimated Construction Cost with Project Length	tip2len				
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw				
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw				


 Shading indicates poor fit cells. The Coast-New Location model parameters were used as a surrogate for all poor fit cells.

Table 12.4 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7				
		Mountain Off-Site Det DOT	Mountain Off-Site Det PEF	Mountain On-Site Det	Mountain New Loc DOT	Mountain New Loc PEF
Intercept		-8.2468	-1.7809	0.9595	-1.4172	1.2102
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO	41.2536	1.3802	-0.1658		5.1669
STIP Estimated Construction Cost	TIP_COST	1.30E-05	1.09E-06	-1.52E-07	7.97E-08	
Project Length	LEN	5.1245	8.0597		-11.8960	2.3055
Bypass Detour Length	N19_DET_LEN_KM	-0.2066	0.0365	0.0150	0.0070	-0.0096
Roadway Percentage of Construction Cost	RW	6.2607	0.4555	-0.3829	10.4362	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN	-1.1218	0.0767		0.1694	-0.1660
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M	-0.0282	0.0397		0.0150	-0.0256
Design Type	n43b	0.1067			-0.0865	-0.0084
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22					
Quadratic form - STIP Estimated Construction Cost	tip22	-4.98E-12				
Quadratic Form – Project Length	len22					
Quadratic Form – Roadway Percentage of Construction Cost	rw22					
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len					-17.8440
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det					
Interaction – STIP Estimated Construction Cost with Project Length	tip2len		-1.15E-05			
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw					
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw					


 Shading indicates poor fit cells. The Coast-New Location model parameters were used as a surrogate for all poor fit cells.

Table 12.4 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7					
		Piedmont Off-Site DOT	Piedmont Off-Site PEF	Piedmont On-Site DOT	Piedmont On-Site PEF	Piedmont New Loc DOT	Piedmont New Loc PEF
Intercept		0.7583	0.7583	0.7583	0.7583	0.6933	0.7952
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO						
STIP Estimated Construction Cost	TIP_COST					-5.95E-08	
Project Length	LEN	-0.4786	-0.4786	-0.4786	-0.4786		-0.8226
Bypass Detour Length	N19_DET_LEN_KM						-0.0156
Roadway Percentage of Construction Cost	RW					1.6378	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN					-0.0519	-0.1045
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M						0.0082
Design Type	n43b						
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22						
Quadratic form - STIP Estimated Construction Cost	tip22						4.55E-14
Quadratic Form – Project Length	len22						
Quadratic Form – Roadway Percentage of Construction Cost	rw22					-2.7718	
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len						
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det						
Interaction – STIP Estimated Construction Cost with Project Length	tip2len						
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw						
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw						


 Shading indicates poor fit cells. The Coast-New Location model parameters were used as a surrogate for all poor fit cells.

Table 12.4 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7			
		Very Mtn Off-Site Det	Very Mtn On-Site Det	Very Mtn New Loc DOT	Very Mtn New Loc PEF
Intercept		1.5996	1.5348	0.7583	-0.6644
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO		0.9645		0.6287
STIP Estimated Construction Cost	TIP_COST		-5.64E-07		1.17E-07
Project Length	LEN	-0.7495	3.3434	-0.4786	6.8212
Bypass Detour Length	N19_DET_LEN_KM		-0.0109		0.0124
Roadway Percentage of Construction Cost	RW	-3.6020	-0.4270		3.3410
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN	-0.0298			
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M	-0.0062	-0.0284		
Design Type	n43b	-0.0086			0.0423
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22				
Quadratic form - STIP Estimated Construction Cost	tip22		7.26E-14		
Quadratic Form – Project Length	len22				
Quadratic Form – Roadway Percentage of Construction Cost	rw22	6.0211			
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det				
Interaction – STIP Estimated Construction Cost with Project Length	tip2len				
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw				
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw				-21.1948


 Shading indicates poor fit cells. The Coast-New Location model parameters were used as a surrogate for all poor fit cells.

Table 12.5 Bridge HLM (with Mean as Surrogate) Model Definition

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7			
		Coast Off-Site Det DOT	Coast Off-Site Det PEF	Coast On-Site Det	Coast New Loc
Intercept		0.6176	0.4658	1.1235	0.7583
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO		-0.2845	-3.5158	
STIP Estimated Construction Cost	TIP_COST		-7.18E-08	-2.48E-07	
Project Length	LEN		-3.8328	0.4128	-0.4786
Bypass Detour Length	N19_DET_LEN_KM			-0.0098	
Roadway Percentage of Construction Cost	RW		0.4055	-1.0825	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN		0.0284		
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M		0.0134	0.0062	
Design Type	n43b		0.0258	0.0221	
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22				
Quadratic form - STIP Estimated Construction Cost	tip22				
Quadratic Form – Project Length	len22		6.0923		
Quadratic Form – Roadway Percentage of Construction Cost	rw22				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det			0.3360	
Interaction – STIP Estimated Construction Cost with Project Length	tip2len				
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw				
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw				


 Shading indicates poor fit cells. The mean cubed root of PE cost ratio was used as a surrogate for all poor fit cells.

Table 12.5 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7				
		Mountain Off-Site Det DOT	Mountain Off-Site Det PEF	Mountain On-Site Det	Mountain New Loc DOT	Mountain New Loc PEF
Intercept		-8.2468	-1.7809	0.9595	-1.4172	1.2102
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO	41.2536	1.3802	-0.1658		5.1669
STIP Estimated Construction Cost	TIP_COST	1.30E-05	1.09E-06	-1.52E-07	7.97E-08	
Project Length	LEN	5.1245	8.0597		-11.8960	2.3055
Bypass Detour Length	N19_DET_LEN_KM	-0.2066	0.0365	0.0150	0.0070	-0.0096
Roadway Percentage of Construction Cost	RW	6.2607	0.4555	-0.3829	10.4362	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN	-1.1218	0.0767		0.1694	-0.1660
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M	-0.0282	0.0397		0.0150	-0.0256
Design Type	n43b	0.1067			-0.0865	-0.0084
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22					
Quadratic form - STIP Estimated Construction Cost	tip22	-4.98E-12				
Quadratic Form – Project Length	len22					
Quadratic Form – Roadway Percentage of Construction Cost	rw22					
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len					-17.8440
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det					
Interaction – STIP Estimated Construction Cost with Project Length	tip2len		-1.15E-05			
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw					
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw					


 Shading indicates poor fit cells. The mean cubed root of PE cost ratio was used as a surrogate for all poor fit cells.

Table 12.5 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7					
		Piedmont Off-Site DOT	Piedmont Off-Site PEF	Piedmont On-Site DOT	Piedmont On-Site PEF	Piedmont New Loc DOT	Piedmont New Loc PEF
Intercept		0.6176	0.6176	0.6176	0.6176	0.6933	0.7952
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO						
STIP Estimated Construction Cost	TIP_COST					-5.95E-08	
Project Length	LEN						-0.8226
Bypass Detour Length	N19_DET_LEN_KM						-0.0156
Roadway Percentage of Construction Cost	RW					1.6378	
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN					-0.0519	-0.1045
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M						0.0082
Design Type	n43b						
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22						
Quadratic form - STIP Estimated Construction Cost	tip22						4.55E-14
Quadratic Form – Project Length	len22						
Quadratic Form – Roadway Percentage of Construction Cost	rw22					-2.7718	
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len						
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det						
Interaction – STIP Estimated Construction Cost with Project Length	tip2len						
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw						
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw						



 Shading indicates poor fit cells. The mean cubed root of PE cost ratio was used as a surrogate for all poor fit cells.

Table 12.5 (Continued)

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Comparison by Tier Grouping Associated with Table 3.7			
		Very Mtn Off-Site Det	Very Mtn On-Site Det	Very Mtn New Loc DOT	Very Mtn New Loc PEF
Intercept		1.5996	1.5348	0.6176	-0.6644
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO		0.9645		0.6287
STIP Estimated Construction Cost	TIP_COST		-5.64E-07		1.17E-07
Project Length	LEN	-0.7495	3.3434		6.8212
Bypass Detour Length	N19_DET_LEN_KM		-0.0109		0.0124
Roadway Percentage of Construction Cost	RW	-3.6020	-0.4270		3.3410
Number of Spans in Main Unit	N45_NUM_MAIN_SPAN	-0.0298			
Horizontal Clearance for Loads	N47_TOT_HOR_CLR_M	-0.0062	-0.0284		
Design Type	n43b	-0.0086			0.0423
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22				
Quadratic form - STIP Estimated Construction Cost	tip22		7.26E-14		
Quadratic Form – Project Length	len22				
Quadratic Form – Roadway Percentage of Construction Cost	rw22	6.0211			
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Project Length	row2len				
Interaction – Right of Way Cost to STIP Estimated Construction Cost with Bypass Detour Length	row2det				
Interaction – STIP Estimated Construction Cost with Project Length	tip2len				
Interaction – STIP Estimated Construction Cost with Roadway Percentage of Construction Cost	tip2rw				
Interaction – Project Length with Roadway Percentage of Construction Cost	len2rw				-21.1948

 Shading indicates poor fit cells. The mean cubed root of PE cost ratio was used as a surrogate for all poor fit cells.

12.2 Regression Results for Roadway PE Cost Ratio Models

Tables contained in this section provide details on the regression parameters generated through MLR and HLM modeling efforts for various roadway PE cost ratio models. The cubed root of PE cost ratio is the response variable for all models.

Table 12.6 Roadway MLR (with Interactions) Model Definition

Parameter	Estimate	StdErr	tValue	Probt
Intercept	0.5196	0.0233	22.2605	0.0000
NEPA_DOC_EA*PLAN_RESP_PEF	-0.1383	0.0556	-2.4882	0.0144
R_SCOPE_OTHER	-0.0817	0.0319	-2.5597	0.0119
LEN*R_SCOPE_R_WIDE	-0.0353	0.0078	-4.5037	0.0000
ROW_RATIO*R_SCOPE_R_WIDE	-0.1192	0.0423	-2.8170	0.0058
FED_FUND*GEO_AREA_PDMT	0.0522	0.0220	2.3789	0.0192
ROW_RATIO*FED_FUND	0.2319	0.0490	4.7295	0.0000
RW*R_SCOPE_R_RRR	-0.1713	0.0337	-5.0881	0.0000
LEN*NUM_LANES	-0.0015	0.0004	-3.6452	0.0004
METRO_AREA*NEPA_DOC_CE	-0.1823	0.0374	-4.8771	0.0000
ST_LEN*TIP_PREFIX_U	-0.6277	0.1831	-3.4280	0.0009

Table 12.7 Roadway HLM (with no tiers) Model Definition

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable
Intercept		0.5480
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO	0.2421
Project Length	LEN	-0.0056
Length of Structures within Project	ST_LEN	-0.3569
Roadway Percentage of Construction Cost	RW	-0.1583
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22	-0.0714

Table 12.8 Roadway HLM (with 1 tier) Model Definition

Selected MLR Numerical Variables	Variable Label	Regression Coefficients for Each Variable Tier 1 Grouping Associated with Table 5.10		
		Coast	Mountain	Piedmont
Intercept		0.7275	0.6335	0.5155
Right of Way Cost to STIP Estimated Construction Cost	ROW_RATIO	0.0944	0.0597	0.4123
STIP Estimated Construction Cost	TIP_COST		-4.46E-09	
Project Length	LEN	-0.0654	-0.0137	-0.0044
Length of Structures within Project	ST_LEN	0.9818		-0.4403
Roadway Percentage of Construction Cost	RW	-0.2884		-0.1338
Quadratic form - Project Length	len22	0.0036		
Quadratic form - Roadway Percentage of Construction Cost	rw22		-0.1691	
Quadratic form - Right of Way Cost to STIP Estimated Construction Cost	row22			-0.1697

12.3 Sample Functional Cost Estimate Worksheet

North Carolina Department of Transportation
Preliminary Estimate

TIP No.
Route
From
Typical Section

Func

County: _____

CONSTR. COST
\$0

Prepared By:
Requested By:

Date
Date

Line Item	Des	Sec No.	Description	Quantity	Unit	Price	Amount
			Clearing and Grubbing		Acre		\$ -
			Earthwork		CY		\$ -
			Drainage Existing Location		Miles		\$ -
			Fine Grading		SY		\$ -
			Pavement Widening		SY		\$ -
			New Pavement		SY		\$ -
			Pavement Resurfacing		SY		\$ -
			"Average Asphalt Wedging		SY		\$ -
			Subgrade Stabilization		SY		\$ -
			1'-6" Concrete Curb and Gutter		LF		\$ -
			2'-6" Concrete Curb and Gutter		LF		\$ -
			4" Concrete Sidewalk		SY		\$ -
			7" Monolithic Islands		SY		\$ -
			Fencing				
			Woven Wire		LF		\$ -
			Chain Link		LF		\$ -
			Erosion Control		Acres		\$ -
			Signing Interchanges				
			Diamond		Each		\$ -
			Half Clover		Each		\$ -
			SPUI		Each		\$ -
			Flyover		Each		\$ -
			Other.....		Each		\$ -
			New RR Signal with Gates		Each		\$ -
			Rubber Railroad Crossing		Each		\$ -
			Upgrade Traffic Signal		Each		\$ -
			Traffic Signal (New)		Each		\$ -
			Traffic Control		Miles		\$ -
			Thermo and Markers		Miles		\$ -
			Structures				
			ML / Creek 'Wx 'L		SF		\$ -
			RC Box Culverts				
			Ex. 3@10x10-50'Extension-3'Fill-90Skew	50	LF		
			Utility Construction				
			Relocate Existing Water Line		LF		\$ -
			Relocate Existing Sewer Line		LF		\$ -
			Misc. & Mob (15% Strs&Util)				\$ -
			Misc. & Mob (45% Functional)				\$ -

Roadway
\$ -

Strs & Util
\$ -

Lgth ___ Miles

Contract Cost \$ -
E. & C. 15%
Construction Cost \$ -

\$ -
\$ -
\$ -

12.4 Analyses of Financial Details

A database for 25 “R” projects consisting of the PE expenses was acquired from NCDOT. Activities carried out within the PE phase were classified as PE sub activities. The 25 project database showed a total of 33 sub activities. Table 12.9 lists the 33 sub activities.

Table 12.9 Listing of PE Phase Sub Activities

PE Sub Activity	Number of Projects with Sub Activity (out of 25)	Sum	Mean	Variance	Std. Dev.
SPECIAL ASSESSMENTS	21	48.34	2.30	70.18	8.37
PHOTOGRAPHY	12	1.76	0.14	0.08	0.29
COMP SUPP SERV EXPENSE	25	1.24	0.05	0.00	0.04
FLYING	9	0.82	0.10	0.00	0.07
PMII PROJECT COST	12	1.47	0.12	0.01	0.09
BSIP PROJECT COSTS	25	19.17	0.77	0.72	0.85
PUBLIC HEARINGS	14	3.82	0.27	0.08	0.28
MATL TEST & GEO INVESTN	21	108.10	5.15	16.52	4.07
MAP & PHO SHEET AND XECT	16	32.76	2.05	4.55	2.13
RIGHT-OF-WAY SURVEYS	16	44.31	2.77	24.63	4.96
PROP DATA FOR LOCN SURV	14	13.72	0.98	1.93	1.40
PRELIMINARY DESIGN	22	320.43	14.56	218.31	14.77
FIELD SURVEYS	18	94.90	5.27	20.15	4.48
OFC ACT-LOCATION SURVEYS	18	71.96	3.99	13.70	3.70
CONSTR PLAN PREPARATION	23	693.02	30.13	474.47	21.80
ROW PLAN PREPARATION	18	149.32	8.30	115.82	10.76
CONTR ENGINEERING SERV	20	561.55	28.07	634.76	25.20
NOT ASSIGNED	14	1.02	0.07	0.01	0.12
REVIEW OF CONTR ENG SER	18	211.00	11.72	313.70	17.71
PROJ MGMT SYSTEM COSTS	15	0.05	0.00	<0.0001	0.00
SAL & EXP PE PERS UTIL.	8	99.99	12.50	1166.38	34.15
CONCEPT RELOC ADV ASSIST	6	0.39	0.07	0.00	0.03
MISCELLANEOUS INVESTIGNS	11	3.65	0.33	0.41	0.64
INTERNAL ORDER SETTLEMENT	4	0.97	0.24	0.08	0.28
ADMINISTRATIVE FUNCTIONS	1	0.02	0.02	---	---
MAINT MGMT SYSTEM COSTS	6	0.00	<0.001	<0.000001	<0.001
TRAINING ACTIVITIES	1	<0.00001	<0.0001	---	---
BICYCLE AND PED FACIL	1	1.12	1.12	---	---
MISC SERV OR OPERATIONS	7	0.40	0.05	0.02	0.14
PAYMENTS TO DENR	3	12.35	4.12	50.58	7.11
TRAINING ACTIVITIES	2	0.00	0.00	<0.00001	<0.00001
UTIL MAKE-READY PLAN PREP	3	0.28	0.09	0.01	0.09

We arranged the data chronologically along with the amount spent on each sub activity during each time period. Trend graphs were plotted to check whether a consistent expense trend existed amongst the projects. A single consistent trend was not observed. Figures 12.1, 12.2 and 12.3 illustrate the expense patterns of the projects from the database. The projects show two common repetitive patterns, a) periodic spikes in expense over the entire duration of PE with occasional exceptionally high expenses and b) front loaded or initial high expenses followed by decreased expenditure as the project progresses.

The total amount spent on each sub activity was calculated as a percentage of the entire project PE expense. The percentage expenditure for each sub activity from all the projects was summed and the total amount spent on each sub activity along with its percentage contribution towards the total PE cost was calculated. The average of the amounts spent on each sub activity and their standard deviation from the mean was calculated to determine the nature of expenditure and to determine which activities, showing the highest and lowest variation, resulted in affecting the PE cost and/or duration. According to statistical observations presented in Table 12.9, the sub activities that showed the highest and most recurring costs were those of construction plan preparation, contract engineering services and review of contract engineering services. These factors are also visible from the trend graph plots. The spikes of exceptionally high expenses throughout the duration of 11 out of the 25 projects are those of contracted engineering services, construction plan preparation or review of contracted engineering services. This is indicative of the fact that the projects which have external contractors (other than department of transportation contractors), show a high and recurring preliminary engineering expenditure. It was also observed that 14 out of the 25 projects showed relatively significant expenses for public hearings. This is indicative of the fact that public opinion has an impact on projects and can be considered as a variable when estimating the PE expense as well as the PE duration. Other factors that can lead to the extension of the project duration include, but are not limited to, litigation, right of way acquisition, utility installation and redesign/rework.

The financial expenditures of additional projects were graphed in Figures 12.4 through 12.21.

Breakdown of expenditure over entire PE phase for R-2405

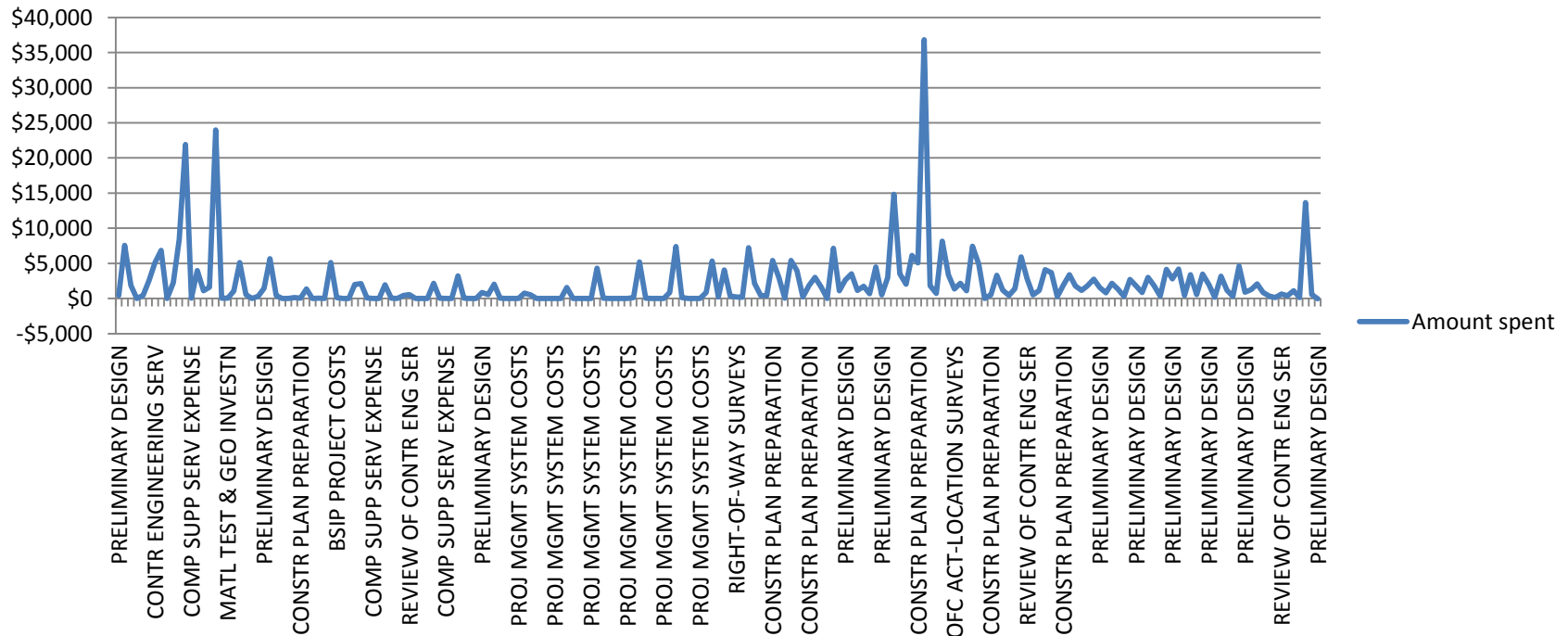


Figure 12.1 R-2405 Breakdown of PE Expenditures over the PE Phase (Trend 1)

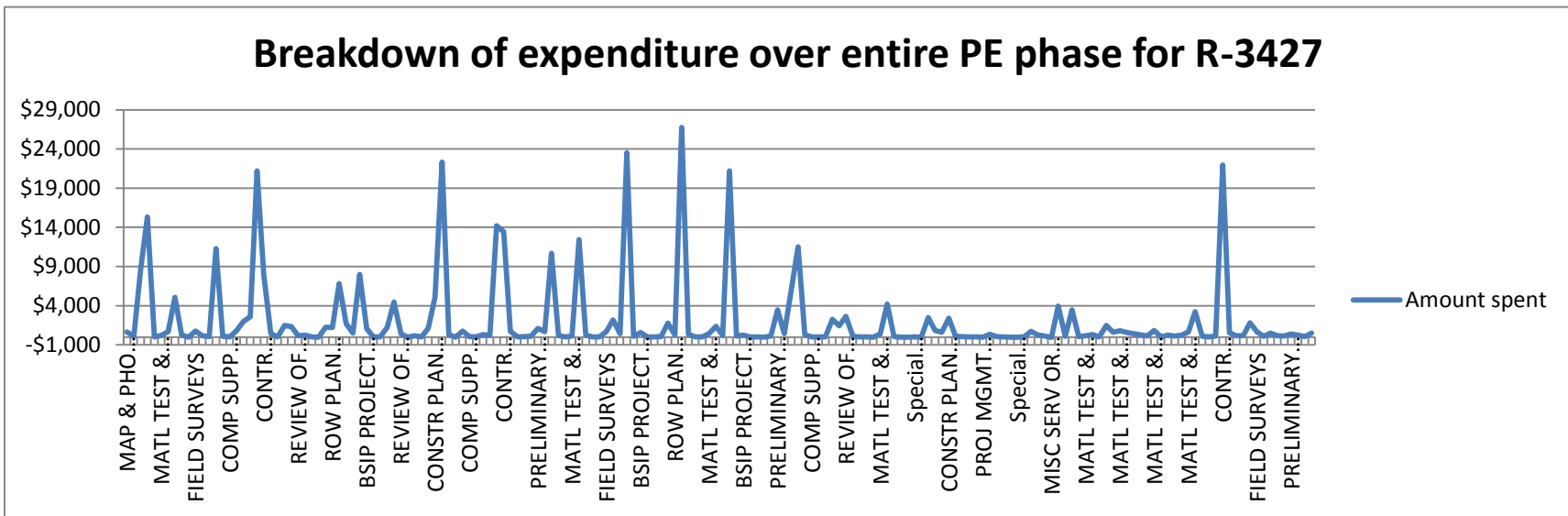


Figure 12.2 R-3427 Breakdown of PE Expenditures over the PE Phase (Trend 1)

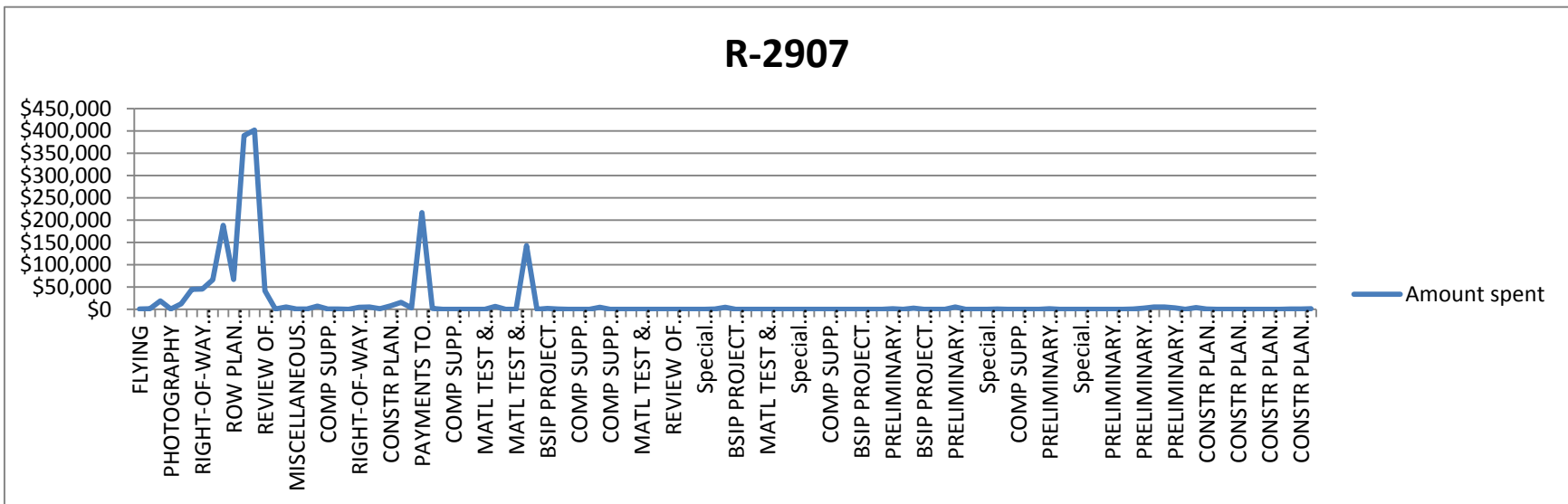


Figure 12.3 R-2907 Breakdown of PE Expenditures over the PE Phase (Trend 2)

Break down of expenditure over entire PE phase for R-2207

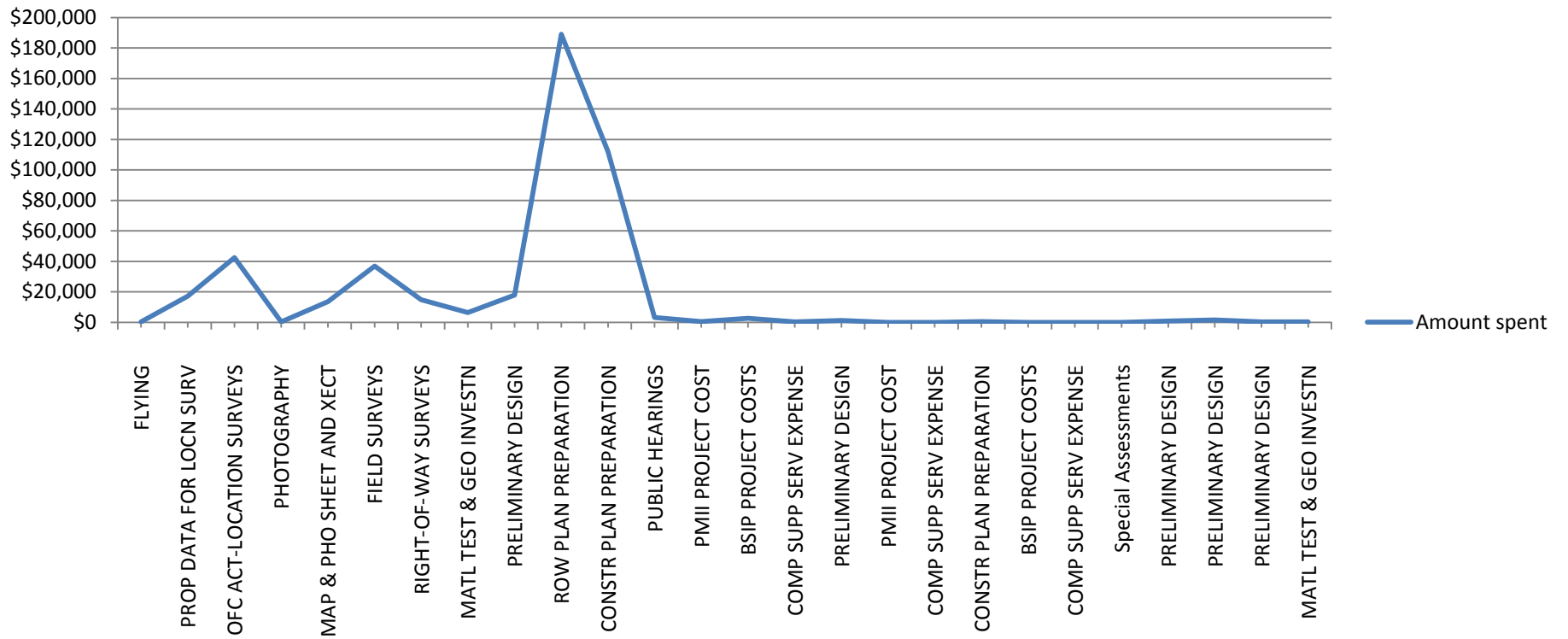


Figure 12.4 R-2207 Breakdown of PE Expenditures

Breakdown of PE expenditure over entire PE phase for R-2248

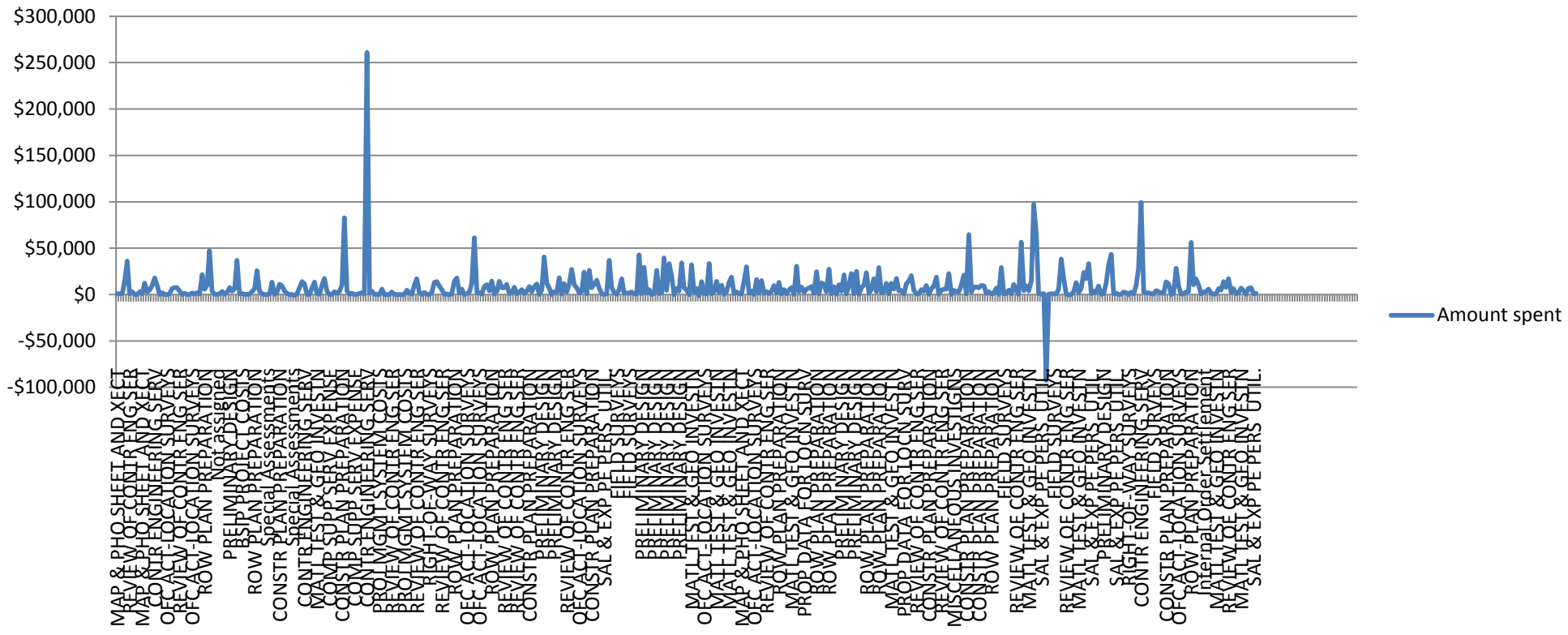


Figure 12.5 R-2248 Breakdown of PE Expenditures

Breakdown of expenditure over entire PE phase for R-3807

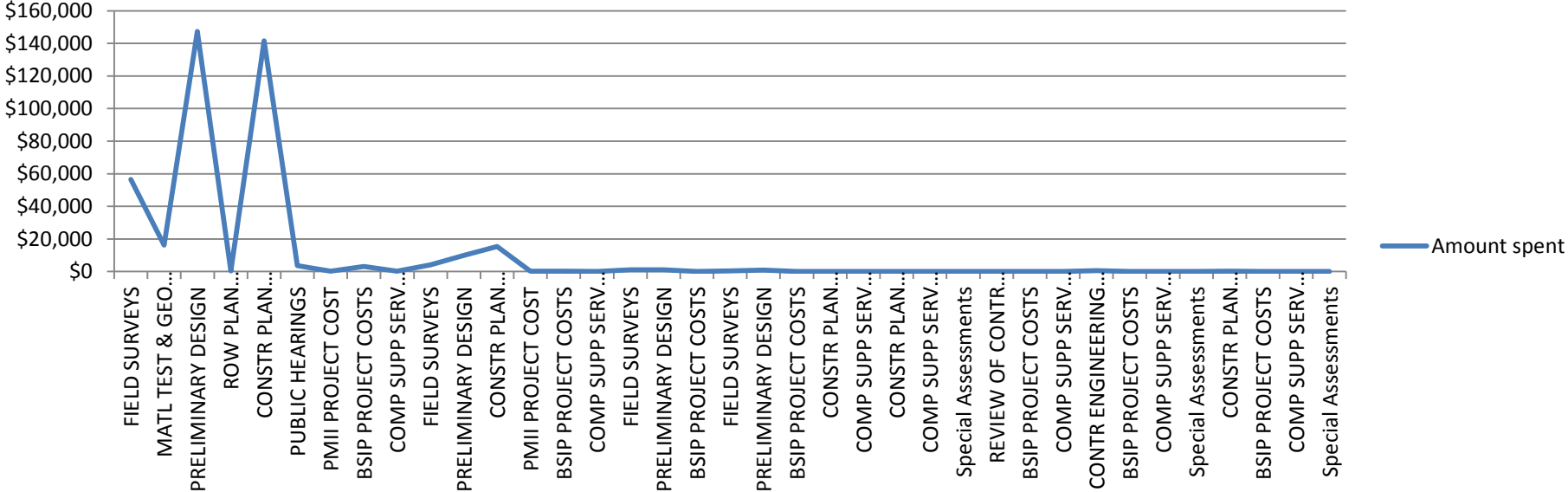


Figure 12.6 R-3807 Breakdown of PE Expenditures

Breakdown of PE expenditure over entire PE phase for R-3303

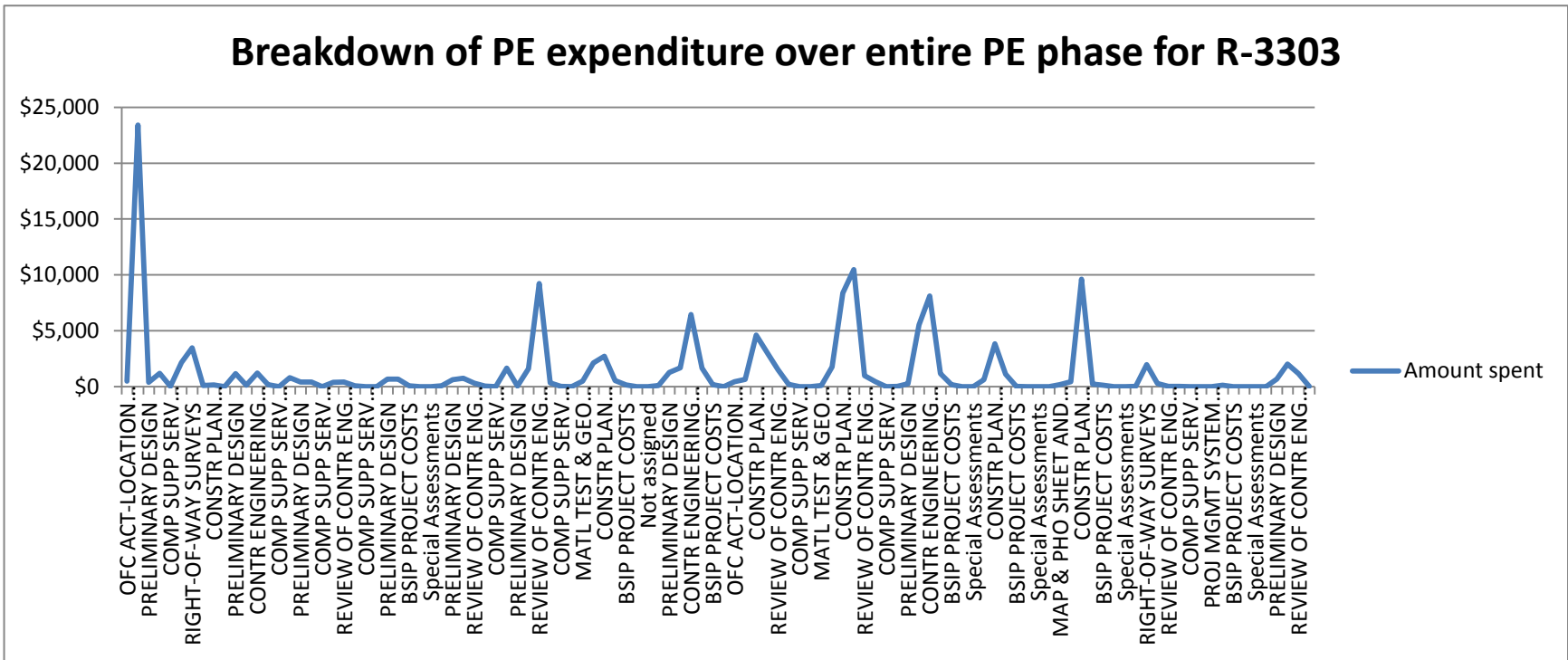


Figure 12.7 R-3303 Breakdown of PE Expenditures

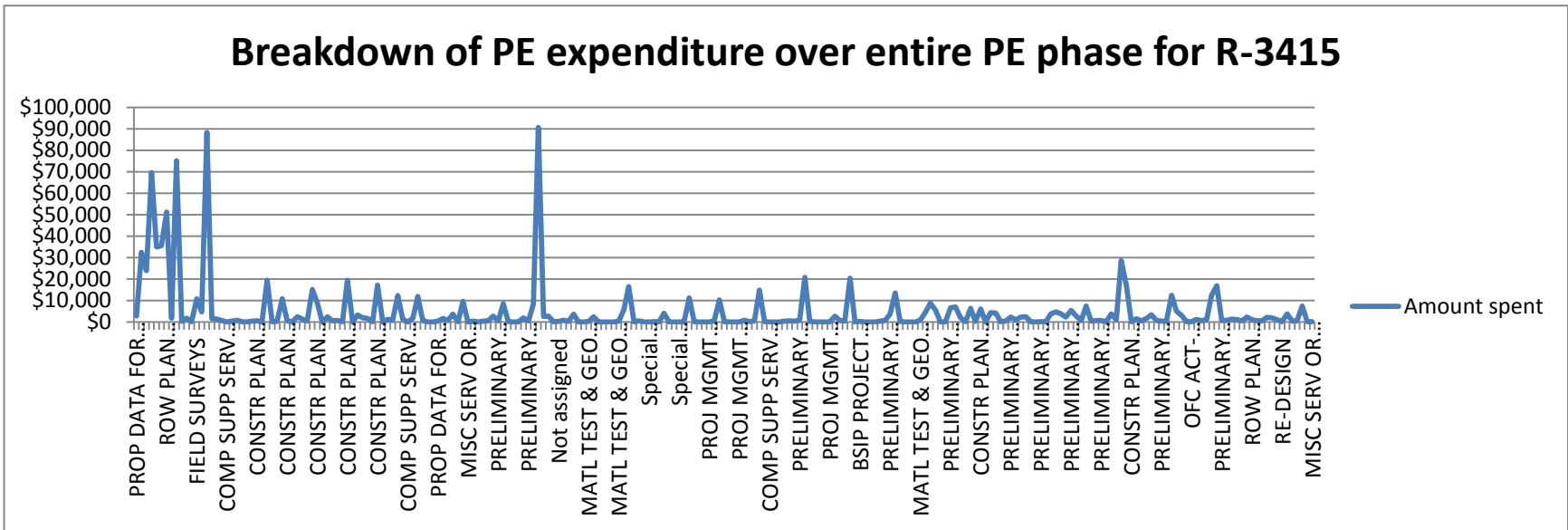


Figure 12.8 R-3415 Breakdown of PE Expenditures

Breakdown of PE expenditure over entire PE phase for R-2904

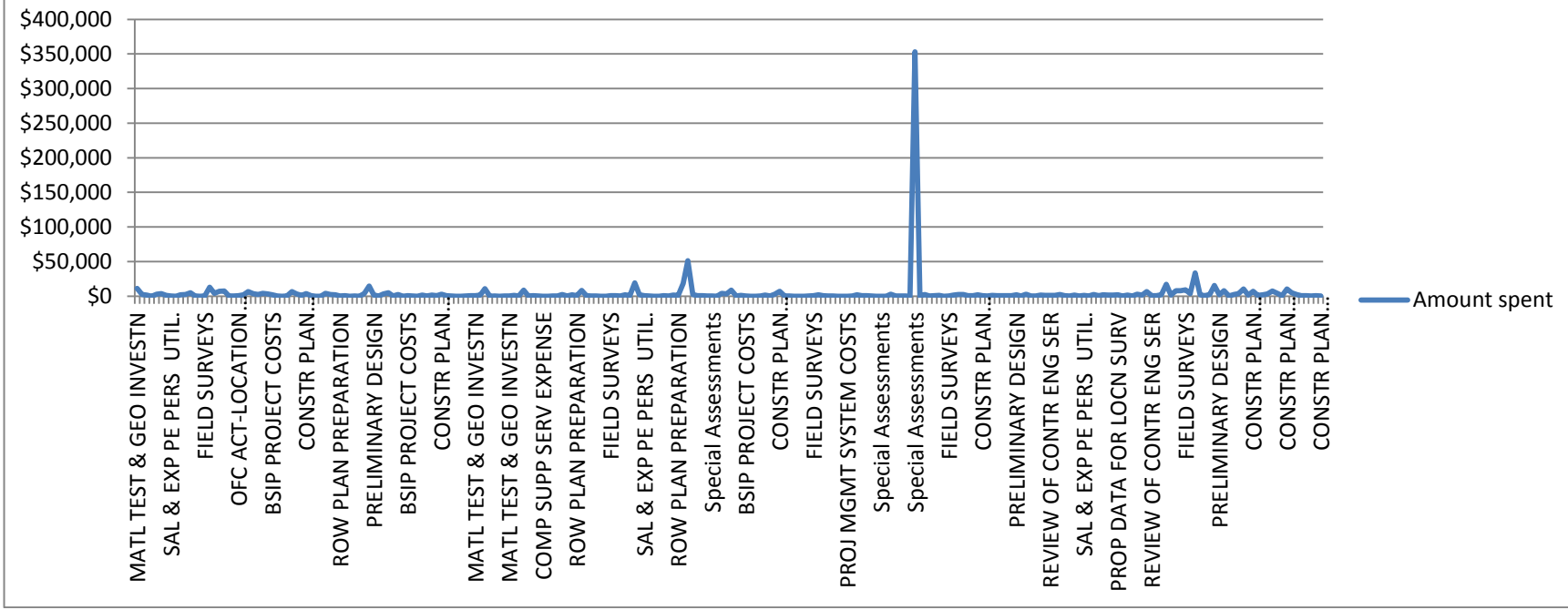


Figure 12.9 R-2904 Breakdown of PE Expenditures

Breakdown of PE expenditure over entire PE phase for R-2823

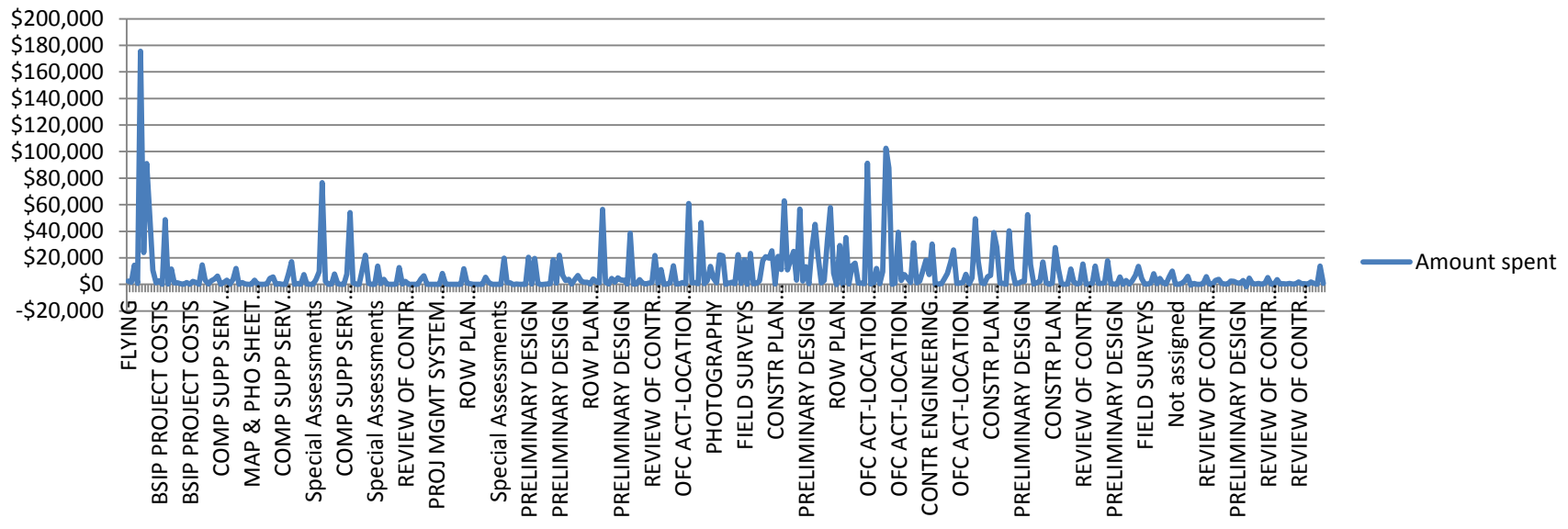


Figure 12.10 R-2823 Breakdown of PE Expenditures

Breakdown of PE expenditure over entire PE phase for R-2643

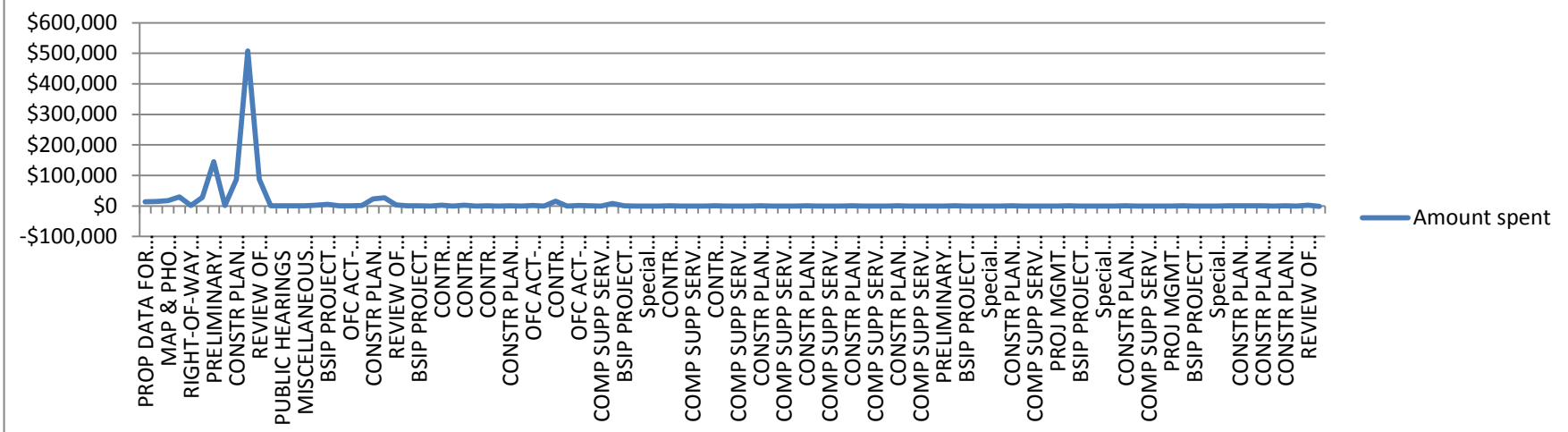


Figure 12.11 R-2643 Breakdown of PE Expenditures

Breakdown of PE expenditure over entire PE phase for R-2616

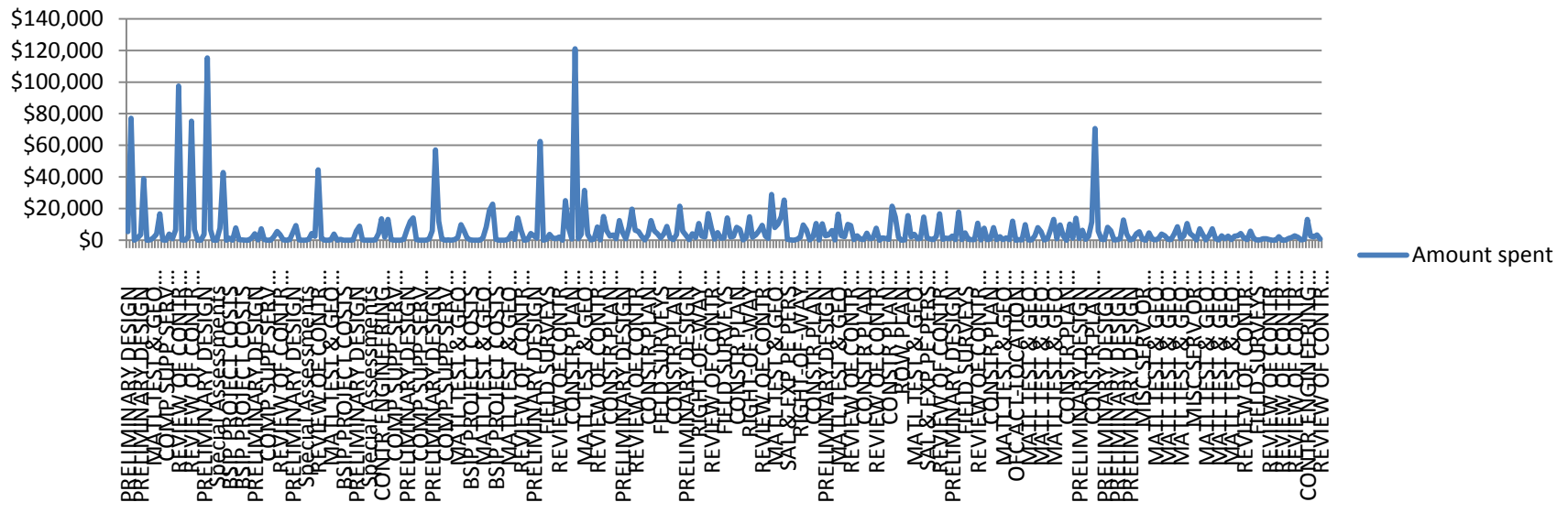


Figure 12.12 R-2616 Breakdown of PE Expenditures

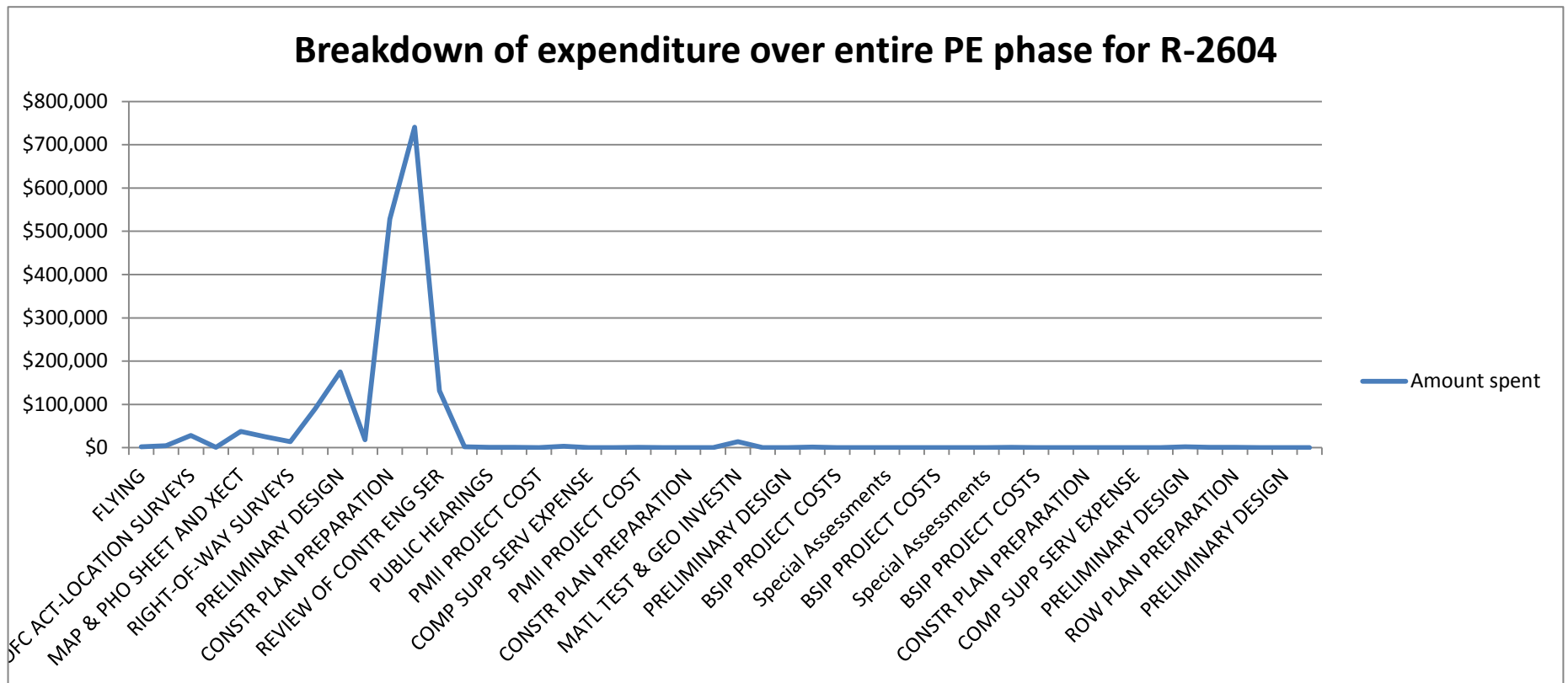


Figure 12.13 R-2604 Breakdown of PE Expenditures

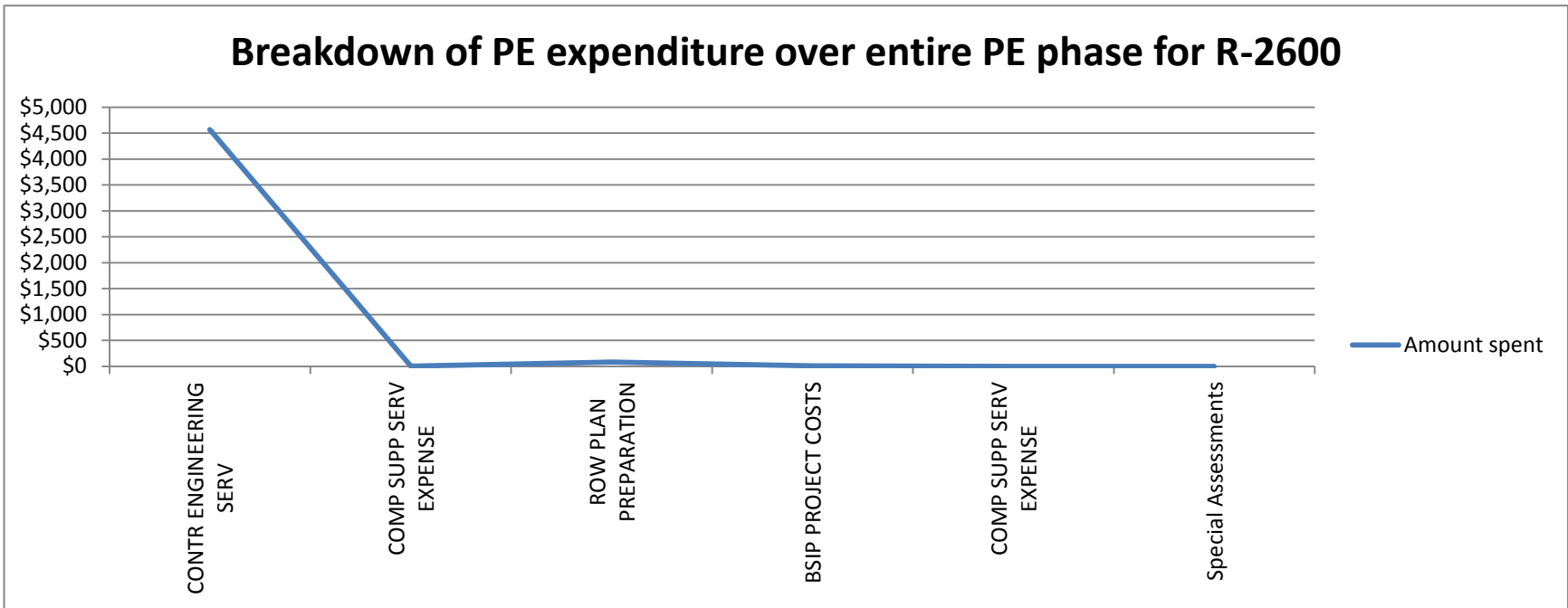


Figure 12.14 R-2600 Breakdown of PE Expenditures

Breakdown of PE expenditure over entire PE phase for R-2555

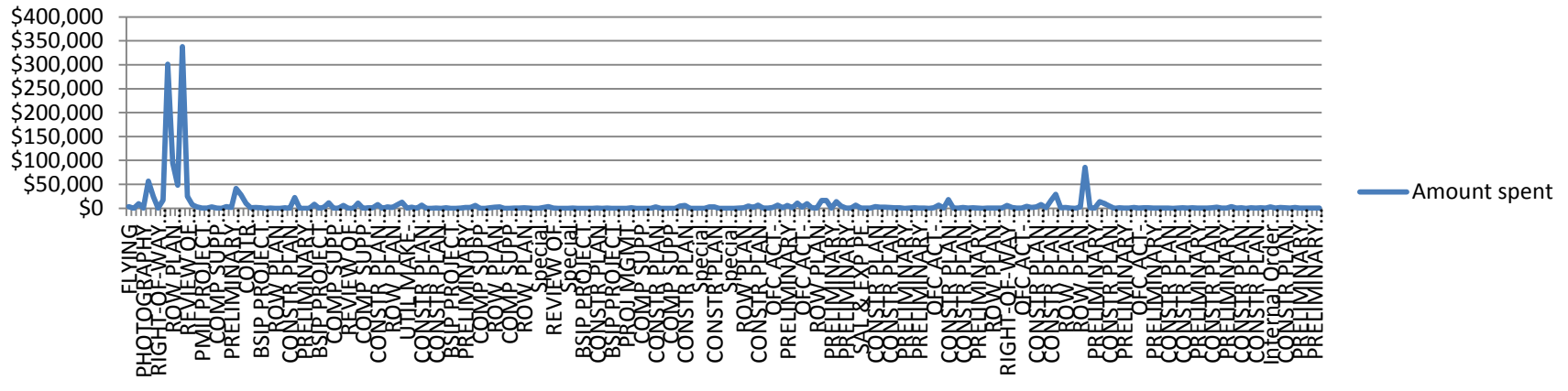


Figure 12.15 R-2555 Breakdown of PE Expenditures

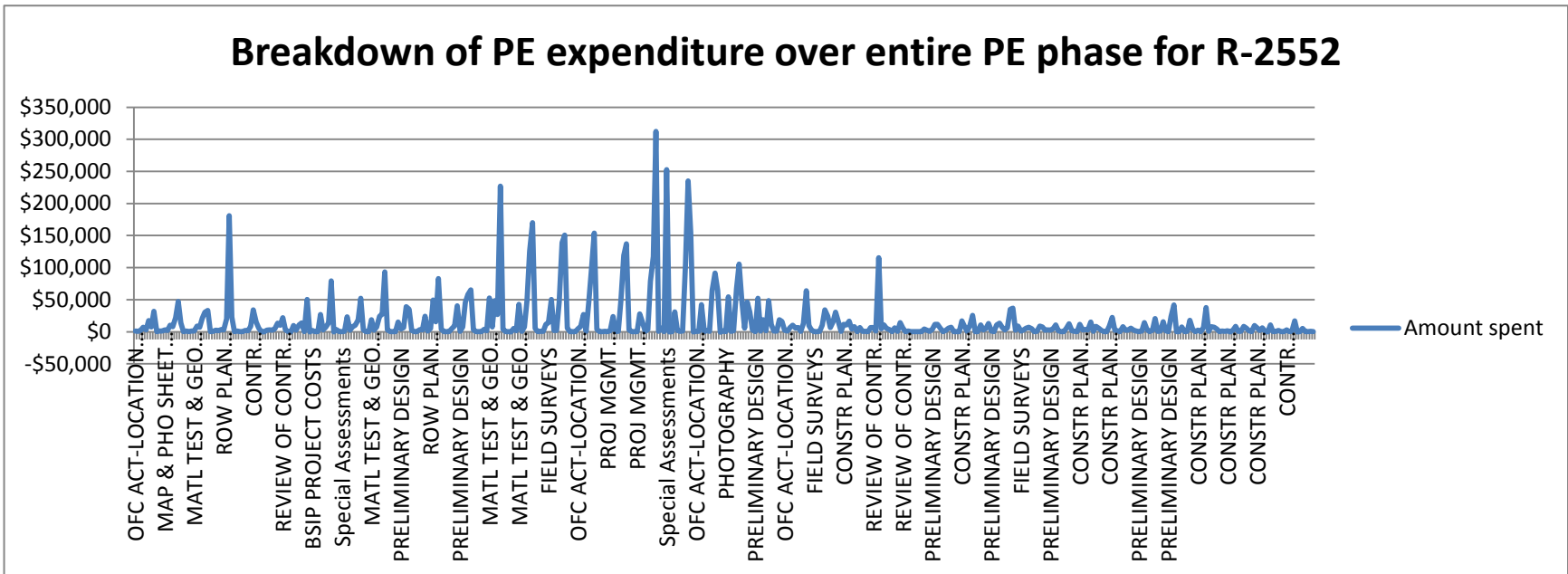


Figure 12.16 R-2552 Breakdown of PE Expenditures

Breakdown of expenditure over entire PE phase for R-2538

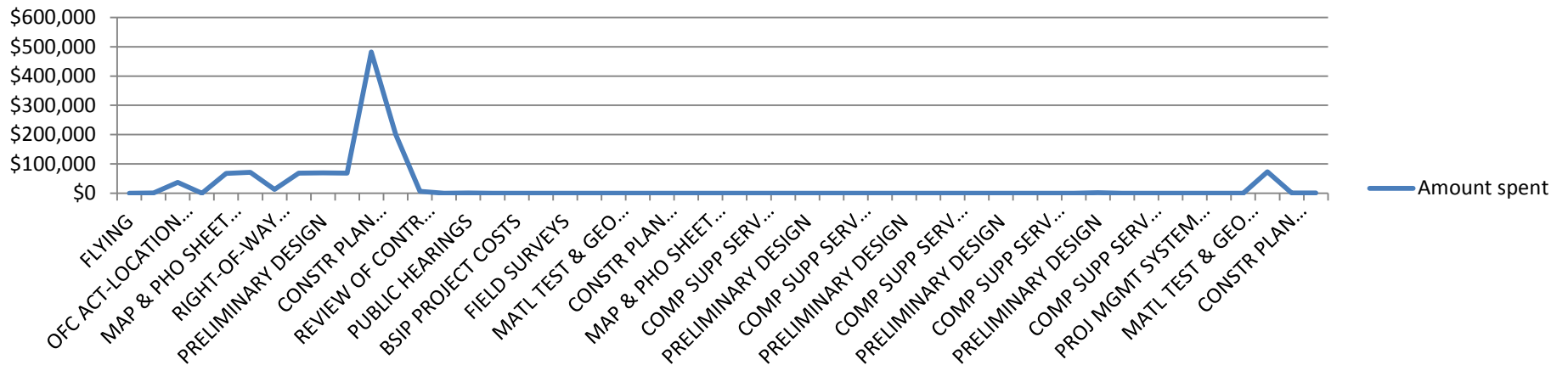


Figure 12.17 R-2538 Breakdown of PE Expenditures

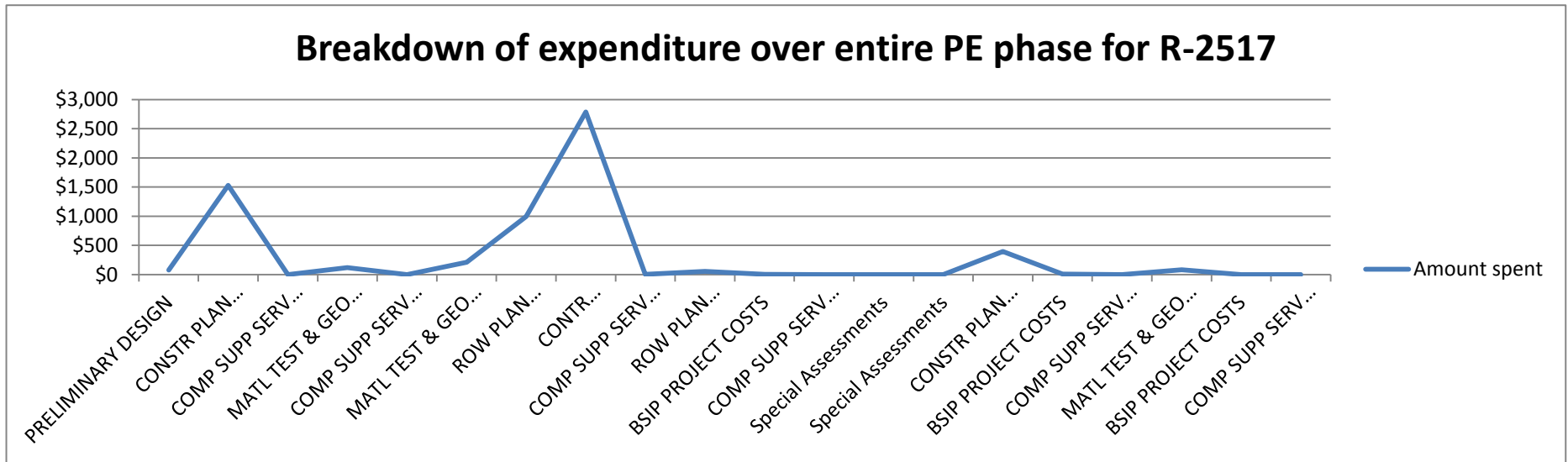


Figure 12.18 R-2517 Breakdown of PE Expenditures

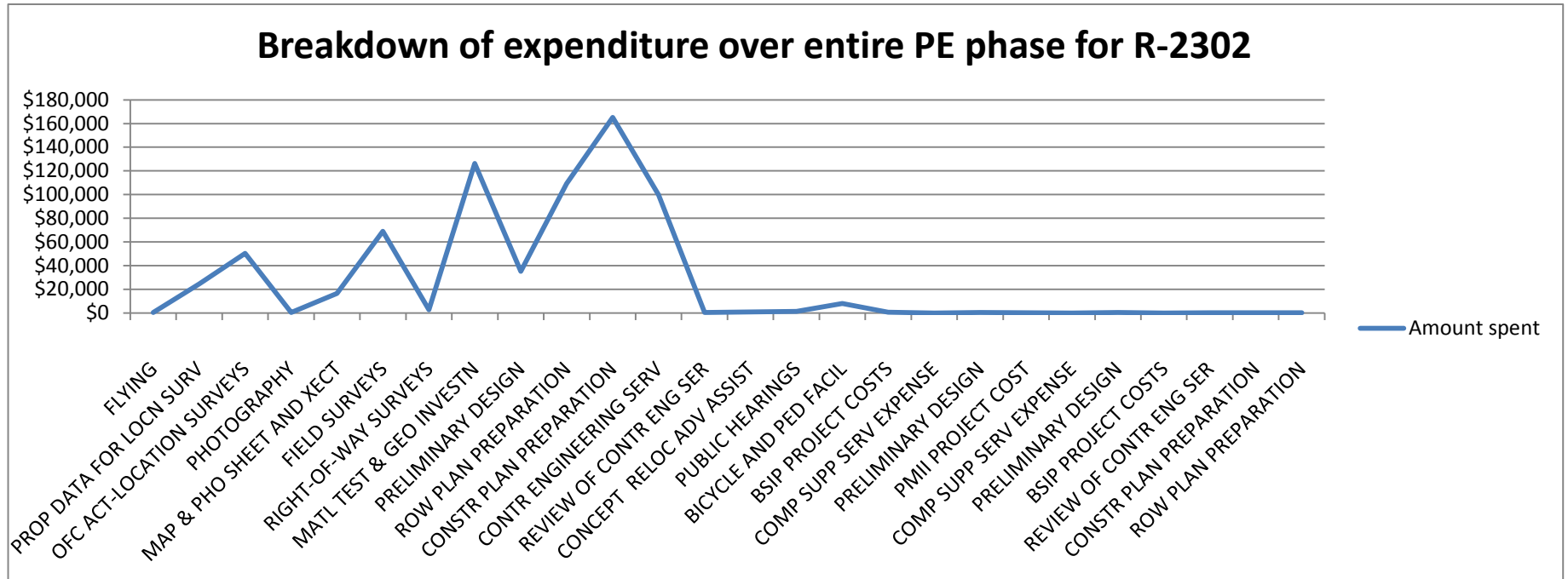


Figure 12.19 R-2302 Breakdown of PE Expenditures

Breakdown of PE expenditure over entire PE phase for R-2236

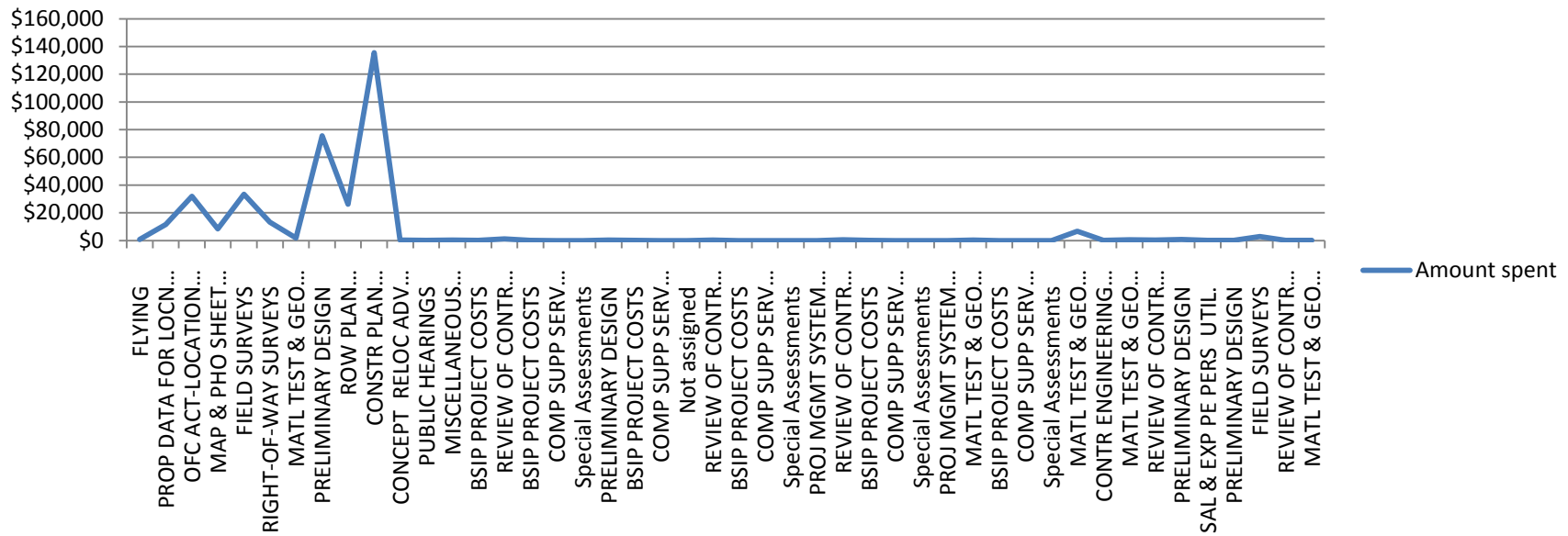


Figure 12.20 R-2236 Breakdown of PE Expenditures

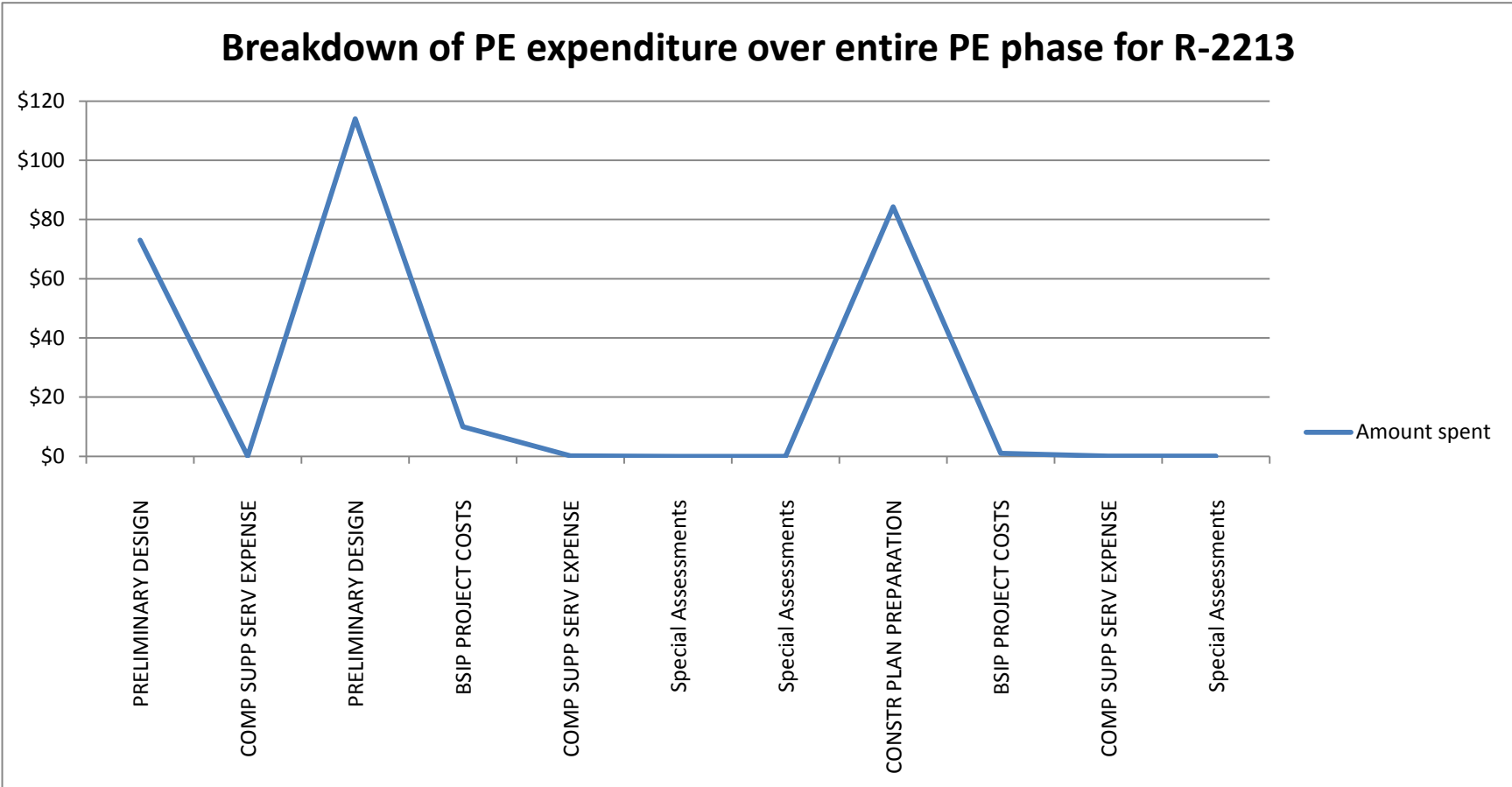


Figure 12.21 R-2213 Breakdown of PE Expenditures

(End of Appendices)