2015

Applying artificial neural networks to top-down construction cost estimating of highway projects at the conceptual stage

Brendon Joseph Gardner
Iowa State University

Applying artificial neural networks to top-down construction cost estimating of highway projects at the conceptual stage

by

Brendon Joseph Gardner

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Civil Engineering (Construction Engineering and Management)

Program of Study Committee:
Douglas D. Gransberg, Major Professor
Hyung Seok "David" Jeong
Peter Savolainen

Iowa State University

Ames, Iowa

2015

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

LIST OF TABLES

Page

ACKNOWLEDGEMENTS

Firstly I would like to thank both Dr. Doug Gransberg and Dr. David Jeong for the inspiration and support to complete this research. The many meetings and chapter reviews were greatly appreciated.

Secondly, many thanks to staff at MDT whom have helped with data collection and sharing of cost estimating practices through meetings and correspondence. I would especially like to acknowledge Lesly Tribelhorn and Kris Christensen.

I would also like to thank my friends for the fantastic experience at Iowa State University. Particularly thanks to Emily and Jorge whom have read, critiqued and supported my work over the past 1.5 years.

ABSTRACT

Conceptual cost estimating (CCE) is a challenging task for highway agencies due to the limited design information available at early stages of project development. As a result, agencies frequently experience large variance from the initial construction estimate to the final cost. Despite the initial estimate’s low level of confidence, it is required for all highway projects as an input to feasibility studies and to establish the project’s budget.

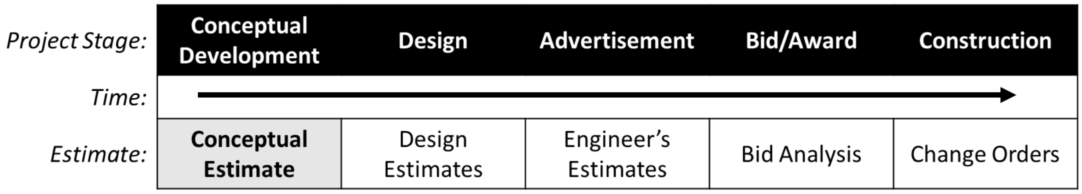
Many authors have explored the use of artificial intelligence and multiple-regression analysis with promising findings to aide CCE. Unfortunately, at this writing, no highway agencies are known to have implemented these data-driven techniques in practice. One of many reasons for this situation is related to a belief that accurate quantities of work are required to produce an accurate estimate. This approach is termed ‘bottom-up’ estimating and is clearly impossible at the initial stage of project development. A second reason relates to the investment necessary to create a reliable database structure that permits high-level statistical analysis. Therefore, this thesis seeks to investigate improvements to data-driven, ‘top-down’ CCE methods to enable practical application.

Firstly, a method to rationally select data used in the model is investigated. The analysis reported in this thesis found that random sampling does not test the true performance of a model for its future application. Secondly, a method to select input variables that have the largest impact on predicting the construction cost but require the least amount of effort is proposed. The models reached a point whereby expending additional effort to include more input variables did not yield an increased performance and debunked the notion that ‘bottom-up’ estimating approaches are intuitively more accurate. This finding is significant for practitioners as resources expended to collect and store additional data points than required is wasted at the conceptual stage.

Finally, a method to express the conceptual estimate stochastically is proposed. The traditional deterministic approach of relying on a specific number communicates false precision. This thesis proposes combining artificial neural networks with bootstrap sampling to create an empirical distribution of the construction costs and better communicate a likely range of project costs.

CHAPTER 1. INTRODUCTION

The first estimate of a highway project's construction cost is defined as the conceptual estimate in the project development timeline shown in Figure 1. At the conceptual stage there is little information known about a project and the detailed design has not yet begun. Further cost influencing information established during project design stages is included when developing the Design Estimates and Engineer's Estimates. Highway agencies are therefore more confident with these later estimates.



Project Stage:	**Conceptual Development**	**Design**	**Advertisement**	**Bid/Award**	**Construction**
Time:	→				
Estimate:	**Conceptual Estimate**	Design Estimates	Engineer's Estimates	Bid Analysis	Change Orders

Figure 1. Construction cost estimating timeline (adapted from Schexnayder et al. 2003)

The development of an effective conceptual estimate can be a challenging task for public owners as these estimates are conducted prior to the design phase with minimal scope definition. Despite the lack of knowledge about a project at the conceptual cost estimating (CCE) stage, these estimates are required by public agencies to estimate the cost of projects for statewide fiscal funding requirements (Anderson et al. 2007, FHWA 2015). This federal requirement is for state departments of transportation (DOT) to develop a State Transportation Improvement Program (STIP) detailing four years of upcoming projects (FHWA 2015).

Flyvbjerg et al. (2002) investigated public transportation projects and found that 86% of projects had experienced cost growths since the initial estimate, on average they were 28% higher than the initial estimate. That study included 258 transportation infrastructure projects from different historical periods, geographical regions and project types, with a combined value of $90 billion. Flyvbjerg et al. also discovered that there have been no improvements in the accuracy of the initial cost estimate from the 70 years of data that was analyzed. In 2003, Schexnayder et al. found that publicity called into question the "ability of departments of transportation to forecast accurately and to control the final cost of their projects."

Cost estimates are typically classified in the following five groups (AASHTO 2013; Turochy et al. 2001):

- Preliminary Engineering (PE),
- Right-of-way (ROW),
- Final Design – Plans, specifications and estimate (PS&E)
- Construction costs (CN), and
- Construction Engineering (CE)

PE concerns the costs associated with project advancement during the planning stage. ROW is all costs associated with land purchase. PS&E is costs associated with producing the final design, specifications and estimation of the construction costs prior to bid. CN is the expenses associated with the construction process. CE covers the monitoring costs incurred with management during the construction phase by the highway agency. This thesis concentrates solely on estimation of the construction costs, or CN amount, specifically at the conceptual stage of project development.

Motivation

One of the key problems at the CCE stage is the 'limited information' known about a particular project's scope during the planning stage (Schexnayder et al. 2003; AASHTO 2013). Importantly, it is at the CCE stage where designers have the greatest ability to influence the end project cost. This introduces the 'cost estimating dilemma' as shown in Figure 2 (Bode 2000). Confidence in CCE enables designers to alter designs and realize savings when they have the greatest ability to influence the cost of the project. This 'dilemma' highlights the importance of an accurate conceptual estimate as the cost of construction can be "impacted significantly by decisions made at the design stage" (Gunaydin and Dogan 2004).

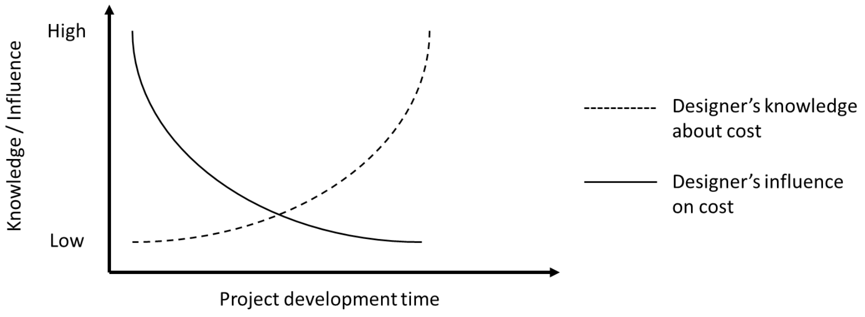


Figure 2. Cost estimating dilemma (adapted from Bode 2000)

Estimating construction costs at the conceptual stages of project development is critical for decision-makers to determine a reasonable project budget and make decisions regarding the project's ultimate feasibility (Harbuck 2007; Lowe et al. 2006; AASHTO 2013). Early construction cost estimates are "the basis for key financial decisions. Thus, the inability to accurately estimate the project costs can result in poor financial decisions" (AASHTO 2013). If the conceptual cost estimate is too high, then a project may be erroneously rejected based on an unfavorable benefit-to-cost ratio. On the other hand, if the cost estimate is too low, then a project may be found to be feasible when in fact it is not and should have been rejected (AASHTO 2013).

Highway agencies need reasonable accuracy in estimating conceptual construction costs to ensure that tentative construction programs optimize available fiscal year funding. Under-estimation during the CCE stage can result in agencies running short of funds to complete its annual construction program. Over-estimating costs can result in too few projects being selected for funding in a given fiscal year, this leads to not having enough projects ready and advertising them before they are truly ready to let, or worse, the loss of federal funding (MDT 2007).

The design budget, a major portion of the preconstruction budget, is typically established as a percentage of estimated construction costs (Jeong and Woldensenbet 2012). It therefore follows that if the construction cost estimate is low, the design budget will also be less than the amount required. The amount of a project's budget allocated to design was found to directly influence its overall construction cost growth from the early estimate (Gransberg et al. 2007). Gransberg et al.'s work observed that up to a point, the greater the investment in design the lower the construction cost growth from its initial estimate. Thus, underfunding the design budget yields the potential for construction budget overruns. As a result, the need to carefully calculate construction costs at an early stage to ensure an appropriate budget for the design and control cost growth to the project becomes even more important.

The four key motivators for studying conceptual estimating, discussed above, are real issues faced by highway agencies. This research is part of a bigger research project to develop an artificial neural network to provide a data-driven conceptual cost estimating tool for Montana Department of Transportation (MDT). Much of the research in this thesis is leveraging the artificial neural network created for MDT.

Content Organization

Chapter 2 contains background for the reader. Specifically the state-of-the-practice of CCE is discussed, background to artificial intelligence in the area of construction cost estimating and application of 'top-down' construction cost estimating. At the end of Chapter 2 the specific problem statements in the report are stated.

Chapter 3 highlights the overall approach and validation methodology. In this chapter the creation of a database and artificial neural network model are introduced. This includes data collection methods, selecting input variables and validation techniques. Later chapters utilize the model created in Chapter 3 to answer the problem statements set in Chapter 2. Chapter 3 finishes with a diagram of the global methodology.

This thesis contains three journal articles shown in Chapters 4, 5 and 6. The chapters are stand-alone documents, each with a specific focus on conceptual estimating of highway projects using data-driven techniques. Whilst the focus of each article differs, they all contribute to the overall research objective. The chapters commence with data sampling techniques for artificial neural networks (Chapter 4); then the focus shifts to quantifying the level of effort expended to conduct the conceptual cost estimate (Chapter 5); finally the report investigates the ability to communicate the conceptual cost estimate stochastically through a range interval (Chapter 6).

Chapter 4 will be submitted for publication in the American Society of Civil Engineers *Journal of Computing in Civil Engineering*. This chapter proposes a method that could be used to sample projects to be included in the artificial neural network database of historic projects. This chapter highlights to highway agencies that performance of data-driven CCE models need testing against a sample of data that is reflective of the future distribution of project types to be predicted.

Chapter 5 will be submitted for publication in the American Society of Civil Engineers *Journal of Construction Management in Engineering*. This chapter specifically focuses on measuring the level of effort expended to conduct the conceptual estimate. More specifically a method is proposed for estimators to focus on collecting input variables which require a low level of effort but have a high impact on the construction cost estimate.

Chapter 6 will also be submitted for publication in the American Society of Civil Engineers *Journal of Risk Management and Uncertainty.* This chapter investigates the bias associated with point estimates developed at the conceptual stages. Further it investigates a method to produce an empirical distribution of expected construction costs through the use of combining artificial neural networks and bootstrap sampling. Research is validated by comparing the actual final construction cost to that expressed by the range estimate.

Chapter 7 summarizes the main conclusions from the papers and addresses the problem statements and objective of this report. Additionally, Chapter 7 discusses the limitations of those conclusions. Finally, Chapter 8 outlines the key contributions to the conceptual estimating body of knowledge and areas for future research.

CHAPTER 2. BACKGROUND

This chapter benchmarks the current state-of-the-practice used to estimate the conceptual cost of projects at highway agencies. It then discusses the application of artificial intelligence to the field of conceptual estimating at highway agencies. Finally the chapter poses the specific research problems investigated in this thesis.

Current State-of-the-Practice

The American Association of State Highway and Transportation Officials (AASHTO) recently released a *Practical Guide to Cost Estimating* (2013) to provide a nationally recognized set of procedures to conduct the cost estimating at highway agencies for all project development stages. That guidebook describes conceptual estimates as an early projection of cost when limited information is known about a project. The suggested method for estimating at the conceptual stage is to develop statistical relationships between cost factors for completed projects and use these to predict future construction costs. This is suggested through the use of parametric cost estimating relationships, such as cost-per-mile of a highway or the cost-per-area for bridge, and adjusted through historical percentage cost factors. The guidebook suggests storing historical data in a spreadsheet or computer software such as AASHTOWare ® Project BAMS/DSS. To reflect uncertainty at the conceptual stage the guidebook developed classifications with accepted uncertainty summarized in Table 1. The accepted estimate ranges at the planning development stage, shown in Table 1, are referenced in later chapters to make comparisons with the performance of the data-driven techniques investigated.

Table 1. Cost Estimating Classification (adapted from AASHTO 2013)

Project Development Phase	Project Maturity (% of the project definition completed)	Purpose of the Estimate	Estimating Methodology	Estimate Range
Planning	0 to 2%	Conceptual Estimating – Estimate Potential Funds Needed (20-year plan)	Parametric	-50% to +200%
	1% to 15%	Conceptual Estimating – Prioritize Needs for Long-Range Plans (10-year plan)	Parametric or Historical Bid-Based	-40% to +100%
Scoping	10% to 30%	Design Estimating – Establish a Baseline Cost for Project and Program Projects	Historical Bid-Based or Cost-Based	-30% to +50%
Design	30% to 80%	Design Estimating – Manage Project Budgets against Baseline	Historical Bid-Based or Cost-Based	-10% to +25%
Final Design	90% to 100%	Compare with Bid and Obligate Funds for Construction	Cost-Based or Historical Bid-Based Using Cost Estimate System	-5% to +10%

The National Cooperative Highway Research Program (NCHRP) produced *Report 574* on *Guidance for Cost Estimation Management for Highway Projects during Planning, Programming and Preconstruction* (Anderson et al. 2007). That NCHRP guidebook was created to identify cost estimating management practices for each phase of project development in order to reduce cost escalation on highway projects. Various management strategies are presented to reduce the risk of cost growth. Under the planning development section methods identified included estimate management, risk management, document quality and estimate quality. The estimate quality section in the report identifies six tools including the use of computer software, conceptual estimation, estimate reviews (internal and external), project scoping and right-of-way.

Surveys investigating current practices have been completed by Byrnes (2002) and Turochy et al. (2001). Both were conducted for estimating construction costs of highway projects at the conceptual stage. Although the actual technique and terminology varies by state, both of these studies found that CCE approaches utilized by highway agencies are generally classed into one the following three categories:

- “cost-per-mile” of typical sections of highway or bridge,
- estimating approximate quantities of major work items, or

- no documented or uniform method, instead using experience and engineering judgement.

The report by Turochy et al. (2001) was completed for Virginia Department of Transportation (VDOT) in response to “attention from news media and elected officials” due to major increases in highway project cost estimates since the planning stage. This further highlights the motivation for creating an accurate conceptual cost estimate and the public scrutiny faced by highway agencies for failing to do so.

It was discovered by Byrnes (2002) when he surveyed all 50 state DOTs that no agencies were at that stage employing sophisticated mathematical models. The same finding was reached by Turochy et al. (2001) with the suggestion that there is a reliance on the experienced personnel at highway agencies to conduct the conceptual estimate. Turochy et al. (2001) specifically identified the potential to develop models for estimating highway project costs through the use of completed cost data with a large number of projects. The details of possible models, proven in the literature, are described in the artificial intelligence section that follows.

Artificial Intelligence

The advancement in digital technology and data storage capacity has meant that state DOTs have an abundance of data available from past projects to estimate the cost of future projects with. The literature shows that two data-driven cost estimating methods, artificial neural networks (ANNs) and multiple-regression analysis (MRA), have been proven to provide reasonable estimates of the conceptual costs of highway projects (Bell and Ghazanfer 1987; Hegazy and Ayed 1998; Mahamid 2011). Both techniques link an historical database of project attributes to the actual construction cost. These relationships identified within the data can then be used to forecast the construction cost of future projects.

MRA is the development of a regression equation to link independent project variables to the cost (Turochy et al. 2001). The equation assigns weights to each of the independent attributes through the method of least error (Turochy et al. 2001). Future construction costs can be estimated using the same equation weights but with the new independent variables. The downside of MRA is that an assumption must be made about the relationship between the terms (Sonmez 2008). Many authors of the MRA literature simplified this required assumption to be a linear regression equation (Bell and Ghazanfer 1987; Sonmez 2011; Mahamid 2011). Alternative relationships could be investigated such a quadratic correlation between terms. Multivariate

Adaptive Regression Splines (MARS) is another such relationship with emerging publications, in this relationship instead of one linear line there are multiple ‘piecewise’ linear sections (Haleem et al. 2013).

ANNs do not require a discrete assumption that a link exists between the construction cost and the variables (Kim et al. 2004). The model uses artificial intelligence to find patterns within the database to link these to the dependent variable (construction cost). The ANN model creates layers of arbitrary data to transform the input variables to the construction cost. Historical data is used to train the ANN model and recognize relationships within the database to the dependent variable. This trained model is then used to forecast future construction costs by looking for similar patterns and predict the dependent variable.

Bell and Ghazanfer (1987) published one of the first MRA models for predicting the cost of highway construction maintenance projects with a database of 174 projects. When validated against test projects it could predict the construction cost to within 17% on average. This error is well within the range recommended in the AASHTO *Practical Guide to Cost Estimating*, for which the conceptual estimate should be in the range of -40% to +100% of the final construction cost (AASHTO 2013) shown in Table 1.

Since Bell and Ghazanfer published their model more than 15 authors have published data-driven CCE models with similar promising results using MRA and ANNs at the CCE stage. In 1992, Sanders et al. published an MRA model with only a 6% error on test projects. Creese and Li, in 1995, published an ANN model with 8.24% estimating error for the construction costs of timber bridges. In 1998 Hegazy and Ayed published an ANN model that could estimate the construction cost of highway projects in Newfoundland, Canada, to within 19.33% of the actual cost. Kim et al. (2004) completed a comprehensive study comparing the performance of ANN, MRA and case-based reasoning to calculate the construction cost of residential buildings in Seoul, South Korea. A total of 530 projects were used in the database, far exceeding the number of projects used by other authors. The estimating accuracy of the model was 3.0% and 7.0% for ANN and MRA models respectively. Details of all 16 studies are analyzed in depth in the literature review sections of Chapters 4 and 5.

Despite promising results from the literature, no DOT is using a data-driven CCE model to assist them in calculating the construction costs of their projects. It is however known that CCE conducted by DOTs lack results with high confidence (Chou et al. 2006; Byrnes 2002;

Walton and Stevens 1997). Turochy et al. (2001) concluded that DOTs are not employing computer model techniques to improve estimate confidence due to:

1. Resistance to replace engineering judgment with computer procedures, and
2. Long term reliance on the skills and experience of planners and engineers.

One benefit of data-driven estimation, such as the ANNs or MRA, is the ability to remove bias and possible pressure to keep estimates under published budget ceilings, a challenge estimators regularly face (Anderson et al. 2007). Flyvbjerg et al. (2002) discovered that underestimation is the rule rather than exception for transport infrastructure projects merely to keep the project from being cancelled before construction begins. Computer tools using historic project information to predict future construction costs can remove the optimism at the CCE stage by relying on real construction data, and taking the emotion and possible bias out of the process.

Multiple researchers have proven the ability of ANNs to produce superior results to MRA in the field of construction cost estimating (Petroutsatou et al. 2012; Kim et al. 2004; Moselhi and Siqueira 1998), some researchers have proven the contrary (Gunduz et al. 2011; Setyawati et al. 2002). Many of the research problems investigated in this thesis use ANNs. This was due to a superior performance of this tool over MRA for the data collected in this research project. Despite this, the use of MRA will be briefly investigated in Chapter 5. This research does not attempt to investigate whether ANN models or MRA models are the most accurate.

Top-down Cost Estimating

One approach to estimate construction cost is for the estimator to break up the project into individual activities and then estimate the cost of each activity based on the resources of materials, labor and plant required (Kim et al. 2012). This approach is termed ‘bottom-up’ estimating. Each resource is then assigned unit rates and the summation of each activity cost is the estimated total project cost. These rates often come from historical bid price averages that estimators have recorded at the highway agency (Byrnes 2002). Byrnes’ research concluded that likelihood of estimate accuracy was directly proportional to the amount of bid tabulation data the estimators included in the database.

An issue with this approach at the conceptual development stage is that quantities are uncertain because the design is far from complete (Kim et al. 2012). A superior approach at the conceptual stage is to focus on the “larger picture” (AASHTO 2013). This is termed a ‘top-down' cost estimate and is commonly used at the conceptual stage when project definition is still fluid (Kim et al. 2012). The 'top-down' estimate focuses on project characteristics such as location, traffic management considerations, utility impacts and other complexities that drive the construction cost (AASHTO 2013).

Top-down cost estimating principles are applied in this research to develop the database used for the construction of an artificial neural network and its resultant cost estimate. Top-down characteristics are used as the input variables to the estimating model. Further details on the types of cost influencing information employed are discussed in Chapter 3, the overall approach to research methodology and validation.

Problem Statement

The literature has shown that artificial intelligence methods can be applied to produce a conceptual estimate with output of suitable accuracy. However, it has been noted that data-driven methods are not being used in practice by highway agencies in the United States. Additionally, conceptual cost estimates are frequently inaccurate and expose highway agencies to public scrutiny over unacceptable construction cost growth.

The main objective of this research is to identify tools that highway agencies may utilize if they choose to adopt data-driven techniques from the literature, thus improving the practical application of data-driven methods. This objective is explored by focusing on the following three questions:

1. *Is there a rational method to sample data that is to be used for artificial neural networks?*
2. *Does adding input attributes (project detail) to 'top-down' estimating methods actually yield further improvements in model performance?*
3. *What data-driven method could be used to better communicate the confidence level attached to the conceptual estimate?*

CHAPTER 3. OVERALL APPROACH TO RESEARCH METHODOLOGY AND VALIDATION

This thesis focuses on conceptual cost estimating of highway projects at one highway agency, the Montana DOT otherwise known as MDT. As such, data used in this study was collected for 189 different highway projects for 5 years of construction at MDT. The section begins by discussing the method used to develop a base ANN used as part of this project. Finally, the global methodology used to meet the thesis objectives are discussed.

Base Artificial Neural Network

Prior to answering the three main research objectives (Chapters 4, 5 and 6) a database of project attributes and construction costs required development. In this section the method to select the project attributes (input variables) and creating the database is discussed. Additionally the validation technique for evaluating the performance of data-driven methods is discussed. Once the database was organized in a commercial spreadsheet, a commercial plug-in to that spreadsheet was used to for the ANN prediction model.

Input variables

Bell and Ghazanfer (1987) concluded that input variables selected have a significant effect on the prediction capability of the model. The same deduction has been reached by at least two other authors of data-driven CCE models (Gunyadin and Dogan 2004; Setyawati et al. 2002). It is during the early stages of creating a data-driven model that attributes need selecting as model-creators usually only have a one-time commitment to collect the cost predictors (Smith and Mason 1997).

Studying previous literature on data-driven CCE models yielded four publications most relevant to highway construction cost estimating. Mahamid (2011) investigated 9 variables in the data-set collected. Al-Tabtabai et al. (1999) also included 9 variables in the data-set collected. Hegazy and Ayed (1998) included 10 input variables. Bell and Ghazanfer (1987) included 2-5 input variables depending on the specific highway project type.

Through meetings held at MDT and the literature studied above, 29 possible input attributes were initially suggested for predicting a typical project's construction cost. The attributes selected for the final model are typically chosen through trial and error, therefore having more input variables available rather than less seemed logical. The technique of selecting the final input attributes through trial and error was reported in at least five other studies on CCE in the construction industry (Creese and Li 1995; Hegazy and Ayed 1998; Bell and Ghazanfer 1987; Gunduz et al. 2011 and Petroutsatou et al. 2012).

Highway agencies typically construct a broad range of project types including bridge construction, pavement preservation, highway maintenance and miscellaneous tasks. The three major work-types conducted at MDT are shown below:

- *Pavement Preservation* – minor rehabilitations and resurfacing
- *Construction* – major highway rehabilitations
- *Bridge* – new bridge construction or major rehabilitation of a bridge

To be able to concentrate data collection efforts and create a methodology, one major project type was selected for investigation in this thesis: pavement preservation projects. This was selected from interviews with MDT as the staff expected that these projects should be the most predictable and provide a suitable test for the estimating methodology.

The desired 29 input attributes, refined from literature review and meetings at MDT, are shown in Appendix A. Each input attribute was aligned with possible measures from the databases discussed in the following section. The 29 attributes were further reduced to the 17 most relevant to pavement preservation cost indicators shown in Table 2. These were selected based on guidance from MDT personnel and data availability for pavement preservation projects. An example is the exclusion of bridge type (steel/concrete), MDT deemed that typically the only bridge work in pavement preservation projects is for deck maintenance and repairs.

Table 2. Proposed input variables trialed in Chapters 4, 5 and 6.

Proposed input attributes
Urban or rural project
Site topography (steep, flat or undulating terrain)
Construction on Native American Reservations
Start and End Stations, Length and Width
Number of bridges in scope
Design Average Annual Daily Traffic (AADT)
Typical Section (depths of surfacing)
Design speed(s)
Intersection signalization and signage
Right-of-way acquisition costs
Traffic Control - closures or detours
Curb & Gutter and Sidewalk
Contract Time
Letting Date
Bridge deck areas
Geotechnical - subsurface & slope recommendations
Extent of Utility relocations and costs

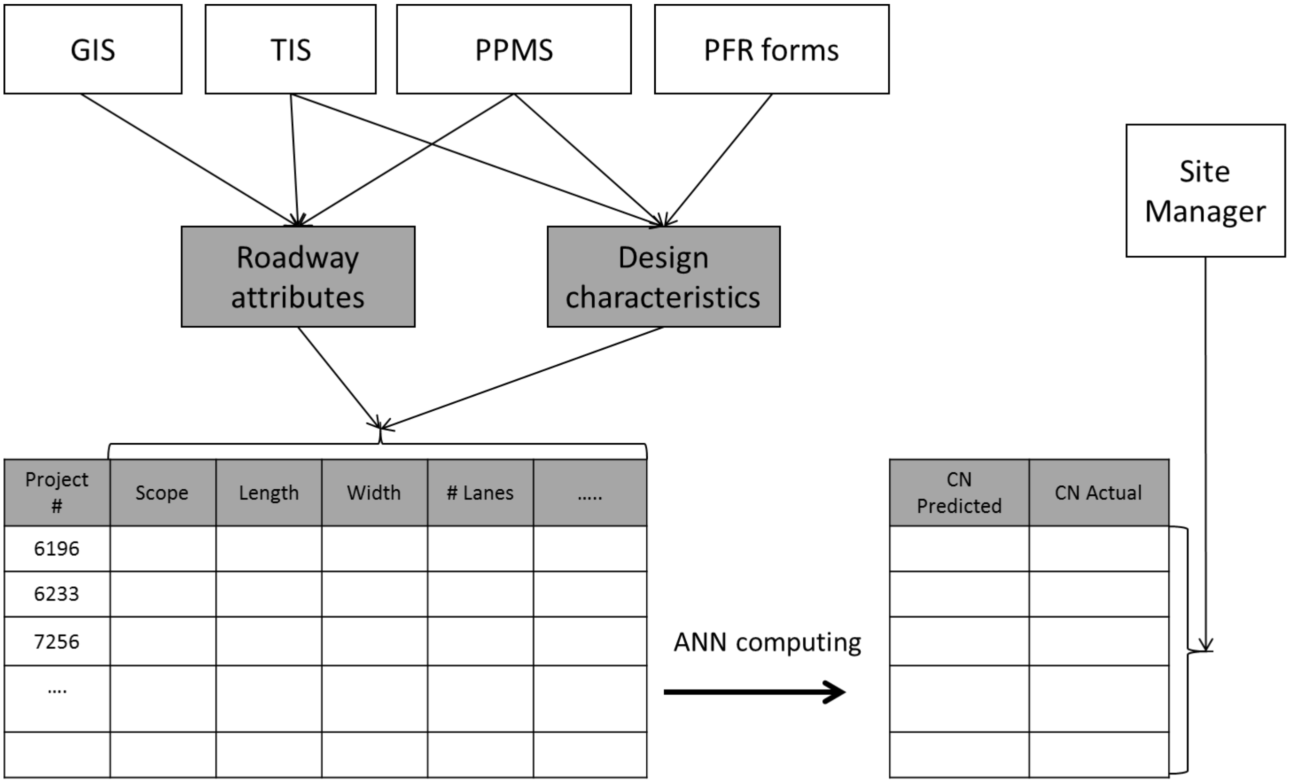
Previous literature published on ANNs and MRAs have used between 2 and 9 input variables. Additionally it was discovered that the final input variables for many publications of data-driven CCE models were selected based on trial and error. The 17 available input variables shown in Table 2 were used as a base in each of Chapters 4, 5 and 6; however, not all 17 input variables were used in each of the models.

Data collection efforts

Because MDT had not yet developed a database specifically for 'top-down' estimating of highway projects, this was included as a task in the overall project MDT research project 8227-001. Multiple databases required combining, along with manual cross-checking, to obtain the attributes necessary for estimating the construction cost with 'top-down' variables. The databases referenced in Figure 3 are:

- GIS: Roadway attributes contained in the geographical information system (GIS) database.
- TIS: Project attributes entered from the construction drawings in the transportation information system (TIS) database.
- PPMS: Data recorded on the Program and Project Management System (PPMS) during the preconstruction activities.

- PFR forms: Conceptual design details completed as a report in the Preliminary Field Review (PFR) forms. This information is textual and the report is required for the transportation committee approval of the conceptual estimate.
- Site Manager: Data stored during the construction phase. Of specific interest in this database was the final construction cost (CN Actual), the dependent variable in the cost estimating model.



Project #	Scope	Length	Width	# Lanes	.....
6196					
6233					
7256					
....					

CN Predicted	CN Actual

Figure 3. Combining multiple databases for cost estimating model

Many of the databases required manual inspection. For example lengths were included in multiple databases (TIS, PPMS and PFR forms), but discrepancies between these numbers required verification. An analysis of the data collected for each input variable is shown in Appendix A. Data extracted from PFR forms required manual extraction and entering. A complexity rating chart was developed for reading the PFR forms, shown in Appendix B, detailing complexity levels and information to be extracted from the forms.

Project construction costs required inflation to a base year to reflect the rising construction costs. The data was collected for construction years 2009-2013. An inflation factor of 3% per annum was applied to the actual construction costs for all projects from the expected mid-point of construction to align with the year 2014 (base-reference). The 3% rate was nominally selected with advice from meetings at MDT and is the rate currently applied to all projects in their construction program. Future research could investigate a more suitable inflation value, however this was not an objective of this thesis.

Validation techniques

In order to validate the usefulness of a data-driven estimating model, the prediction model must be tested. Prediction ability of a model is most easily tested with projects where the final construction cost is known in order to compare the predicted cost to the actual. A test selection of projects from the database must be retained from training the ANN or MRA model. Typically a randomly selected 20%-30% of the data is retained for testing the model (Petroutsatou et al. 2012; Moselhi and Siqueira 1998). For this project 20% was selected in a selection process shown in Chapter 4, the method of which is a major contribution of this thesis.

The error in the data-driven CCE models collected for comparison was calculated using the Mean Average Percentage Error (MAPE) of the testing data. This method is commonly used by authors of data-driven CCE models in the construction industry (Petroutsatou et al. 2012; Gunduz et al. 2011; Mahamid 2011; Lowe et al. 2006; Kim et al. 2004). Calculation of the MAPE is furnished using Equation 1 (Mahamid 2011).

$$MAPE(\%) = \left(\frac{100\%}{n}\right)\sum_{i=1}^{n}\left|\frac{P_i - A_i}{A_i}\right| \tag{1}$$

where:

n = Number of data-points used to test the model

P_i = Predicted construction cost using the data-driven CCE model for the ith project

A_i= Actual construction cost from the historical records for the ith project

Global Methodology

The overall methodology and validation techniques used in this research is illustrated in Figure 4. The base model introduced in Chapter 3 is used to answer all three research problem statement questions. As shown in Figure 4, the motivation for the problem statements were discovered through content analysis and literature reviews into data-driven CCE models and current state-of-the-practice.

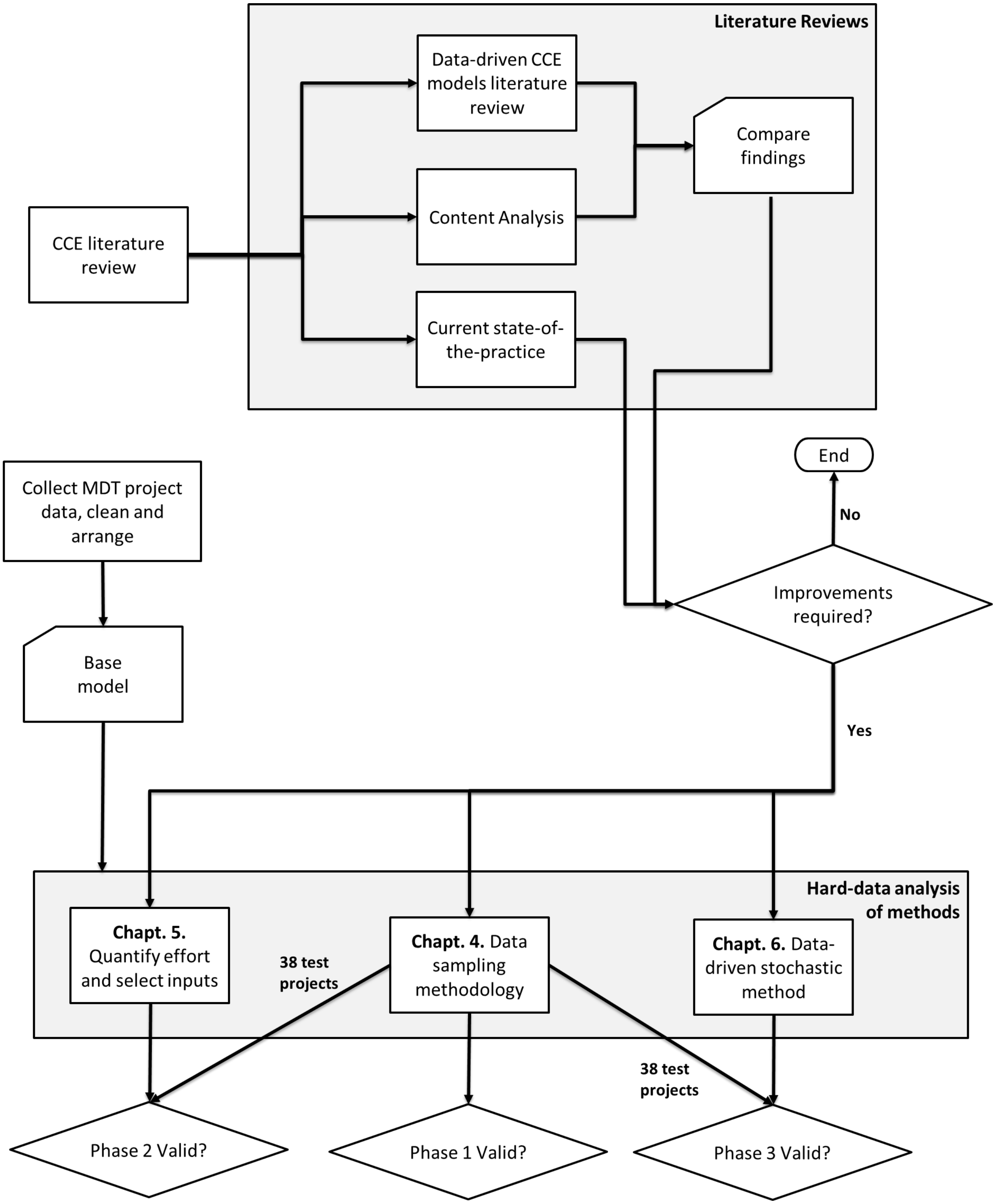

Literature Reviews
CCE literature review
Data-driven CCE models literature review
Content Analysis
Current state-of-the-practice
Compare findings
End
No
Improvements required?
Yes
Collect MDT project data, clean and arrange
Base model
Hard-data analysis of methods
Chapt. 5. Quantify effort and select inputs
38 test projects
Chapt. 4. Data sampling methodology
Chapt. 6. Data-driven stochastic method
38 test projects
Phase 2 Valid?
Phase 1 Valid?
Phase 3 Valid?

Figure 4. Global methodology covered in this thesis

CHAPTER 4. RATIONALLY SELECTING DATA FOR HIGHWAY CONSTRUCTION COST ESTIMATING AT THE CONCEPTUAL STAGE

Gardner, B., Gransberg, D. D., and Jeong, H. S. (2015). “Rationally Selecting Data for Highway Construction Cost Estimating at the Conceptual Stage.” To be submitted to ASCE *Journal of Computing in Civil Engineering*.

Abstract

Over the past 30 years there has been little improvement in construction cost estimating confidence, despite significant advancement in computing capabilities and data availability. During this period the literature reveals a number of highly accurate prediction models, however many are supported by databases containing very few data points. The practicality of these models is limited due to their narrow scope and lack of defined sampling techniques used to select their data points. Models to estimate construction costs at early stages of project development using artificial neural networks and multiple-regression analysis have been developed for some time, but they are not being used in practice by US state DOTs. This paper investigates how data point selection limits the practical performance of these models and a contributing reason why sophisticated models have not yet been implemented by DOTs. A total of 20 conceptual cost estimating models, using artificial neural networks and multiple-regression analysis, were assessed in this study. While a data-driven conceptual cost estimating model may appear accurate, not appropriately sampling the data inputs will result in a model with little practical application and therefore not suitable for use in industry. This study found that data used to train conceptual cost estimating models need to include attributes reflective of the projects in the total population of data. As a result, this research proposes a rational method to sample project data.

Introduction

Previous literature has proven the ability of ANNs and MRA to predict the conceptual construction cost of projects to suitable accuracy, this was discussed in the artificial intelligence section of Chapter 2. However, it was also noted in Chapter 2 that no highway agency is currently known to utilize this data-driven, ‘top-down’ artificial intelligence method to conduct their conceptual cost estimate. This is despite the proven ability of data-driven methods and the continued lack of confidence with the conceptual development stage estimate of construction cost (Flyvbjerg et al. 2002; Schexnayder et al. 2003). This chapter investigates the practicality of models published in the literature and investigates contributing reasons why they have not yet been implemented in practice.

Background

Literature supports the case that more data in the prediction model results in improved reliability and accuracy. When Bell and Ghazanfer (1987) created an MRA model using 174 highway projects their research concluded - "larger data-sets tend to reinforce the reliability of the model." This judgement is supported by many authors of data-driven CCE models (Setyawati et al. 2002, Gunaydin and Dogan 2004, Tatari and Kucukvar 2011 and Gunduz et al. 2011) where these authors had between 16 and 74 projects in their databases and used a mixture of ANN and MRA for their prediction models.

In 1998 Elhag and Boussabaine recommended that future CCE models should exploit more than the 30 training data points they used in their research to improve the model accuracy. Following this, in 2002, Emsley et al. created a model with nearly 300 projects to specifically address the deficiencies in the ANN created by Elhag and Boussabaine. Other data-driven CCE models created with a notable size of database: Kim et al. (2004) and Lowe et al. (2006) used 530 and 286 historical projects respectively for their databases.

Weaknesses in the size of training data contributing to the limited practical application of data-driven CCE models has been suggested but not yet thoroughly investigated. Setyawati et al. (2002) recommended that the effects of more data in building and construction estimating need to be further studied. This paper aims to contribute to understanding the size of training data selection and model reliability in relation to the construction industry.

Objective

The objective of this paper is to evaluate the use of data-driven CCE models to help determine the limiting factor for practical use in industry. As such this paper explores 20 construction CCE models using ANN or MRA to determine the impact that the quantity of data utilized for training has on model accuracy. Further, this research investigates a rational sampling method for when the entire data population is not utilized. Of the CCE literature reviewed there were no reports on the sampling method used for training or testing the model or size of the total population of historical projects available.

Methodology

Literature on published CCE models involving ANN and MRA were reviewed. It was important to identify only models that were relevant to this study. Three criteria were used to ensure this:

1. the study is related to the construction industry,
2. input variables are obtainable at the early design stage,
3. the output variable is a construction cost estimate of the project.

If the input variables of the data-driven CCE models were simply the bill of quantities then it was deemed a ‘bottom-up’ or a detailed estimate of the construction cost and these models were excluded from the study.

A commercial search tool for document content (Bazeley and Richards 2000) was used to search for relevant publications, organize and record the analysis. A broad search was conducted initially of all the collected publications. The number of case studies was then reduced to only 16 publications containing 20 data-driven CCE models with the necessary information to conduct an effective content analysis. The estimating error of the data-driven CCE models and the number of data points used were recorded for comparison and to investigate alignment with literature suggestions on this topic.

Results

The data gathered from CCE publications are shown in Table 3 and outline the brief scope for the types of projects being predicted. Some publications analyzed their database using both ANNs and MRA to compare the relative performance of the two different techniques, whilst others just performed one technique. The model error was calculated using the MAPE method presented in Chapter 3, Equation 1. Where authors had not used this method then our research team recalculated the error to enable direct comparisons of performance.

Table 3. Construction cost estimating models studied

CCE literature	Data points	ANN estimating error	MRA estimating error	Brief project scope
Petroutsatou et al. (2012)	149	4.65%	–	Tunnels in Greece
Mahamid (2011)	131	–	13.0%	Highway (various sizes)
Gunduz et al. (2011)	16	5.76%	2.32%	Light rail track works in Turkey
Lowe et al. (2006)	286	–	19.30%	Buildings in United Kingdom
Petroutsatou et al. (2006)	149	–	9.6%	Tunnels in Greece
Kim et al. (2004)	530	3.0%	7.0%	Residential Buildings in Seoul, South Korea
Gunaydin and Dogan (2004)	30	7.0%	–	RC 4-8 story residential buildings in Turkey
Emsley et al. (2002)	288	16.6%	–	Buildings
Setyawati et al. (2002)	41	13.4%	9.2%	Education Building Construction
Al-Tabtabai et al. (1999)	40	9.1%	–	Highway Construction
Hegazy and Ayed (1998)	18	19.33%	–	Highway Construction in Newfoundland, Canada
Elhag and Boussabaine (1998)	30	17.80%	–	School Construction
Moselhi and Siqueira (1998)	34	10.77%	14.76%	Steel framed low-rise buildings
Creese and Li (1995)	12	8.24%	–	Timber Bridges
Sanders et al. (1992)	11	–	6.0%	Urban Highway Bridge widening in Alabama
Bell and Ghazanfer (1987)	174	–	17.0%	Highway Construction Maintenance projects

– = data not applicable to that publication

Since Bell and Ghazanfer (1987) concluded that "larger data-sets tend to reinforce the reliability of the model" DOTs investigating the possibility of data-driven cost estimating would expect equal if not more training data to be used in the data-driven CCE models for reliability and confidence. Figure 5 shows that only three authors in the study population used more than the 174 historical construction projects that Bell and Ghazanfer used in their data-driven CCE model in 1987. This is surprising given the explosive computing capabilities and data storage capacity that has occurred since Bell and Ghazanfer published their results. Of the data-driven CCE models studied six authors reached the same conclusion as Bell and Ghazanfer in 1987, yet there are still many published models using very few historical construction projects in their ANN or MRA analysis.

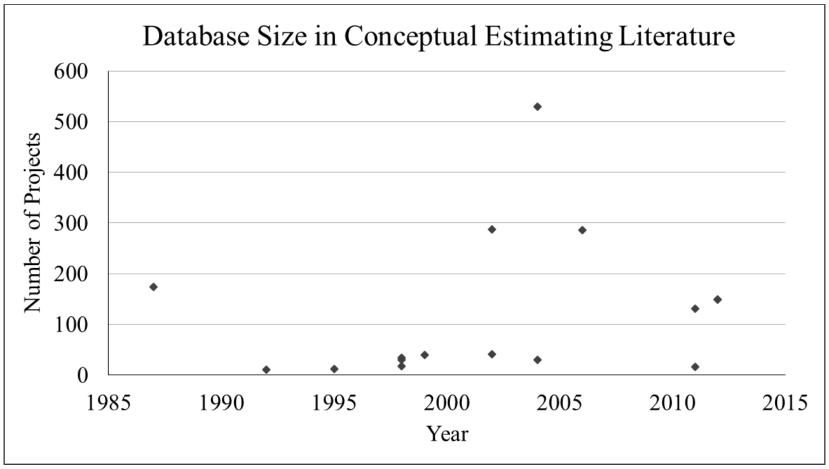


Figure 5. Timeline showing the database used in data-driven CCE models.

Literature from data-driven CCE models support the hypothesis that lack of data will result in unreliable CCE (Bell and Ghazanfer 1987; Elhag and Boussabaine 1998; Setyawati et al. 2002; Gunaydin and Dogan 2004; Tatari and Kucukvar 2010; Gunduz et al. 2011) and could therefore be a reason for limited industry use. However, findings from the content analysis of the 20 data-driven CCE models investigated in this study show when the accuracy of the prediction model is plotted against the number of data points, in Figure 6 there is little to no trend. The arrow shows the direction of the trend expected from literature findings, as the number of data points used in the model then the estimating error should decrease. There is an unexplained cluster of points in the bottom left of the plot; these case studies are circled and report high accuracy with a low number of data points used.

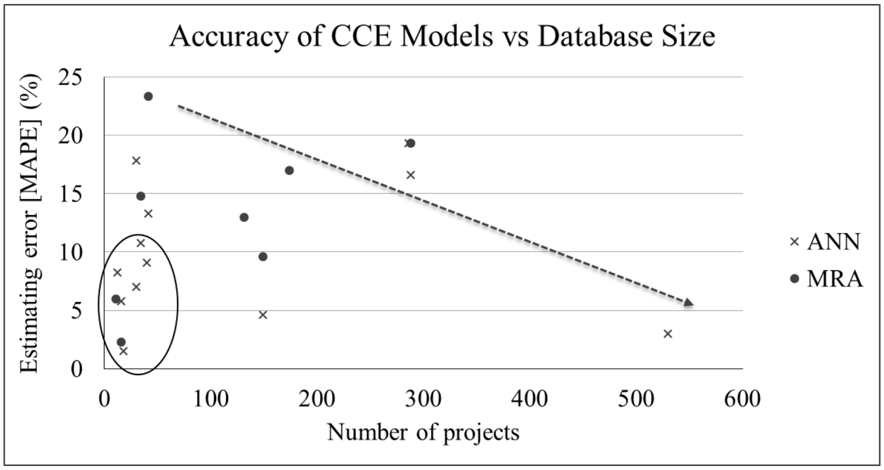


Figure 6. Accuracy of data-driven CCE models published and database size

The content analysis results from Figure 6 conflict literature suggestions, whereby increasing the database within the CCE models will result in improved reliability and accuracy (Bell and Ghazanfer 1987; Elhag and Boussabaine 1998; Setyawati et al. 2002; Gunaydin and Dogan 2004; Tatari and Kucukvar 2010; Gunduz et al. 2011). An explanation for this could be that some of the published data-driven CCE models have been built for projects of very narrow scope.

Creese and Li (1995) created a model specifically for timber bridges using only 12 projects. Sanders et al. (1992) limited scope of their data-driven CCE model to bridge widening only, using 11 projects. Sanders et al. recognized that the model was only useful for interstate bridge widening's stating that the "model presented in this report obviously has very limited application."

Gunduz et al. (2011) created a model for light rail track works with only 18 projects and achieving nearly 2% prediction accuracy. Validation of the light rail model was based on only two projects. Additionally, the light rail model estimated the trackworks portion of the rail projects only and did not account for other infrastructure in the project (Gunduz et al. 2011).

Data-driven CCE models that are only accurate for a very narrow scope of work do not provide general utility due to the extremely limited group of projects on which they can be applied. Typical DOT projects range in scope from simple to complex and would therefore require many different data-driven CCE models to meet their needs. Furthermore, even if the models could theoretically be built, many if not most would not contain enough data points to be reliable.

It leads one to suspect that CCE publications using a small number of data points in their analysis may not have included the entire population of historical projects for the defined scope and purpose of the estimating model. While the literature does not fully explain the rationale for not using the entire population, there are potentially two practical reasons for this:

1. the researchers did not have access to the complete agency project databases, or
2. the effort of collecting each project was significant and tedious resulting in a small number of historical projects used in the analysis.

Of the CCE literature reviewed there were no reports on the sampling method or size of the total population of historical projects used for training or testing the model.

Discussion of Results

Literature study supports the hypothesis that increasing the number of training data points in CCE models improves the accuracy and reliability (Bell and Ghazanfer 1987; Elhag and Boussabaine 1998; Setyawati et al. 2002; Gunaydin and Dogan 2004; Tatari and Kucukvar 2010; Gunduz et al. 2011). However a content analysis of 20 data-driven CCE models found no trend in the improvement of performance with increased number of data points. Instead, this study found that some estimating models were reporting very accurate results using few data points to train their data-driven CCE models. Further analysis revealed that these models may be of very narrow scope, limiting the practical application for use by DOTs.

Published work in the manufacturing (Bode 2000) and aeronautical industry (Rajkumar and Bardina 2003) reached the same conclusion; more data improves accuracy of the data-driven model. In these fields more data used in training produced improved predictions, however this improvement had diminishing returns after a point. Rajkumar and Bardina produced over 7000 data points in the laboratory for their ANN model studying aerodynamic coefficients.

The challenge with data collection in the construction industry is the availability of data. Historical data used in CCE models comes from completed projects which can cost millions of dollars each. The number of projects that can be included in the database is limited to those completed each year, which is often quite low due to the high costs of each. More importantly, each construction project is normally unique in many ways due to the scale of the transportation infrastructure. Unlike products in the manufacturing industry, data cannot simply be regenerated in a laboratory thousands of times. The effort required to collect construction project data produces the need for a rational data selection method, allowing an individual to accurately represent the entire project population with a sample.

This research next investigates and then proposes a possible sampling method by studying the distribution of key attributes in a project population to rationally sample the data. The purpose of this is to propose a method going forward for sampling the data to improve model credibility. Such a method could increase the application of data-driven CCE models for DOTs.

Rational sampling method

Proposed Technique

A rational sampling method should be used to select data-points for data-driven CCE models when the entire population of data is not going to be utilized. This ensures that the data sample appropriately represents the population being modeled, and information is not unintentionally misleading. The proposed technique is shown in Figure 7. First the population of historical projects is defined in terms of scope and size. Defining the scope of the project allows readers and practitioners to understand what the data-driven CCE model can be used for (it's purpose). It is also important to understand the sample of projects actually used in the prediction model relative to the total population. This is similar to reporting on a non-response rate by statisticians when completing surveys (Dillman et al. 2009; Fink 2009).

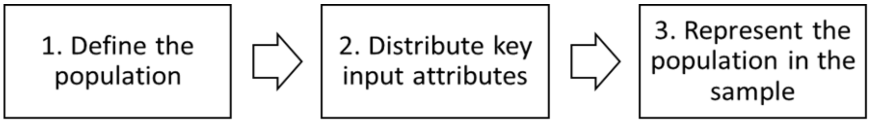

1. Define the population
2. Distribute key input attributes
3. Represent the population in the sample

Figure 7. Proposed rational sampling steps

The distribution of key input variables must also be studied. These are anticipated to be input variables that have the greatest contribution to the end accuracy of the model. Not selecting a representative distribution of key attributes in the sample may limit the practical application of data-driven CCE models for predicting the construction cost of the population in the future.

Next, if the entire population of data is not going to be used in the CCE model then a sample size needs to be nominated. It is justifiable to not use the entire population of data due to computing limitations or time and effort restraints to collect the entire database for all attributes, especially when the population is large with a broad scope. Finally the distribution of key attributes in the population needs representation in the sample to be reflective of the population. To demonstrate how this rational method could be applied an illustrative example is provided in the following section using an ANN data-driven model.

Illustrative Example

Step 1: Define the population: A total of 850 projects were made available to the research team from MDT for analysis. This database included all highway projects completed from 2007 until 2015. The population was further defined to pavement preservation projects only. This left a total of 431 historical projects available. Five consecutive years of projects in the design phase from 2009-2013 were selected and the population further defined to chip seal, thin lift overlay or mill & fill projects, the three main major work-types, all less than $5M in value. A total of 189 projects remained for analysis – this was our research population of data.

Step 2: Distribute key input attributes: The base ANN model developed in Chapter 3, with a database of 189 projects defined above, was trained with a randomly selected 80% of the projects. This trained model was then tested with the remaining 20% of projects, these test projects were not used to train the model. From the tested model two of the input variables were deemed to be the most sensitive to the construction cost, this was the highway classification and length of the projects. The distribution of these two attributes across all 189 projects was analyzed visually and is shown in Figure 8a.

Table 4. Input variables used

Proposed input attributes
Urban or rural project
Site topography (steep, flat or undulating terrain)
Construction on Native American Reservations
Start and End Stations, Length and Width*
Number of bridges in scope
Design Average Annual Daily Traffic (AADT)
Highway Classification*
Typical Section (depths of surfacing and aggregate)
Traffic Control - closures or detours
Curb & Gutter and Sidewalk
Contract Time
Bridge deck areas
Geotechnical - subsurface & slope recommendations
Extent of Utility relocations and costs

**denotes attributes analyzed*

Step 3: Represent the population in the sample: A test sample of 38 projects (accounting to 20% of the database) was separated from the 189 projects in the database. The 38 projects were selected and removed by iteratively selecting projects until distribution of the two key attributes (from Step 2) aligned with the distribution in the entire population. This left 151 projects available to train a model. Selecting test data reflective of the population in this proposed method will test the true performance of the cost estimating model against its intended end-use.

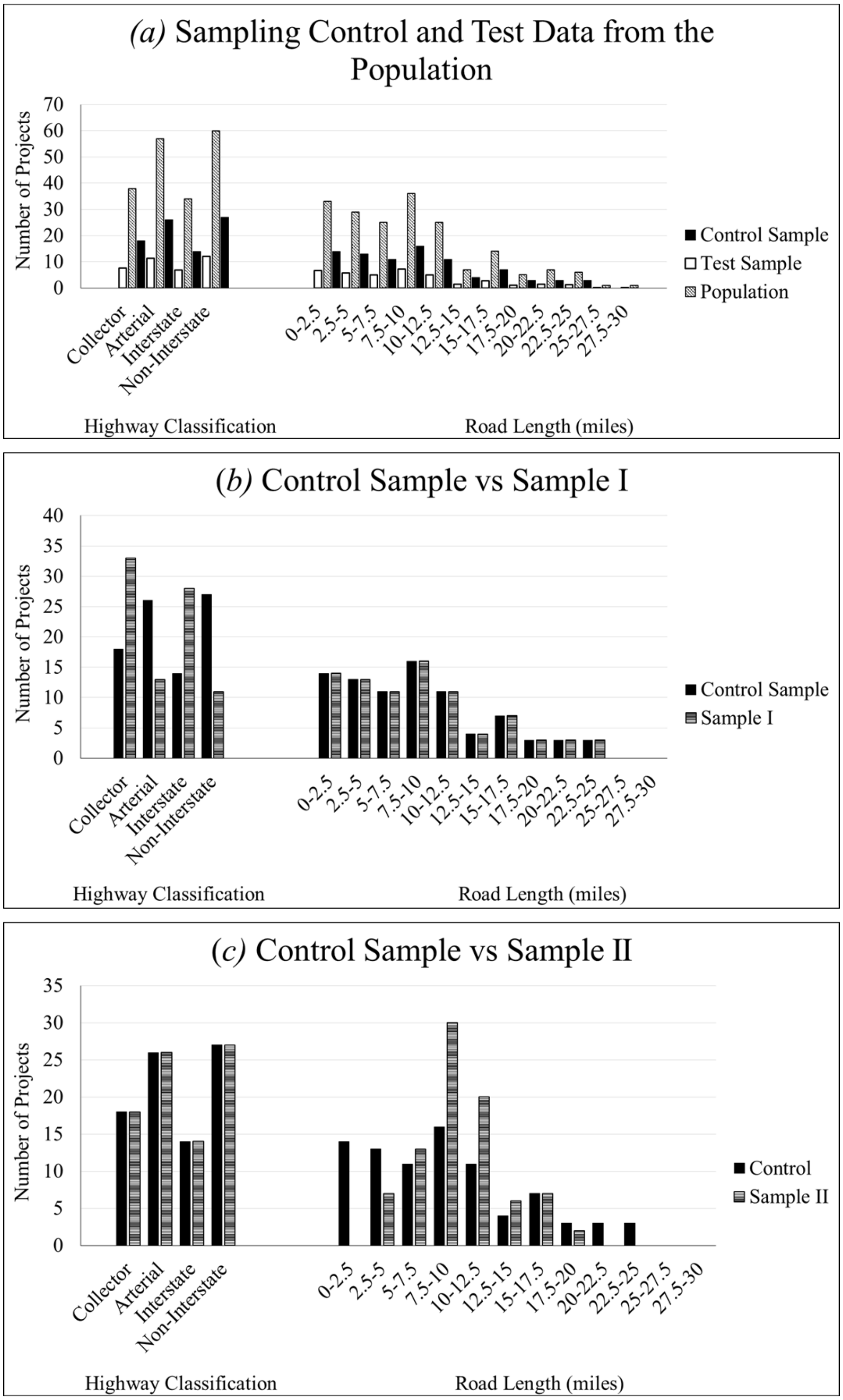


Figure 8. Distribution of Sample I, II and III based on two key input variables

Next, a control sample of 85 projects was selected from the remaining 151 available projects in the training data-set. This was completed iteratively in order to match the distribution of highway classification and length from the population in the control sample. The scale match of the distribution for projects in the control sample against the population is shown in Figure 8a.

For the purposes of validating this method two additional samples of 85 projects were selected from the 151 training projects, these are Sample's I and II. In each of the samples one of the two key attributes were deliberately misrepresented relative to the control sample. The highway classification was misrepresented in Figure 8b (Sample I) and the lengths of the projects were misrepresented in Figure 8c (Sample II) relative to the distribution in the control sample.

Results: The 14 attributes from the 151 remaining historical projects were then used to train the ANN model against the actual construction costs from the database. Two different artifical neural network configurations were trialed. The Generalized Regression Neural Network (GRNN) was found to perform superior to the Multi-Layer Feedforward (MLF) network also available in the software. The historical projects not included in the training of the artificial neural network were then tested in the model. The plot of predicted construction costs versus the acual construction cost for the 38 test data-points is shown in Figure 9.

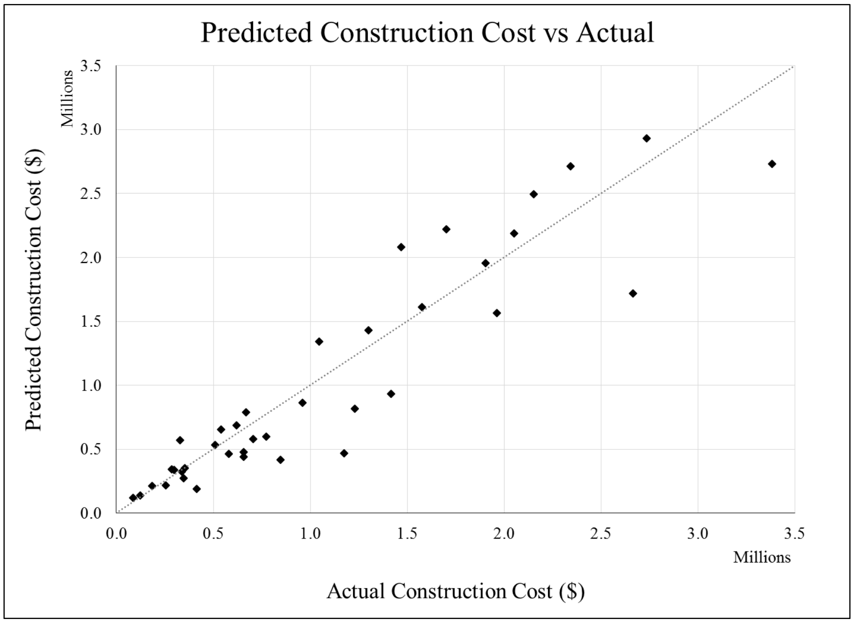


Figure 9. Validating the artificial neural network with the test data

The MAPE for the test 38 projects was 22.9%. This is well within the expected accuracy of the construction estimate at the planning stage suggested in the AASHTO *Practical Guide to Cost Estimating* (2013) where -40% to +100% is accepted at the conceptual development stage, shown in Table 1 (Chapter 2). Further improving the accuracy of this model was not the goal here, so research into sampling this population of 151 training projects continued. A separate model was trained and tested for the Control Sample, Sample I and Sample II. The same 38 projects were used to the test the error of these trained models.

Results of all four ANN models created are shown in Table 5. It was not surprising that no single sample out-performed predicting the construction costs of the 38 test projects than using the entire population (151 projects) to train the model. This is in agreement with literature, from the construction industry and other fields, that states the use of more data improves the accuracy and reliability of the model (Bell and Ghazanfer 1987; Setyawati et al. 2002; Gunaydin and Dogan 2004; Tatari and Kucukvar 2010; Gunduz et al. 2011; Rajkumar and Bardina 2003; Bode 2000).

Table 5. Error in the testing data

Sample	MAPE with the test data
Entire Population (151 projects)	22.9%
Control Sample (85 projects)	32.5%
Sample I (85 projects)	38.0%
Sample II (85 projects)	40.5%

Sample I and Sample II also performed notably worse at predicting the construction cost in comparison to the Control Sample. On visual inspection of Sample II (Figure 8c) the distribution of ‘length’ attributes was much more significantly misrepresented than the ‘highway classifications’ in Sample I (Figure 8b) as the extreme highway length values (high and low) were truncated. This only resulted in a small increase in the estimating error from 38.0% to 40.5%. This finding suggests that some attributes are more sensitive to sufficient representation in the sample database and do not need to exactly match that of the population. Further research needs to be completed to find a relationship between the level of representation in the sample required to appropriately predict the construction cost without using the entire population of data.

Other industries are focusing on ‘big-data’ for the data-analytics and decision analysis. The transportation industry is currently lagging behind in its use of historical data, specifically in the area of cost estimating. Data-driven techniques for CCE of highway projects have proven results in the literature. However, when DOTs are searching for published data-driven CCE models they need to be aware of the limits to their practical application; a data-driven CCE model may appear to perform well but without rational sampling of the data and suitable scope definitions a DOT cannot be confident in these techniques.

Conclusion

Literature from both construction and manufacturing industries support the concept that more data increases the accuracy and reliability of data-driven CCE models (Bell and Ghazanfer 1987; Setyawati et al. 2002; Gunaydin and Dogan 2004; Tatari and Kucukvar 2010; Gunduz et al. 2011, Bode 2000; Rajkumar and Bardina 2003). Despite this widely held belief, a content analysis of 20 data-driven CCE models revealed that some models had a very low prediction error despite using few projects to train the model. A reason for this result is the narrow scope of the projects included in the database and lack of test data. These two attributes make the use of data-driven CCE models undesirable for use by DOTs.

Despite the small databases in the CCE models, the literature has remained silent on methods used to select the data used. To help improve the validity of CEE models for future industry use, this paper suggests a rational method to effectively represent a database without using all data points. An illustrative example using artificial neural networks was provided to demonstrate how such a method would be applied in practice. It was found that key attributes need sufficient representation in the sample of data.

Regardless of the vast improvement in computing technologies over the past 30 years, no great advancement in CCE accuracy has been made, preventing DOTs from using these technologies within their work. This paper found contributing reasons for this decision to be that many published data-driven CCE models have a very narrow scope, lack of confidence in the sizes of some databases used and no sampling method used for selection of projects.

CHAPTER 5. QUANTIFYING EFFORTS FOR DATA-DRIVEN CONCEPTUAL COST ESTIMATING FOR HIGHWAY PROJECTS

Gardner, B., Gransberg, D. D., and Jeong, H. S. (2015). “Quantifying Efforts for Data-Driven Conceptual Cost Estimating For Highway Projects.” To be submitted to the ASCE *Journal of Construction Engineering and Management.*

Abstract

A modern dilemma has emerged in light of ever improving technological advances; enlarged data-collection efforts do not yield a proportional increase in knowledge. Storing more data than is necessary, without receiving any useful additional benefit, is not only resource intensive but also requires additional funding to collect and manage it. Data-driven models using historical project attributes to estimate future construction costs, such as multiple-regression analysis and artificial neural networks are both proven techniques found in the literature that highway agencies could adopt for conceptual estimating. This research noted that the literature using these techniques have been solely focused on estimating model performance with little to no focus on the level of effort required to conduct the conceptual estimate. It is commonly believed that using more input data enhances estimate accuracy. However, this paper will test the concept that using more input variables than necessary in the conceptual estimate overcomplicates the conceptual model without a commensurate increase in accuracy. Conceptual estimates using the minimum amount of input data to produce an estimate with a reasonable level of confidence is more cost effective than current practices. It allows designers and estimators to focus their time on advancing project development, instead of investing time into projects that may never advance past the initial conceptual stage. Furthermore, reducing data requirements saves highway agencies time and money on storage of unnecessary project information. This paper quantifies the effort expended to undertake estimates for both artificial neural network and multiple-regression analysis models used for the conceptual estimate. The paper concludes that input variables which have a large influence on the final predicted cost and require a low amount of effort are desired in data-driven conceptual cost estimating models.

Introduction

In public works, the budget for a project is often established at a point in project development where the estimator has the least amount of design detail from which to compute an estimate (Bode 2000). Taking federally-funded highway projects as an example, the budget is formally set when the project is assigned a federal project identification number (PIN) and included in the STIP (FHWA 2015; Anderson et al. 2007). The estimate is usually used during early planning stages to conduct initial feasibility studies, and both engineers and planners realize that the accuracy of the initial cost estimate is a function of the level of design detail available at the time of the estimate. To account for the anticipated change in project scope as the development process proceeds, a standard contingency based on a percentage of the total estimate is added (Minassian and Jergeas 2009). This kind of estimate is termed a 'top-down' estimate because it relies on parametric cost factors such as lane-miles, location, project type, etc. rather than a 'bottom-up' estimate whose basis are the quantities of materials needed on the project (Kim et al. 2012).

The conundrum faced by engineers in public works is that in order to receive the authorization to expend funds to advance the project to completion the official budget is based on a figure derived with the least amount of project-specific technical information (Bode 2000; FHWA 2015). If the figure is too conservative, the project may not be receive authorized funding necessary to advance to the next preliminary engineering stage. As a result, it becomes important to take the initial cost estimate seriously and utilize the available information that has the highest influence on the bottom-line while not allocating precious time and resources to a project that ultimately will not advance. Additionally, the time period to conduct the estimate is typically limited in the feasibility stage (Gunduz et al. 2011), but the estimate requires sufficient accuracy for benefit-cost analysis and prioritizing budgets (Anderson et al. 2007). Therefore, the objective of this paper is to explore a solution that can be used to complete critical initial estimates with high impact data that requires the minimum level of effort for the estimator to obtain.

Conceptual Cost Estimating Effort at Highway Agencies

Highway agencies cannot afford to over-invest their design and planning resources in projects at the conceptual stage. If less effort can be expended at the conceptual stage, then an estimator's time can be better applied in the later design estimating stages shown in Figure 1 (Chapter 1). Any investment in the project at the conceptual stage could be rendered worthless if a project is not selected for further development following a benefit-to-cost analysis or a needs assessment. In the context of structural steel buildings only 15 percent of those that reach the conceptual stage ever get constructed (Moselhi and Siqueira 1998).

No matter the CCE technique employed by highway agencies or suggested in the literature, a particular level of project scope definition (or conceptual design effort) is required in order to conduct a cost estimate. Sanders et al. (1992) observed this balancing act between efforts expended and estimate accuracy, stating “there is an inverse relationship between the accuracy of an estimate and its preparation cost. At some point, increased accuracy cannot justify the additional costs incurred.” The earlier that the initial estimate is developed, the lower the level of effort expended on project definition required for CCE, which translates into lower costs and fewer resources. This means that estimators and designers can focus their efforts on projects which have advanced past the planning stage and are likely to reach construction.

Data-driven CCE Models – Prior Studies

CCE techniques reviewed in this research include both ANN and MRA models. The benefit of data-driven techniques is the ability to use historical project information for forecasting and the speed at which this can be achieved. Gunduz et al. (2011) recognized this stating “reliable cost estimates are required within a very limited time period at the feasibility stage,” and the research in their paper concentrated on the use of ANN and MRA models to produce fast and accurate results.

Performance of data-driven CCE models is subject to variations in model architecture and parameters; this includes the input variables used, number of hidden layers and nodes in the ANN model, and data-set size. The effects of model architecture and parameters have been studied in data-driven CCE models published in the literature (Setyawati et al. 2002; Mahamid 2011; Petroutsatou et al. 2012). A detailed content analysis on the number of input variables is completed in the next section.

Literature analysis

Previous authors of data-driven CCE model research have remained silent on the effort to collect, store and use databases to conduct the cost estimates. As a result, this research analyzed the data-driven CCE models published in the literature to observe how many input variables are being used and resultant error. Collection and storage of data from historical projects requires time and resources of which highway agencies have a limited quantity. Further cost influencing information gathered later in the project life-cycle can be included in more detailed ‘bottom-up’ design stage estimates. The literature analysis was a starting point of this research to see if additional inputs improve estimating accuracy.

The same 16 publications on data-driven CCE models from Chapter 4 were studied. From each of the publications both the MAPE and the number of input variables used to produce their best performing model was collected. The results of the content analysis are shown in Table 6.

Table 6. Construction cost estimating models studied to understand input variables

Author	Input Variables	ANN estimating error	MRA estimating error	Brief Project Scope
Petroutsatou et al. (2012)	5	4.65%	–	Tunnels in Greece
Mahamid (2011)	9	–	13.0%	Highway (various sizes)
Gunduz et al. (2011)	17	5.76%	2.32%	Light rail track works in Turkey
Lowe et al. (2006)	12	–	19.30%	Buildings in UK
Petroutsatou et al. (2006)	5	–	9.6%	Tunnels in Greece
Kim et al. (2004)	9	3.0%	7.0%	Residential Buildings in Seoul, Korea
Gunaydin and Dogan (2004)	8	7.0%	–	4-8 story residential buildings in Turkey
Emsley et al. (2002)	5	16.6%	–	Buildings
Setyawati et al. (2002)	2	13.4%	9.2%	Education Building Construction
Al-Tabtabai et al. (1999)	9	9.1%	–	Highway Construction
Hegazy and Ayed (1998)	10	19.33%	–	Highway Construction in Newfoundland, Canada
Elhag and Boussabaine (1998)	4	17.80%	–	School Construction
Moselhi and Siqueira (1998)	4	10.77%	14.76%	Steel framed low-rise buildings
Creese and Li (1995)	3	8.24%	–	Timber Bridges
Sanders et al. (1992)	10	–	6.0%	Urban Highway Bridge widening in Alabama
Bell and Ghazanfer (1987)	5	–	17.0%	Highway Construction Maintenance projects

Note: – = indicates that data is not applicable to that publication

The results from the literature content analysis found in Figures 10a and 10b show that previous publications are achieving lower error through more input variables. Both plots in Figure 10a and 10b show diminishing returns with a smaller reduction in error as each input variable is added, this is highlighted by the best fit curves being negative power curves. The relationship is much stronger with the MRA models (Figure 10b) in the literature with the power curve coefficient of determination (R^2) value being 0.7211. When the obvious outlier in the ANN group (Figure 10a) is removed then the R^2 value in that plot increases from 0.1462 to 0.335.

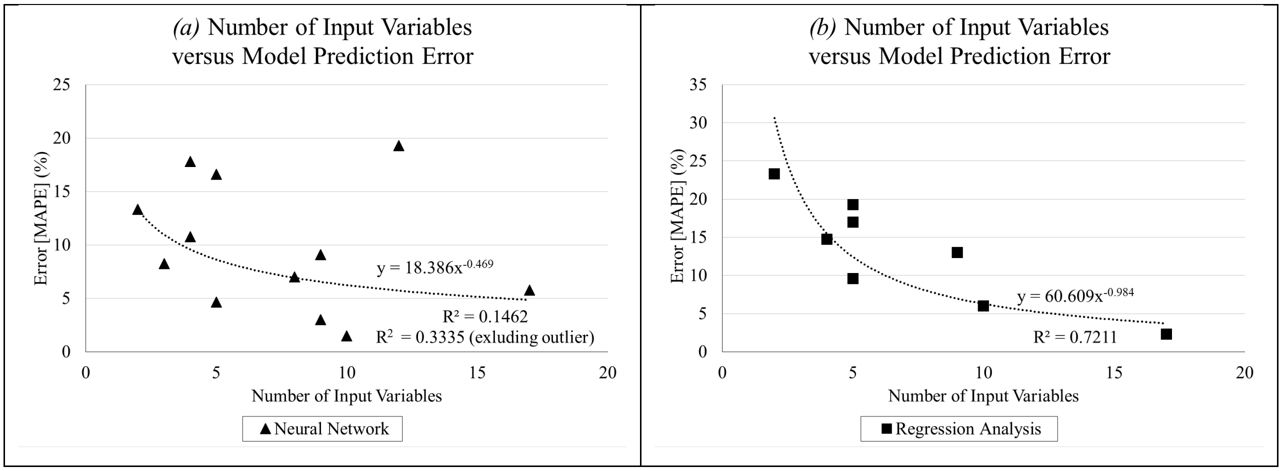


Figure 10. Literature analysis of inputs versus error

A weakness of this conclusion is that the literature is for projects of many different scopes. Additionally, none of these past studies have compared their input variables with the perceived level of effort to obtain them for every project, meaning that effort required to populate the model and its performance have not been directly compared. This leads one to infer that the results reported in the literature contain an underlying assumption that each input variable requires equal estimating effort. Therefore, that in the body of knowledge will be filled by the results of this chapter, which will specifically quantify input variable effort and prove that not all input variables require the same level of effort to compute.

The requirement to minimize CCE effort is also recognized in other industries outside of construction. Verlinden et al. (2008) created an ANN to calculate the cost of sheet metal manufacturing for customers; the research recognized the necessity to provide customers of sheet metal a swift quotation, albeit at the cost of possibly reduced accuracy. In another study, Walczak (2001) created an ANN to predict a foreign exchange rate. Walczak's study found there was no need to utilize the entire available database and that only a few years of data was necessary to provide reasonable confidence. Walczak concluded that this would have a significant effect on model development cost savings, where "the cost is not only financial, but also the development time and effort."(Walczak 2001).

Research objective

This paper proposes a new CCE objectives hierarchy, illustrated in Figure 11, to evaluate the performance of data-driven CCE models. Previous data-driven CCE models are focused on the prediction accuracy (Objective 1), but this research investigates the effort expended (Objective 2) in gathering the input information for the models.

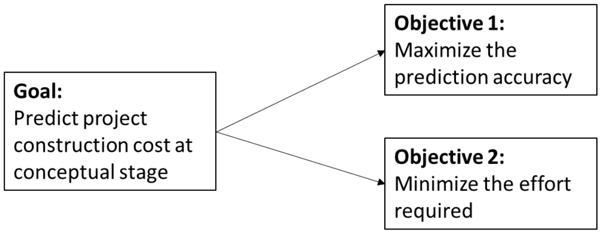

Goal:
Predict project construction cost at conceptual stage
Objective 1:
Maximize the prediction accuracy
Objective 2:
Minimize the effort required

Figure 11. Proposed dual-objective hierarchy tree for conceptual cost estimates

The objective of this paper is to evaluate the effort expended for data-driven CCE models. Specifically the paper focuses on two questions:

1. Can a framework be created to select inputs that help meet the dual-objective goal of maximum performance with minimal effort?
2. Is there an optimum number of input variables that highway agencies should be collecting to minimize the effort for data-driven CCE models?

The outcomes of this research should help both researchers and practitioners to focus on both objectives during the CCE stage, allowing them to estimate the projects construction cost at an early stage of project development with the least amount of effort but with the optimal performance.

Research Methodology

To validate the input selection framework and determine if an optimum level of input variables exist a combination of perceptional survey data was used with real project data to predict the construction cost. The research steps are shown in Figure 12 below. In step 1, a survey was conducted to grasp perception on the level of effort required for different inputs to the conceptual estimate. The dual-objective input selection method, proposed as part of this research, was then utilized in step 2. Next, the estimating error for each model was recorded using the proposed input selection order (step 3a) and then it was repeated using the input selection order in reverse (step 3b). Finally step 4 compares the cumulative perceived effort for each construction cost estimate to the estimating error achieved. In this step the proposed input selection method (step 3a) is compared to completing the task in reverse order (step 3b) in order to validate framework effectiveness.

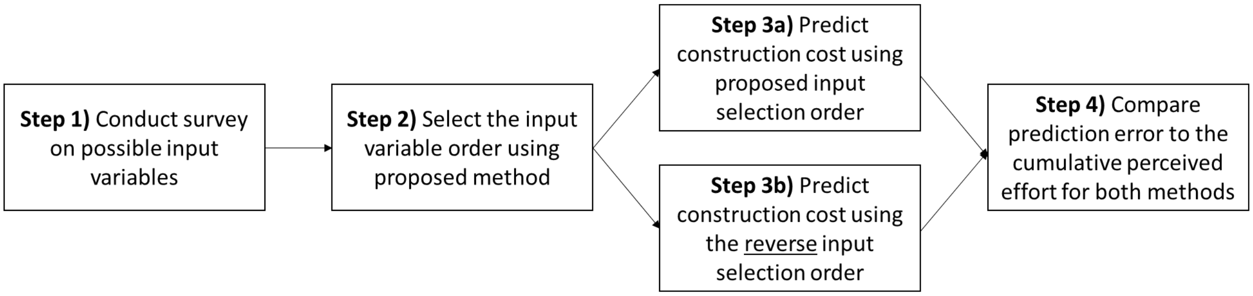

Step 1) Conduct survey on possible input variables
Step 2) Select the input variable order using proposed method
Step 3a) Predict construction cost using proposed input selection order
Step 3b) Predict construction cost using the reverse input selection order
Step 4) Compare prediction error to the cumulative perceived effort for both methods

Figure 12. Research steps

Survey

A survey was conducted at MDT to understand the perceived level of effort required to estimate the construction cost of a project at the conceptual stage. Firstly, two days of interviews at MDT established the key attributes of a project that influence the construction cost to aid the survey development, this was discussed in the base-model development (Chapter 3). Following these interviews, and a review of literature, 29 variables were identified that have an influence on the construction cost of MDT's highway projects, these are shown in Table 7. The research team then assigned the attributes into one of three categories:

1. *Roadway:* an attribute associated with information about the proposed project location.
2. *Design:* an attribute determined during the design process.
3. *Construction administration:* attribute is related to the construction activity.

These categories were selected to reflect the location where the data was being received from at MDT. For example the majority of roadway characteristics were generally sourced from the Data and Statistics Bureau at MDT which store Geographical Information Systems (GIS) on roadway attributes.

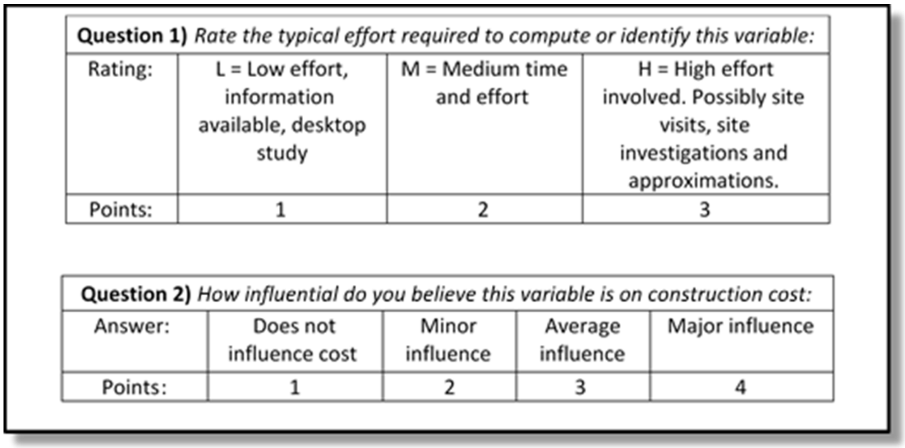
Table 7. Cost influencing attributes identified at MDT

Design related attribute		Roadway information attribute	
1	Design AADT	19	Urban or rural project
2	Design speed	20	Construction on Native American Reservations
3	Start and end stations, length and width	21	Site topography
4	Intersection signalization and signage	22	Existing surfacing conditions and depths
5	Horizontal and vertical alignment	23	Number of intersections in project
6	Extent of changes to the existing intersections	24	Number of bridges in the project scope
7	Typical section	Construction administration attribute	
8	Curb, gutter and sidewalk	25	Traffic Control - closures or detours
9	Bridge type and complexity	26	Environmental permitting requirements-wetlands
10	Volumes of excavation and embankment	27	Letting Date
11	Geotechnical - subsurface & slope recommendations	28	Context sensitive design issues, controversy
12	Bridge deck area	29	Contract time
13	Hydraulic recommendations and culverts		
14	Storm sewer extents		
15	Bridge span lengths		
16	Foundation complexity of the bridge		
17	Right-of-way acquisition and costs		
18	Extent of utility relocations and costs		

Survey respondents were asked, amongst other questions, to answer the following on each of the 29 attributes identified:

1. rate the typical effort required to compute or identify this variable, and
2. how influential do you believe this variable is on the construction cost of a project?

The entire survey template is shown in Appendix C. Questions were designed with an ordinal (categorical) scale where respondents are required to select the most suitable answer as shown in Figure 13 (Fink 2009; Fowler 2009).



Question 1) *Rate the typical effort required to compute or identify this variable:*			
Rating:	L = Low effort, information available, desktop study	M = Medium time and effort	H = High effort involved. Possibly site visits, site investigations and approximations.
Points:	1	2	3

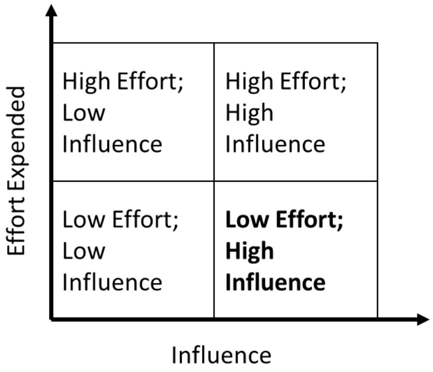
Question 2) *How influential do you believe this variable is on construction cost:*				
Answer:	Does not influence cost	Minor influence	Average influence	Major influence
Points:	1	2	3	4

Figure 13. Ordinal scale used for the two survey questions

The survey was distributed at MDT through an email link to all 84 preconstruction personnel that were deemed suitably qualified to respond. A total of 35 responses were received with four of these excluded as non-responses. This resulted in a 37% response rate. Responses were received from all five bureaus and from a large range of job titles. Whilst there is "no agreed-upon standard for a minimum acceptable response rate" (Fowler 2009) the researcher team were satisfied that the 37% response rate was reflective of the entire population.

Input variable selection

To meet the dual-objective goal during CCE it was proposed that input variables be selected starting with those that require a low level of effort to compute or identify but also have a high influence on the construction cost of the project. This is shown in Figure 14 below with the input variables suggested to be selected in the bottom right hand quadrant.



High Effort; Low Influence	High Effort; High Influence
Low Effort; Low Influence	**Low Effort; High Influence**

Figure 14. Selecting input variables to meet the dual-objectives of CCE

To validate this selection process the research team combined the perceptional survey results with performance of a data-driven CCE model created specifically using projects that the survey respondents design and manage at MDT. Two data-driven CCE modeling techniques, ANN and MRA, were utilized with the database developed in Chapter 3 to predict the construction costs of projects. Input variables were systematically added to the data-driven CCE model starting with those in the bottom right quadrant of Figure 14 to meet the dual-objectives of the main CCE goal. Further inputs were added based on their distance from the bottom right quadrant in Figure 14, this is explained in more detail later on in this paper. In each of the models the performance and total perceived effort from all input variables used were recorded.

Results

Survey response

The average results of the survey from 31 respondents are shown in Figure 15, the numbers relate to the 29 attributes from Table 7. Respondents rated the effort on a 1-3 ordinal scale whilst the influence of this variable on the construction cost was rated on a 1-4 ordinal scale, these scales are shown in Figure 13. As such quadrants were arbitrarily assigned on both scales to visually divide up the results and aid the input variable selection process. The units on both axis correspond to the ordinal response scale from Figure 13, they are referred to as "points" from here on.

Visually, there are a number of interesting results which can be observed in Figure 15. Firstly, only 5 of the 29 attributes shown in Table 7 fall in the bottom right quadrant of the plot: attributes MDT perceive as requiring a low amount of effort to collect which also have a high influence on the construction cost of the project. It was not a surprise that three are roadway characteristics, easily identified once a project has been selected and its location confirmed. These characteristics include whether the project is going to be in an urban environment, the topography of the road and the number of bridges within the limits of the project. Bridge deck area was the only design factor identified in the bottom right quadrant.

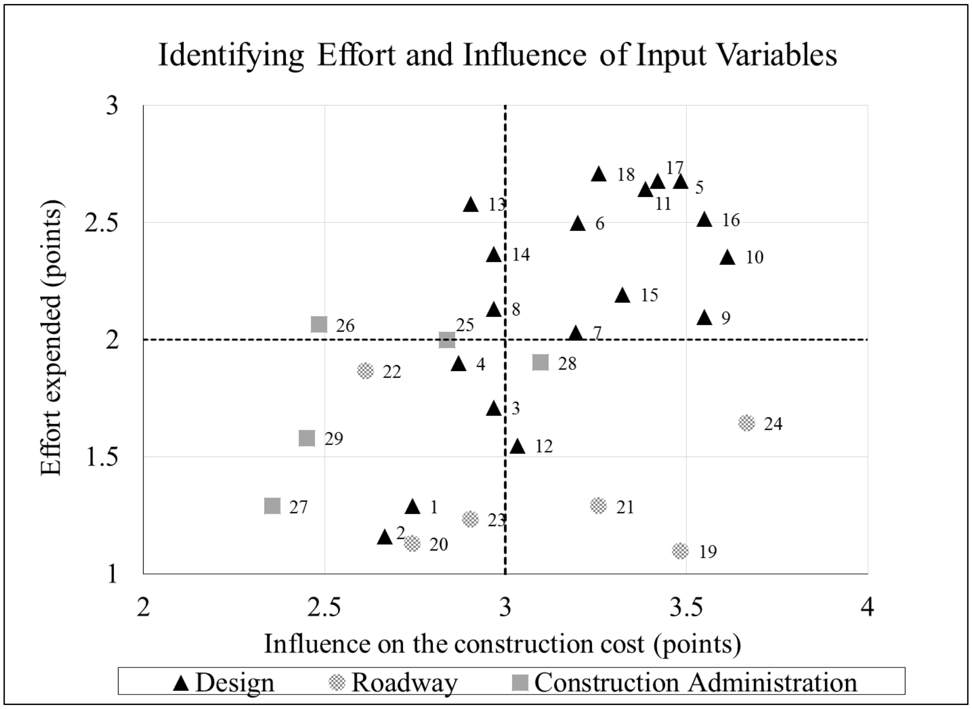


Figure 15. Results of MDT cost estimating survey

Secondly, all the attributes in the top right quadrant of the Figure 15 are design factors. This is intuitively logical as design requires significant effort to be expended and the outcome should have a large effect on the construction cost. Finally, very few variables occupy the top left quadrant. Those that do occupy this quadrant are bordering other quadrants inferring that any attribute requiring a significant amount of effort to be expended by MDT is going to have a significant influence on the construction cost of the project. This observation is also reinforced by the fact that two-thirds of all variables are in the bottom left or top right quadrant (i.e variables are either low-effort/low-influence or high-effort/high-influence variables).

Case-study

The findings from the survey were used to validate the dual-objective input variable selection method proposed as part of this research. The research team proceeded to build a data-driven CCE model, which has the least amount of effort with suitable performance. As such as many of the 29 attributes were included in the model, one at a time, starting with the variable closest to the most preferred to the least preferred variables (as shown in Figure 16). The formula to calculate each distance was based on the Euclidean distance, and shown in Equation 2 (Danielsson 1980).

$$Distance\ to\ ideal\ input\ (points) = \sqrt{(x_i - A)^2 + \ (y_i - B)^2} \qquad (2)$$

where,

x_i = the average perceived cost influence from the survey.

A = 4, the maximum construction cost influence based on the ordinal survey rating and the ideal value as shown in the survey questions (Figure 13).

y_i= the average perceived effort from the survey.

B = 1, the minimum effort rating based on the ordinal survey rating and the ideal value as shown in the survey questions (Figure 13).

i = the input attribute being measured, ranges from 1 to 29.

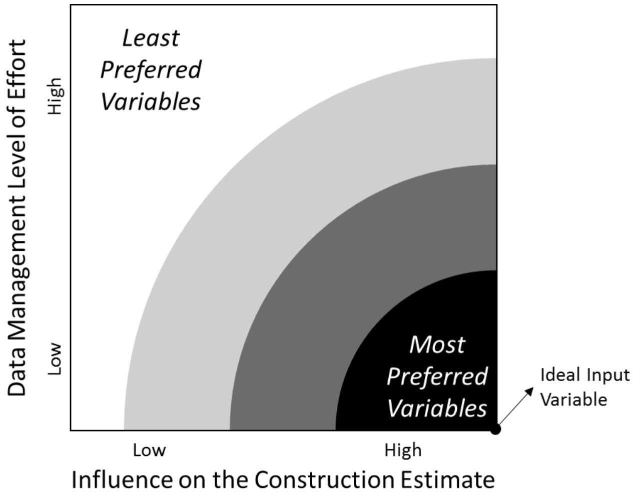

Least Preferred Variables
Most Preferred Variables
Ideal Input Variable
Data Management Level of Effort
High
Low
Low
High
Influence on the Construction Estimate

Figure 16. Preference for selecting input variables

The research team then used the base ANN with 17 input variables. In this chapter the database is tested with both ANN and MRA models. Because the perceptive survey for effort and cost influence was created for generic project types, some of the project attributes were not relevant to pavement preservation projects. As a result, 13 of the 17 input variables were chosen as the perceived effort would have been most relevant to pavement preservation were selected. These were selected based on guidance from MDT personnel and the ranked order is shown in Table 8 from the most preferred input variable to the least preferred.

Table 8. Input variables selection order and distance from ideal input

Proposed input variable selection order	Average perceived influence (points)	Average perceived effort (points)	Distance to ideal input (points) Refer to Equation 2
19. Urban or rural project	3.48	1.10	0.56
21. Site topography (steep, flat or undulating terrain)	3.26	1.29	0.80
3. Start and End Stations, Length and Width	2.97	1.71	1.25
1. Design AADT	2.74	1.29	1.29
7. Typical Section (depths of surfacing and aggregate)	3.19	2.03	1.31
2. Design speed(s)	2.67	1.16	1.34
4. Intersection signalization and signage	2.87	1.90	1.44
25. Traffic Control - closures or detours	2.84	2.00	1.53
8. Curb & Gutter and Sidewalk	2.97	2.13	1.53
29. Contract Time	2.45	1.58	1.65
27. Letting Date	2.35	1.29	1.67
11. Geotechnical - subsurface & slope recommendations	3.39	2.65	1.76
6. Extent of Utility relocations and costs	3.26	2.71	1.86

Input variables were added by selecting them in the order starting with the shortest distance from the ideal input variable to the largest distance. The average survey results for the influence and effort are shown in Table 8 along with the calculated distance to the 'ideal input variable' shown in Figure 16. Each time a new input variable was added to the model the MAPE of the model with the test data was recorded. To verify the usefulness of the input selection method the process was repeated in the reverse order (starting with the largest distance from the ideal input variable).

To be able to compare the results from all the models, the same 151 projects selected in Chapter 4 were used to train each model and the same 38 projects were used to test the model and calculate the MAPE. The 38 projects were selected through the sampling methodology developed in Chapter 4.

ANN Results

A commercially available ANN modelling software package was used to train and then test the database. Initially, only one input variable with the shortest distance to the 'ideal input variable' shown in Figure 16 was used to train and then test the first model. Input variables were then added to the model one at a time, getting further from the 'ideal input variable'. Each time the MAPE and cumulative effort points of the prediction model was recorded. The process was then repeated until all 13 input variables were included in the ANN model. The process was then conducted in reverse order by adding input variables in the opposite fashion. Figure 17 illustrates the results of each approach.

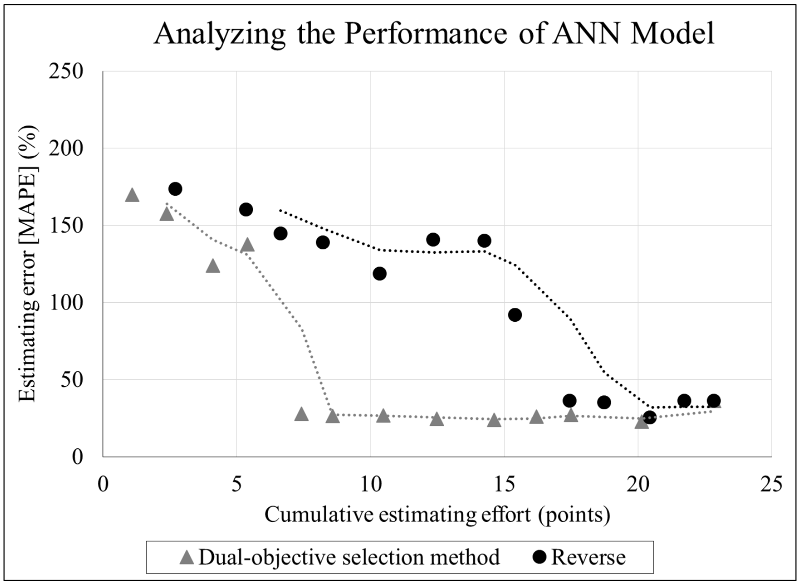


Figure 17. ANN performance and effort expended

Figure 17 shows that when input variables are added in the order suggested by this method then the model can more quickly reach reasonable accuracy with less effort. This method minimized the number of input variables required to achieve the lowest possible MAPE. Once the first 6-8 variables, from Table 8, were added to the model then adding further inputs yielded no further reduction in estimating error. The corresponding model reached around 25% MAPE estimating error with a cumulative effort of 7.5 points. With the reverse order of input variable selection a comparable level of error was not reached until around 17.5 to 20 points of effort. This is over twice the level of estimating effort for the same performance. Both methods show that there is a point where adding additional input variables, or expending more effort, results in diminishing returns and little or no improvement in performance in predicting construction costs for the additional effort. When the point of diminishing returns is reached the overall goal of the estimating model is reached: maximum performance with minimal effort. This also effectively debunks the notion that increasing the number of input variables will increase the accuracy of the estimate.

The authors speculate that selecting input variables which require a low level of effort essentially means that variable is known to a high degree of confidence at the early estimate stage. Two examples are the 'length' of the project and if the project will be in an 'urban or rural' setting. These two variables both require a low level of effort, thus are known to a high degree of confidence at the early stage. Because these two variables were also perceived by MDT as having a high influence on the construction cost then the input selection process proposed in this research picked these two variables amongst the first 6-8 variables.

On the contrary, design variables require a high level of effort at the early stage. Although they have high influence on the construction cost many were excluded from the first 6-8 variables. Most design factors do have a perceived high impact on the construction cost, but, at the early stage there is a low level of confidence with those numbers. Two such examples are the geotechnical complexities and utility replacements required. At the early stage highway agencies only have a very vague estimate of those variables, thus the confidence in the top-down number is very low at the conceptual stage. However, it is recognized that their designed outcome does have a significant impact on the cost. The data inputs for design variables in the conceptual estimating model are sourced from project information at the early stage, thus they are not inputs known to a high level of confidence and contain plenty of variability from this initial estimate to the final estimate. This is unlike variables such as the 'length' or 'urban/rural' input variables which are known to a high level of confidence at the early stage and also have a high impact on the construction cost.

MRA Results

The same database was next used with commercial software for MRA with a linear assumption. When the process was repeated with MRA the rational selection method proposed in this research also proved successful to meet both objectives, as seen in Figure 18. It is evident that the ANN model’s performance was superior to the MRA, 25% error using ANN compared to 50% with MRA. These errors are both within the range suggested by the AASHTO *Practical Guide to Cost Estimating* (2013) at the planning stage. The superior performance of ANN is in agreement with several data-driven CCE models found in the literature (Petroutsatou et al. 2012; Kim et al. 2004; Moselhi and Siqueira 1998). However, this conclusion is not universal in the construction literature with some authors reporting the opposite findings (Gunduz et al. 2011; Setyawati et al. 2002). The ongoing debate with both techniques was the reason that this chapter differed from Chapters 5 and 7 by testing the database with both MRA and ANN in order to contribute to other literature findings.

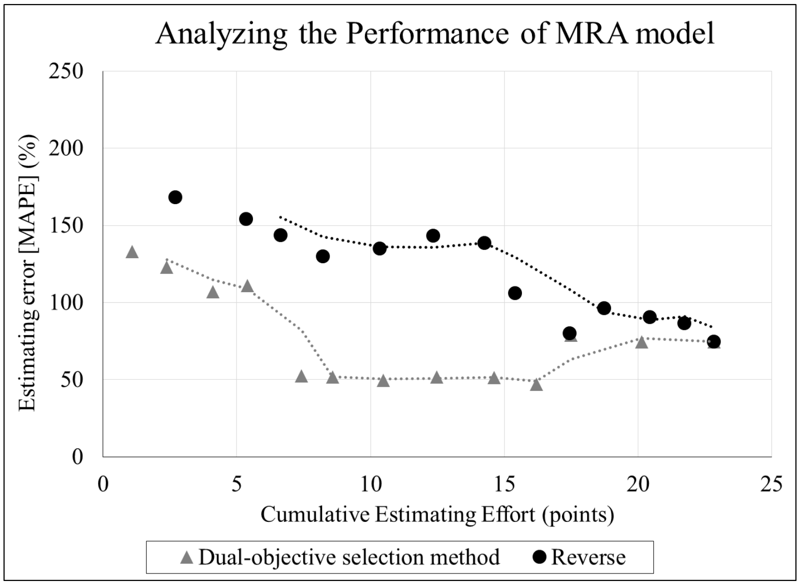


Figure 18. MRA performance and effort expended

It is interesting to note that with the MRA model when using the reverse order of input variables never reaches the optimal prediction accuracy of around 50%. Also the regression analysis actually performs better with fewer input variables and at 10 input variables, the MAPE starts increasing. Without a rational input variable selection method, such as trial and error commonly employed in the literature (Hegazy and Ayed 1998; Kim et al. 2004), one may conclude that a given set of data is not capable of predicting the construction costs to reasonable accuracy.

Discussion

The research in this paper has shown that data-driven CCE models do not need to include all project attributes to predict the construction cost to reasonable accuracy at an early stage of project development. If highway agencies are going to employ data-driven methods for CCE then the implications of this research highlight:

1. A rational input selection method, such as the one suggested in this paper, can be used to yield suitable input variables with low effort and contribute to acceptable performance.
2. Once highway agencies are confident in the input variables required to estimate the conceptual cost of projects, the collection of further information is unnecessary. It only consumes data storage space and requires time/effort from personnel whose efforts could be better applied elsewhere.
3. The results imply that suitable confidence in estimating the conceptual costs of projects can be achieved with lower project definition if the correct input variables are selected.

The final implication of this study is the most important: at the conceptual stage of a project life-cycle, an early estimate with readily available input variables can achieve satisfactory accuracy. This is better than a slightly more accurate result at a later stage of design development. It should be noted that this research is based on the analysis of perceptional data from a single DOT agency and as such, its conclusions cannot be generalized without regard to a specific agency's attribute impact and effort perceptions being checked. Nevertheless, the overarching concept of using the high impact/low effort variables should be true for most, if not, all public transportation projects.

Conclusion

ANN and MRA models constructed for this research both reached the goal with the dual-objectives of low effort and high accuracy faster using the input selection method proposed in this research. Adding further input variables using either model technique resulted in diminishing returns of the model performance. Findings from this research have positive implications for practitioners willing to employ data-driven conceptual cost estimating techniques.

The paper’s primary contribution for both researchers and practitioners is to highlight for the first time that while increasing the number of input variables in an early estimate may appear to enhance estimate accuracy on an intuitive basis, this is not necessarily true. The MDT case study showed that for both the ANN and MRA approaches that adding detail to the model reached a point of diminishing returns at roughly 6 to 8 high impact/low effort variables.

CHAPTER 6. STOCHASTIC COST ESTIMATING OF HIGHWAY PROJECTS AT THE CONCEPTUAL STAGE USING BOOTSTRAP SAMPLING

Gardner, B., Rueda, J., Gransberg, D. D. (2015). "Stochastic Cost Estimating of Highway Projects at the Conceptual Stage using Bootstrap Sampling." To be submitted to the ASCE-ASME *Journal of Uncertainty and Risk.*

Abstract

Conceptual cost estimating is typically completed early in the project life-cycle when very little design work has been completed. Because little information is known at this early stage, conceptual estimates usually deviate substantially from actual construction costs. The conceptual estimate is not expected to be highly accurate; however when expressed as a deterministic value, it often leads to a false inference of accuracy by those not familiar with the vagaries of conceptual cost estimating, making it difficult for the agency to explain cost growth as the project proceeds through the development process. Communicating the conceptual estimate stochastically allows the agency to produce a probability distribution of the likely construction cost and address the level of confidence it has in the given estimate. Named probability distributions are readily available for developing a stochastic estimate on many commercial software's to communicate uncertainty. However, instead of fitting available distributions, this research generates an empirical distribution to express a range in construction costs for individual projects. Creating empirical distributions eliminates assumptions required for selecting an existing distribution. This paper describes the development of a stochastic data-driven model, which combines artificial neural networks and bootstrap sampling to estimate construction costs and their associated uncertainty at the conceptual stage. This study used 189 highway projects to train and test the estimating model.

Introduction

The difficulty with conceptual cost estimate accuracy is demonstrated in the AASHTO *Practical Guide to Cost Estimating* (2013), which cites the accepted uncertainty of the early estimate in a range of -40% to +100% from the initial cost estimate to the final construction cost. This corresponds to a project scope definition of 1-15%, as shown in Table 1 (Chapter 1). That AASHTO publication also acknowledges the difficulty in quantifying uncertainty associated the cost at the conceptual stage. It is known that many highway agencies experience substantial cost growth from this initial estimate to the final construction cost (Flyvbjerg et al. 2002; Schexnayder et al. 2003; Chou et al. 2006).

Reflecting the construction cost as a point estimate (i.e. a given number) does not portray the estimator's confidence, or lack thereof, in the estimate, nor does it indicate the potential for cost growth. Therefore, those using the estimate in the planning and programming process may be over confident in its accuracy. The following section discusses the bias and optimism associated with point estimates, it then goes on to discuss the benefits of reflecting the construction cost stochastically.

Optimism and bias associated with conceptual estimates

Bias from the estimator and the tendency to be over-optimistic in construction costs has been found to directly attribute to construction cost growth. Bias and over-optimism was discovered as one of the 18 primary factors contributing to construction cost escalation by Shane et al. (2009). Over-optimism was "often viewed as the purposeful underestimation of project costs to ensure that a project remains in the construction program" (Shane et al. 2009). In that study interviews were conducted with over 20 public highway agencies to identify the key factors which led to highway construction cost escalation.

There is a proven link in the literature for which an optimistic estimate of construction cost can lead to inadequate design funds for a project and further exacerbate construction cost growth. Typically, the design budget is established as a percentage of the initial construction cost estimate (Jeong and Woldesenbet 2012). Therefore if the construction budget is optimistic (low), so too is the design budget. Gransberg et al. (2007) investigated the relationship between the design budget and cost growth from the initial estimate. The study established that, up to a point, the greater the percentage assigned to design, the lower the cost growth measured with respect to the conceptual estimate. It therefore follows that an optimistic design budget, assigned as the result of an optimistic construction cost estimate, will more likely lead to cost growth from the initial estimate due to design activities being underfunded.

In the study by Flyvbjerg et al. (2002) it was found with overwhelming statistical significance that cost estimates presented at the pre-design stage are systematically and intentionally misleading, and not caused by error. The study by Flyvbjerg et al., discussed in Chapter 1, included 258 transportation infrastructure projects from different historical periods, geographical regions and project types. Three main reasons for the statistical significance were investigated; this was: economic self-interest, appraisal-optimism, or misleading forecasts for political reasons to get projects started. The conclusion of that research was that the pre-design cost estimates were deliberately low to get projects started and hence the reason for 9 out of 10 projects experiencing cost growth.

This paper proposes the use of data-driven methods to produce stochastic estimates and increase the level of cost transparency. Using historical project data to forecast costs and assign contingencies removes any psychological elements or bias that may be inherent to the estimator. Additionally, if the output is reported correctly, it should reduce any deliberate deception from project promoters whom omit project risks and other potential costly elements in a traditional point estimate (deterministic estimate) in order to get the project started.

Stochastic range estimating – the objective

Most highway agencies currently express their conceptual estimate as a point estimate with a contingency assigned as a percentage of the construction cost (Molenaar 2005, Byrnes 2002, Turochy et al. 2001). Byrnes reported that DOTs add a contingency ranging from 5-45% depending on project type and uncertainty; similar contingency factors were also reported by Turochy et al. (2001). The problem with point estimates is that they communicate a false sense of confidence in the cost estimate, making it difficult to assess their quality (AASHTO 2013) and potentially leading to forecast bias by those using the estimate to make financial decisions (Chelst and Canbolt 2012). Firstly, when the conceptual estimate is expressed as a point estimate, it appears accurate to those with no knowledge of the limitations of the estimate itself. Hence, there is a perceived illusion of control and predictability. Secondly those using the point estimate in a benefit-to-cost analysis or for budgeting, fail to acknowledge the possible extreme values or range in numbers that the final construction cost could eventually experience. Finally, Chelst and Canbolt (2012) state that there can be tendency for an anchoring bias, where "the forecaster becomes too anchored to the first estimate to develop a wide range that is reflective of actual dispersion" of the costs. Chelst and Canbolt go on to state that "the preferred technique is to initially focus on estimating both good and bad extremes."

Providing an estimate range is often thought to show less confidence in the cost and forethought than a point estimate. However, a probabilistic range actually requires the estimator to draw on a wide spectrum of experiences to define a range as well as to explore its associated probabilities (Chelst and Canbolt 2012). Point estimates on the other hand simply require specific assumptions and corresponding numbers to justify that forecast (Chelst and Canbolt 2012).

This research investigates a stochastic range estimating method to improve communication of the conceptual cost estimate to those that are unfamiliar with its basis and limitations. The paper's objective is to explore a method which permits highway agencies to utilize databases of historic project information for the following purposes:

1. To forecast the final cost at the conceptual stage and,
2. To assign a range of expected costs to help communicate the uncertainty associated with the conceptual estimate and,
3. To compare cost estimating transparency of the point estimate to that of the stochastic approach.

This chapter utilizes the same database developed from MDT projects and introduced in Chapter 3. The method is tested with ANN modeling, however the principles could be extended to MRA models or projects of different scope.

Background

Holistic risk approach

There are two problems with the current technique of assigning contingency as a percentage of the construction cost estimate. Firstly, the contingency required is not necessarily directly proportional to the construction cost; contingency should depend on other factors such as project type and complexity (Gransberg et al. 2011). Secondly, if the construction cost estimate is low, then the assigned contingency will also be low, further exacerbating the cost growth of the project. On the other hand if the construction cost estimate is high, then the contingency will be too high, unnecessarily tying up additional fiscal year funding which might have been used to fund additional projects.

An alternative approach to assigning contingency as a percent of the construction cost estimate is to use a 'bottom-up' method by creating a project specific risk register. All possible risks, likelihoods, and consequences are assigned a possible value and contribute to the overall contingency fund of the project. The problem with a risk-register is that at the early stages very little information is known about the project, making it difficult to conduct an elemental 'bottom-up' estimate of all the risks. Additionally when one conducts a 'bottom-up' assessment one must still make an allowance for risks that have yet to be identified (Kaplan and Garrick 1981). Since the conceptual estimate and its associated risk assessment, are produced at an early stage of project development, the allowance for unknown risks would be difficult to quantify. This 'bottom-up' approach should be reserved for later, more confident, estimates when more information is known about a particular project, and its risks can be better itemized.

An emerging technique, investigated in this research project, is to take a more holistic ('top-down') approach to assign the contingency (Sillars and O'Connor 2007). Sillars and O'Connor created such a cost-risk procedure for the Federal Transit Administration (FTA). This was in response to the 'bottom-up' risk register method not performing well and lacking the required variability of ranges. At the conceptual stage a 'top-down' holistic approach intuitively makes sense due to the difficulty with identifying all possible risks until the design is complete. The current state-of-the-practice, assigning contingency based on construction cost, is still a holistic approach, but it is directly proportional to the confidence in the conceptual cost estimate.

This research aims to leverage the 'top-down' cost estimating approach developed in Chapters 3 and 4 to not only calculate the construction cost, but also an associated contingency based on the risk profile of the decisions makers. Data-driven estimating models found in the literature generally express the result as a point estimate (Sonmez 2008). This research investigates the use of combining ANNs with bootstrap statistical sampling to create a stochastic range of the construction costs for highway projects.

Bootstrap sampling method

The bootstrap method provides a simple process to resample the original data-set (Chernick 1999). Utilizing the bootstrap method to sample a database enables one to answer a key question in data-analysis and statistics: how accurate are the results of the estimate? (Efron and Tibshirani 1993; Davison and Hinkley 1997). Efron and Tibshirani (1993), summarized many of the bootstrap applications discovered since the 1980s including the ability to create empirical distributions, calculating standard errors, integration with regression analysis and confidence intervals.

The bootstrap data-set is created by sampling the original data-set, shown in Figure 19. There a two methods to sample the original data-set shown in Figure 19 (process A) (Efron and Tibshirani 1993; Davison and Hinkley 1997):

1. sampling without replacement (WOR) or,
2. sampling with replacement (WR).

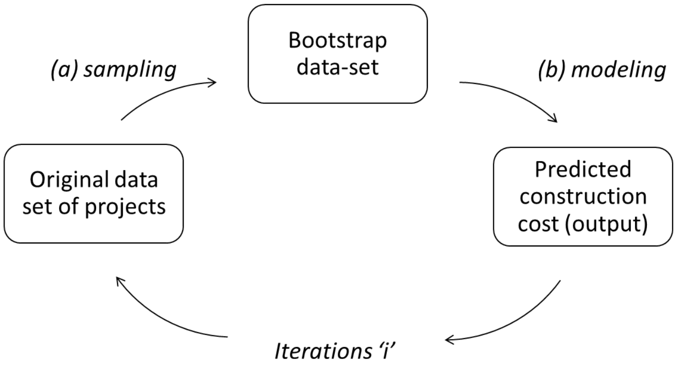

Bootstrap data-set
(a) sampling
(b) modeling
Original data set of projects
Predicted construction cost (output)
Iterations 'i'

Figure 19. Bootstrap process (developed from Efron and Tibshirani (1993)

Extracting a nominated percentage of projects from the original data-set is sampling without replacement (WOR). In this process 'n' is defined as the size of the bootstrap sample and 'N' is the number of data points in the original data-set. The bootstrap data-set cannot exceed the size of the original data-set (N>n). Additionally, every project in the original data-set can only occur once in the bootstrap data-set. The sample fraction is simply defined by f=n/N (Efron and Tibshirani 1993; Davison and Hinkley 1997).

The second method to sample the projects is with replacement (WR). Once a project has been included in the bootstrap data-set then it is returned to the original data-set of projects to enable it to be selected again (Sonmez 2011; Efron and Tibshirani 1993; Davison and Hinkley 1997). Sampling WR means that some data in the bootstrap set can appear zero times, some appear once, some appear twice or more (Sonmez 2008).

Davison and Hinkley (1997) argue that sampling WOR is the simplest method, Efron and Tibshirani (1993) argue the opposite. Provided that the bootstrap sample is much smaller than the population size then the probability of sample repetitions will be small anyway (Efron and Tibshirani 1993). This research tests sampling WOR method, this is because the bootstrap method is being used to create confidence intervals and not as a method to deal with lack of data used in other studies (Tsai and Li 2008).

Once the bootstrap sample of projects is created, the construction cost (output) can be calculated by modeling (process B). Two methods presented above were ANN or MRA to predict the construction cost. Because ANN and MRA are data-driven estimating techniques then the output will vary with the input of projects. Therefore, a range estimate can be created if there is methodical control of the data-set (inputs) going into the data-model to get accordingly varied construction cost (outputs).

The final step is to iterate, as shown in Figure 19. Iterating the bootstrap sampling process many times allows one to obtain multiple construction cost outputs with different costs. A probability distribution function of the construction costs (outputs) can be created either in a discrete method (probability mass function) or by converting the discrete outcomes to a continuous function (probability density function). The probability distribution function is commonly called a stochastic estimate because the expected construction costs have probabilities associated with them (Bedford and Cooke 2001).

Tsai and Li (2008) used the bootstrap method combined with an ANN to estimate the cost of manufacturing ceramic powder. Their study specifically pursued this technique to address the small training data-set that they had by creating virtual samples. Tsai and Li's study found that using the bootstrap method to create virtual samples actually reduced the ANN error and made the predictions more stable. They argued a benefit of bootstrap sampling combined with ANN modeling was the improvement in accuracy when little data was available through the use of virtual samples. Instead of stabilizing a small data-set, this chapter makes use of the bootstrap approach to create a stochastic cost estimate, the details of which are covered in the next section.

Stochastic estimating – previous studies

Kaplan and Garrick (1981) recognized the benefits of a probabilistic curve when quantifying risk by stating that "a single number is not a big enough concept to communicate the idea of risk. It takes a whole [risk] curve." The benefit of stochastic estimating has been explored by various authors since then, but few in the field of highway construction cost estimating. FHWA, in their cost estimating guidance (2007), allow highway agencies to express their conceptual estimates as a range with indicated levels of confidence, thus it is logical to draw increased attention of the ability of highway agencies to communicate their conceptual estimates through a range.

In 2005 Molenaar created a stochastic cost estimating method for Washington State Department of Transportation (WSDOT) specifically for projects greater than $100M in cost. WSDOT are now successfully implementing this practice. Molenaar concluded that the "stochastic method better conveyed the uncertain nature of project costs at the conceptual phase of project development." The stochastic method was trialed on 'Highway Megaprojects' and although the method was effective, the cost of the process was in the order of $3M for WSDOT due to workshops, development costs and feedback sessions. Molenaar's research concluded that the benefit was better management of public funds and possible gains in public confidence through transparent communication. That research solely concentrated on megaprojects and if highway agencies are to adopt this method then they need to employ a risk-analyst expert. The research reported below instead focuses on typical projects for highway agencies and should not require the employment of a specialist to manage.

Sonmez (2008) used bootstrap sampling with replacement to calculate a probabilistic conceptual cost estimate of a building. The number of projects used to train the regression model was 19. The technique was deemed valid when the one building project, with which the model was validated with, was enclosed within the 90% probability level. A total of 1000 iterations were completed where the construction cost of the test project was calculated in each iteration with a bootstrap data-set of 20 projects. Each of the 19 projects available to make the bootstrap sample was included either nil, once, twice or many times to fill the 20 training spots, thus sampling WR was used. Sonmez stated that further studies should include larger data-sets, this chapter contributes to the limitation outlined by Sonmez through the use of 189 projects in the database as opposed to 20.

In other fields, researchers used the bootstrap procedure to represent uncertainty for incremental cost-effectiveness ratios for endoscopy clinical procedure (Lord and Asante 1999). The authors stated that health economists have a "responsibility to present estimates of the degree of uncertainty surrounding the results of economic evaluations." They indicated that decision-makers place too much reliance on point estimate results presented. This communication issue and perceived confidence is therefore not only experienced in the construction industry.

Other techniques to produce a stochastic estimate, without the use of bootstrap sampling, do exist. Monte-Carlo simulation can be used simulate outcomes to produce probability in a commercial spreadsheet. In 2004 Sonmez used this approach to create a range estimate using normal distribution. However, in that research Sonmez did outline the inherent assumptions regarding the distributions and expected errors. This conclusion further supports the use of bootstrap to create an empirical distribution as it “enjoys the advantage of not relying on assumptions or calculations of the original distributions” (Dupret and Koda 2000).

Methodology

To compare the cost estimating effectiveness of a stochastic estimate with a point estimate then both methods of estimating the construction costs were completed. The methodology differences for the two different models are shown in Table 9. The ANN model for the point estimate was that developed in Chapter 4. No adjustment to the model architecture, input attributes or modeling software were made between the models; the only exception being the projects that were used to train the ANN.

Table 9. Model details for point estimate and stochastic estimate

	Point Estimate	Stochastic Estimate
Number of projects in testing database	38	38
Number of projects in training database	151	121
Number of iterations	1	100
Output	Point estimate	Confidence interval
Validation	MAPE	Actual CN within confidence interval

The three main steps taken to create the point estimate and stochastic estimate output are detailed:

1. ANN data-driven model from Chapter 4 used predict construction cost as a point estimate for 38 test projects. All 151 projects were used to train the ANN as shown in Table 9.
2. Bootstrap samples of 121 projects were used to train the ANN model instead of the entire data-set. A total of 100 iterations were completed (i.e. 100 point estimates) with randomly selected bootstrap samples. The construction cost of the same 38 projects was predicted on each iteration. The combination of all construction costs formed the stochastic estimate and this was converted into a confidence interval.

3. The output of the point estimate and stochastic estimates were compared.

A point estimate provides a single number and a stochastic estimate provides a range of numbers. The difference in the form of the estimate output makes comparison difficult. As such our research team validated the two models differently, this is shown in Table 9. The ability to communicate the cost estimate confidence was compared between the point estimate and that of the range estimate. For both estimating methods the performance of the estimating tool was measured against the actual construction cost, this is the validation technique in Table 9. Comparison with the actual CN cost to the point estimate was calculated using the MAPE (Equation 1, Chapter 2). The performance of the range estimate could not be measured using the MAPE as the output was a range of numbers. Instead, for validation of the stochastic estimate the actual construction cost was compared to the range estimate to see if it was enclosed within the maximum and minimum extreme values.

Data Analysis and Results

The results section is divided into two parts. The ANN model outlining the results from developing a data-driven point estimate (Results I). In the second part the point estimate is further developed into a stochastic estimating model (Results II).

Results I: point estimating model

The point estimate was calculated using all 151 projects to train the model and the same input parameters as presented in Chapter 4. The same 38 projects were used to test the model and calculate the MAPE, the error from each of the individual 38 projects is shown in Table 10. The MAPE of all test projects was calculated through Equation 1, this was 23% and shown in Table 10, well within the recommended performance in the AASHTO *Practical Guide to Cost Estimating* (2013) at the conceptual stage.

It could be perceived by a project promoter that given a point estimate, the construction cost should be enclosed by a range within 23% of that number. But this is not correct. The MAPE was calculated based on the *average* error from the actual construction cost. If one enclosed a range +/-23% from the actual construction costs only 24 out of the 38 estimates would fall within this range, as shown in Table 10. Thus this finding shows that the MAPE does not reflect the confidence of each individual project, our model much more confidently predicts the construction costs of some projects when compared to others. The empirical method produced in the following section creates individual contingencies for each project based on the confidence in that project and associated data.

Table 10. Point estimate versus actual construction cost

Unique project number	Predicted point estimate	Actual construction cost	Estimating Error	Enclosed within +/- 23% bounds of the predicted
7907	$ 2,190,506	$ 2,049,786	7%	Yes
7655	$ 687,360	$ 618,878	11%	Yes
7648	$ 1,610,835	$ 1,577,284	2%	Yes
7629	$ 935,281	$ 1,416,928	34%	No
7622	$ 2,931,223	$ 2,735,769	7%	Yes
7616	$ 2,714,477	$ 2,341,870	16%	Yes
7613	$ 274,872	$ 346,417	21%	Yes
7611	$ 815,565	$ 1,228,248	34%	No
7610	$ 788,482	$ 668,753	18%	Yes
7608	$ 478,445	$ 655,898	27%	No
7601	$ 2,494,663	$ 2,153,096	16%	Yes
7471	$ 419,294	$ 845,535	50%	No
7462	$ 577,875	$ 706,344	18%	Yes
7444	$ 1,956,166	$ 1,904,516	3%	Yes
7405	$ 136,058	$ 121,409	12%	Yes
7306	$ 191,456	$ 413,068	54%	No
7108	$ 469,082	$ 1,173,722	60%	No
6988	$ 121,798	$ 85,237	43%	No
6974	$ 2,732,350	$ 3,380,123	19%	Yes
6959	$ 535,376	$ 508,032	5%	Yes
6952	$ 1,567,018	$ 1,963,090	20%	Yes
6948	$ 324,069	$ 337,096	4%	Yes
6944	$ 865,742	$ 960,662	10%	Yes
6942	$ 655,190	$ 541,157	21%	Yes
6927	$ 1,431,002	$ 1,300,320	10%	Yes
6894	$ 2,080,816	$ 1,469,483	42%	No
6811	$ 336,661	$ 296,926	13%	Yes
6799	$ 211,790	$ 182,946	16%	Yes
6795	$ 354,359	$ 351,910	1%	Yes
6523	$ 463,207	$ 578,304	20%	Yes
6503	$ 218,961	$ 255,169	14%	Yes
6501	$ 1,340,614	$ 1,044,308	28%	No
6499	$ 597,541	$ 772,972	23%	No
6266	$ 570,293	$ 327,928	74%	No
6253	$ 440,025	$ 656,403	33%	No
6237	$ 344,405	$ 285,501	21%	Yes
5752	$ 2,218,890	$ 1,701,527	30%	No
5751	$ 1,717,133	$ 2,663,697	36%	No
MAPE (calculated through Equation 1, Chapter 3):			22.9%	

Results II: stochastic estimating model

Table 11. Range estimate results for 38 test projects

Project Number	Minimum Value Predicted	Probability Level				Maximum Value Predicted	Actual Construction Cost
		5%	15%	85%	95%		
7907	$824,741	$1,406,550	$1,728,648	$2,825,781	$2,870,946	$3,581,856	$2,049,786
7655	$430,625	$467,737	$572,985	$694,898	$696,662	$717,304	$618,878
7648	$542,000	$999,585	$1,199,560	$2,094,691	$2,412,435	$3,556,034	$1,577,284
7629	$895,547	$922,928	$923,321	$1,126,959	$1,221,602	$1,529,054	$1,416,928
7622	$1,133,263	$1,546,317	$1,568,898	$3,031,642	$3,032,167	$3,032,169	$2,735,769
7616	$1,153,138	$1,174,832	$1,628,757	$2,714,176	$2,715,070	$2,737,307	$2,341,870
7613	$161,313	$194,891	$229,911	$301,865	$329,689	$384,002	$346,417
7611	$474,971	$483,203	$529,673	$1,032,264	$1,246,094	$1,456,776	$1,228,248
7610	$235,422	$488,716	$584,155	$753,068	$801,898	$1,248,959	$668,753
7608	$330,430	$355,491	$420,568	$518,382	$543,090	$630,549	$655,898
7601	$1,440,817	$1,440,837	$2,492,953	$3,431,572	$3,431,577	$4,038,078	$2,153,096
7471	$316,712	$355,137	$366,945	$558,984	$1,002,218	$2,511,961	$845,535
7462	$344,431	$480,753	$548,546	$668,353	$759,549	$1,204,432	$706,344
7444	$1,173,390	$1,232,745	$1,580,326	$2,735,104	$3,536,238	$4,051,083	$1,904,516
7405	$89,920	$104,680	$121,730	$164,815	$185,335	$310,316	$121,409
7306	$144,090	$160,617	$167,621	$234,521	$281,810	$2,283,585	$413,068
7108	$145,940	$372,513	$472,067	$627,791	$666,937	$2,271,069	$1,173,722
6988	$97,859	$104,047	$111,191	$148,464	$162,408	$402,573	$85,237
6974	$1,550,002	$1,773,396	$1,844,132	$3,065,984	$3,621,122	$3,891,009	$3,380,123
6959	$233,175	$308,543	$405,207	$545,208	$554,351	$570,260	$508,032
6952	$527,431	$603,247	$1,001,401	$2,048,786	$2,319,392	$2,657,287	$1,963,090
6948	$248,460	$270,439	$288,816	$391,078	$444,432	$1,077,672	$337,096
6944	$299,942	$466,895	$524,385	$1,254,101	$1,323,058	$2,891,232	$960,662
6942	$263,154	$377,331	$502,391	$692,654	$736,706	$766,056	$541,157
6927	$826,662	$913,651	$1,128,157	$1,529,055	$1,529,055	$3,150,506	$1,300,320
6894	$680,576	$749,276	$1,197,166	$2,304,526	$2,959,622	$3,327,156	$1,469,483
6811	$299,087	$313,086	$338,888	$605,708	$674,011	$1,238,211	$296,926
6799	$154,221	$158,102	$169,601	$214,793	$229,768	$296,274	$182,946
6795	$241,073	$287,675	$359,852	$545,451	$596,202	$857,057	$351,910
6523	$256,790	$362,354	$410,310	$522,215	$551,703	$605,589	$578,304
6503	$147,859	$169,334	$186,775	$245,650	$528,303	$1,006,085	$255,169
6501	$558,065	$896,055	$906,936	$1,342,951	$1,476,607	$1,529,052	$1,044,308
6499	$387,612	$439,490	$453,757	$615,456	$650,599	$1,382,243	$772,972
6266	$200,185	$315,382	$400,045	$661,697	$665,759	$1,173,788	$327,928
6253	$143,152	$199,808	$291,538	$556,654	$628,034	$1,359,631	$656,403
6237	$183,489	$198,812	$268,941	$385,624	$439,155	$558,939	$285,501
5752	$1,000,091	$1,255,209	$1,543,764	$4,249,406	$5,036,280	$5,275,446	$1,701,527
5751	$971,781	$1,261,650	$1,541,001	$2,203,069	$2,502,674	$4,257,199	$2,663,697

Range estimate results for all 38 test projects are shown in Table 11. The minimum and maximum values were the two extremes predicted during the 100 iterations in bootstrap samples.

The empirical probability levels 5%, 15%, 85% and 95% indicate probabilities that project costs will be below that value. The 90% confidence range of estimated construction cost is the range between the 95% and 5% probability levels, similar confidence levels can be obtained also by subtracting the high and low probability levels to calculate the confidence range. There are some interesting outcomes shown in Table 11:

- 35 of the 38 test projects fall within the minimum and maximum expected extremes predicted throughout the 100 bootstrap samples.
- 27 of the 38 test projects fall within the 5% and 95% expected cost.
- 18 of the 38 test projects fall within the 15% and 85% expected cost.

From these results it is apparent that as the confidence range is narrowed then more projects fall outside of the range. Thus to best represent the uncertainty then one should quote both the maximum and the minimum values.

Figure 20 displays the stochastic estimate for four selected projects. Project 6799 is a chip-seal project only and is known to a very high degree of certainty. This is shown in Figure 20 by the narrow range of expected construction costs. Projects 6952 and 7907 were mill and fill projects with length 6.2 and 7.5 miles respectively and the final surface was chip-seal surface. Due to the similar characteristics they are parallel with project 7907 slightly higher in predicted and actual costs due to the slightly longer length.

Project 5752 displays the least certainty and this is displayed visually with the widest range in expected construction cost. The stochastic ANN model has predicted a drastically different range for this project compared to both projects 6952 and 7907, this is despite reasonably similar actual construction costs for all three of three projects (5752, 6952 and 7907). Project 5752 was 8 miles in length, included asphaltic levelling, asphaltic isolation lift, asphaltic resurfacing lift followed by a chip-seal surface. The complexities and unknowns were all high with the other major difference being inclusion of bridge work. The modelling process has recognized the many high complexities and unknowns when calculating the cost of project 5752 and therefore produced a huge range in construction costs.

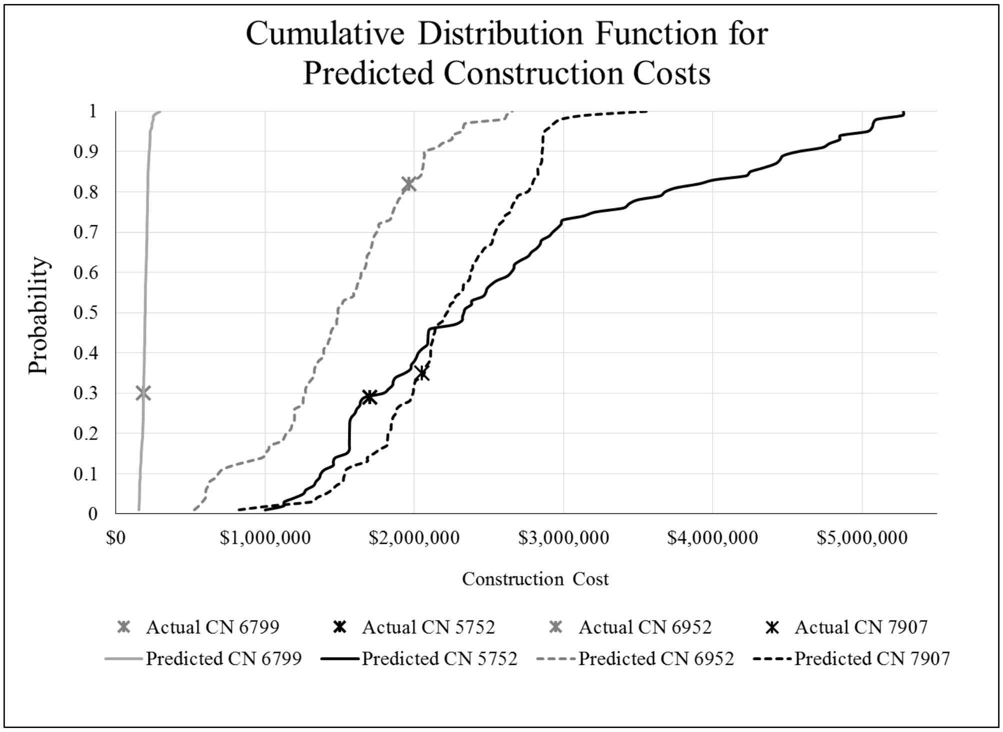


Figure 20. Visual representation of estimate confidence for four projects

The actual construction costs for each project, shown in Figure 20, fall within the confidence intervals for their respective ranges predicted with the model. The four plots in Figure 20 lead one to conclude that the distribution of expected construction costs are not constant. If one were to assign a distribution, then the assumptions of that named distribution would not work on all projects, this further highlights the benefits of the empirical process presented in this paper.

Discussion

A limitation of ANN results is that it is essentially a 'blackbox' where one cannot easily decipher the reason for cost variation. The literature confirms that this is a common downside to ANNs (Kim et al. 2004; Hegazy and Ayed 1998). The project costs are estimated based on pattern recognition, and perhaps the pattern recognition, or lack thereof, is providing the confidence intervals. When more data is added to the ANN then one may become more confident in the range of possible project costs.

In developing a stochastic and point estimating model with the same set of data it has become apparent that:

- The point estimate results provide no rational means to assign an individual contingency for each project based on the result. Thus the point estimate provides no improvement to the current state-of-the-practice for assigning contingency.

- Producing a stochastic estimate visually aided the comparison of expected construction costs for various projects.
- Given the large variations in the empirical distributions then it is apparent that a single set distribution could not easily be added to each project to assess the confidence levels.

This research presented here is an example of how a highway agency could embrace this estimating principle for cost transparency, utilization of existing databases and to express the actual confidence in each estimate. Changing the culture of project estimating from point estimates to estimating ranges will require a major attitude shift. , “It is more challenging to determine the investment in the presence of significant uncertainty [as opposed to point estimates] as to the project’s return on investment. It requires a corporate culture and leadership that can tolerate and even embrace this ambiguity” (Chelst and Canbolat 2012).

The commercial software used to train and test the artificial neural network was not compatible to bootstrap sampling, as such the iterations were completed manually and it was time consuming limiting the iterations to 100. More iterations or a larger data-set should better enclose the actual costs around the extremes, although 35/38 is 92% of the time correct. Further studies could extend the data-sets, conduct more iterations and investigate the sampling fraction used (80%) along with trialing sampling WOR compared to WR.

Conclusion

Point estimates are simply one number with no indication of the level of confidence behind that number. In later estimating stages, quantities are known, and highway agencies can be more confident and can express the estimate in that form. For the earlier estimate stages, where confidence is lower, the estimate should be expressed in a manner that describes the estimator’s confidence and providing a range does just that. This research has shown the power that developing an empirical distribution has for expressing the point estimate as a distribution of likely costs. This research found that not all projects have the same level of confidence, as such individual contingencies require a rational basis for their amount rather than a fixed percentage of construction costs.

CHAPTER 7. CONCLUSIONS AND LIMITATIONS

Conclusions

This section presents the main findings from each of the three research papers in Chapters 4, 5 and 6. Chapter 4 presented a method to rationally sample data that could be used for data-driven techniques such as artificial neural networks or multiple-regression analysis:

- Firstly, when all available 151 data points were used to train the model, the error in testing the model on the remaining 38 projects was the lowest. This finding aligns with literature suggestions where more data for testing and training the model will increase accuracy and the reliability of that model.
- When less than 151 data points were used to train the model, the error in testing the remaining 38 projects was least when the distribution of key input attributes were reflected in the sample of data.

Findings from Chapter 4 were used to rationally select the 38 projects to test the model against. This MAPE reported of 22.9% is therefore reflective of the error for future project predictions. The same 38 projects were used to test the model in Chapters 5 and 6.

Chapter 5 focused on quantifying the efforts to conduct the conceptual development stage estimate. The effort collected was perceptive through a survey at a highway agency. The major findings of Chapter 5 were:

- Selecting input variables that have a high influence on the construction cost but require a low level of effort to calculate or identify was proven to be a rational selection method.
- The case-study showed that once 6-8 variables were added to the model then further detail yielded no reduction in the estimating error.
- Highway agencies do not need to store and collect more input variables than required. In doing so only increases the demand on data storage and efforts to collect the data with little to no increase in performance.

Chapter 5 proved for the first time that not all project attributes need to be known to calculate the construction cost at the conceptual stage to reasonable accuracy. This result is positive for practitioners wanting to implement data-driven techniques.

Chapter 6 leveraged the artificial neural network created in Chapters 3 and 4 by combining the method with bootstrap sampling. The purpose of this was to express the conceptual estimate as range as opposed to a point estimate. Point estimates can result in overconfidence and not communicate the lack of uncertainty associated with the conceptual cost estimate.

- Bootstrap sampling combined with artificial neural networks were proven as a suitable method to produce a range estimate for highway projects.
- The range estimates better communicated the expected construction costs at the conceptual stage as opposed to a point estimate. There was an improved ability for the cost estimate to have a contingency assigned not simply based as a percentage of the construction cost.
- The empirical distribution produced confidence intervals for all 38 test projects. Because the distribution was empirical and specific to each project then no assumptions were necessary, typically required when fitting a named distribution.

Limitations

The limitations in Chapter 4 start with the content analysis completed. There may be other relevant literature in the field of early estimation for construction projects which were not considered in the content analysis of this chapter. The results of this section are therefore limited to the 16 publications investigated. Chapter 4 then makes use of the database developed in Chapter 3. The results from the ANN cannot be generalized to include all highway projects. The same database was utilized throughout Chapters 5 and 6 and therefore these all have this same limitation.

In Chapter 5 the content analysis from Chapter 4 was extended, thus very similar limitations exist. Additional to the content analysis, perceptive data for estimating effort and influence of the input variables was collected through a survey in Chapter 5. This survey was conducted at MDT and completed by 31 employees suitably qualified to do so. The results of this survey cannot be extended beyond data from this agency. The results of Chapter 5 are therefore only relevant to the data collected for that agency.

Chapter 6 leverages the data-driven ANN model created in Chapters 3 and 4. Bootstrap sampling was combined with ANN. The ability to communicate the estimate as a range was only demonstrated through 38 test projects and as such the results are limited to that data.

CHAPTER 8. CONTRIBUTIONS AND RECCOMENDATIONS FOR FUTURE RESEARCH

Contributions

The major contribution of this research was that for the first time it was proven that at the conceptual estimating stage once enough information is known then adding further detail does not enhance the estimate accuracy. This is significant for practitioners willing to trial data-driven CCE techniques. Practitioners can instead concentrate on creating an accurate database with those variables that have a high impact on construction cost and do not require a high-level of effort. The specific contributions are outlined for each chapter below.

Chapter 4 for the first time identified that some previously published literature on data-driven CCE models are reporting models with such low prediction error capabilities, yet are powered by databases with very few projects. This was not in line with literature reports that larger databases increased the accuracy and reliability of estimating models. Furthermore, it was identified that no literature had reported their method to select the data for their models in a rational way. As such, this research introduced a method that could be used if an entire database were not to be used for estimate modelling. It also contributed to the body of knowledge further proving the statement that ‘larger databases increase the reliability and accuracy of a model’, this later point being contrary to the content analysis of select publications.

Chapter 5 attempted to quantify the level of effort required to conduct the conceptual estimate, previously this has never been attempted in the field of construction cost estimating. A new objectives hierarchy tree was proposed at the CCE stage, that being creating a model to predict with reasonable accuracy but require a low level of effort. Previous research has only focused on reducing the estimating error. To address this, a methodology to select input variables which meet the dual-objective framework was proposed. The paper’s primary contribution was significant for both researchers and practitioners – for the first time it was proven that while increasing the number of input variables in an early estimate may appear to enhance estimate accuracy on an intuitive basis, this is not necessarily true. Once around 6 to 8 high impact/low effort variables were included in both the MRA or ANN models then further input variables yielded diminishing returns in the estimate error.

Finally, Chapter 6 compared stochastic cost estimating to the point estimate which is the typical format at the conceptual stage. The benefits of bootstrap sampling combined with artificial neural networks was displayed with data from a highway agency for the first time. The ability for the stochastic estimate to reflect the true confidence in the estimate at the conceptual stage was displayed. Much literature has produced data-driven models that construction practitioners could use to calculate a point estimate at the conceptual stage, instead Chapter 6 challenges the overconfidence associated with a point estimate. Specifically, the challenge is laid for highway agencies to not assign contingency based on a fixed percentage of the construction cost. Data-driven methods such as proposed in this chapter display the ability for this to occur.

Recommendations for Future Research

The ability for data-driven techniques, such as ANN and MRA, to calculate the conceptual cost of projects has been proven in the literature but has not yet been implemented by highway agencies. This thesis contributes to methods and improvements in order for that to occur. Further research in this area could continue as follows:

- All techniques and methods could be proven with larger databases for increased reliability. Additionally the methods could be extended beyond the scope of works displayed in this research.
- The stochastic results produced in Chapter 6 was completed manually for 100 iterations due to the incompatibility of the software's. Future research would conduct more than 100 iterations of the stochastic estimating method by creating a software or method to combine the ANN modeling with bootstrap sampling.
- It is known that the commercial software selected to conduct the ANN or MRA modeling will affect the output. As part of this research only one commercial software was tested. Future research would test the methodologies and practices on alternative software to validate results or investigate better performance.
- A method to predict the likelihood of construction growth from the initial estimate (CGIE) would benefit the estimate at the conceptual stage. It has become apparent that the confidence in estimating the construction costs at the initial stage is hugely variant for each project. A data-driven method could be investigated to recognize patterns between CGIE and the types of projects which produce higher variations. This could be used in combination with the point estimate to produce a cost estimate with associated contingency or a confidence rating index. Alternatively, it could be correlated with the

estimate ranges produced in Chapter 6 as further validation of the range estimates produced for construction cost estimate for each of the 38 test projects.

- In this thesis there was a larger focus on the use of ANN and only one chapter produced results using MRA. In that chapter an assumption of linearity was made to produce the predictions. An emerging method called Multivariate Adaptive Regression Splines (MARS) is being trialed in other fields with success. The prediction model produces a series of “piecewise” linear relationships (Haleem et al. 2013). Further research could extend this into the field of highway construction cost estimating.
- The database was created with an assumed inflation rate applied to all construction costs. This rate of 3% was based on experience from the highway agency that provided the data, MDT. Of the publications studied in this research there was silence on the method or inflation rate applied to their construction costs. Further research could investigate a method to predict the inflation rate, as opposed to using historical averages, for best prediction results.
- Much literature studied on artificial neural networks select the input variables based on expert opinion and trial and error. In this research a method was proposed to select the variables based on the perceptive level of effort and influence on the construction cost. Future research could investigate decision analysis methods, for example multi-attribute utility theory (MAUT), to select the best input variables for their models.

BIBLIOGRAPHY

American Association of State Highway and Transportation Officials (AASHTO). (2013). *Practical Guide to Cost Estimating, First Edition,* American Association of State Highway and Transportation Officials Washington, DC.

Al-Tabtabai, H., Alex, A. P., and Tantash, M. (1999). "Preliminary Cost Estimation of Highway Construction Using Neural Networks." *Cost Engineering,* 41(3), 19-24.

Anderson, S., Molenaar, K., and Schexnayder, C. (2007). *Final Report for NCHRP Report 574: Guidance for Cost Estimation and Management for Highway Projects During Planning, Programming and Preconstruction,* National Cooperative Highway Research Program (NCHRP), Transportation Research Board of the National Academics, Washington D.C.

Bazeley, P., and Richards. R. (2000). "The NVivo: Qualitative Project Book." *SAGE Publications Inc.*, California. 1-199.

Bell, L. C., and Ghazanfer, A. B. (1987). "Preliminary Cost Estimating For Highway Construction Projects." *AACE Transactions,* C6.1-C6.4.

Bedford, T., and Cooke, R. (2001). "Probabilistic Risk Analysis: Foundations and Methods," *Cambridge University Press,* Cambridge, UK.

Bode, J. (2000). "Neural networks for cost estimation: simulations and pilot application." *International Journal of Production Research,* 38(6), 1231-1254.

Byrnes, J. E. (2002). "Best Practices for Highway Project Cost Estimating." M.S. thesis, Arizona State University, Mesa, AZ.

Chelst, K., Canbolat, Y. B. (2012). "Value-Added Decision Making for Managers." *CRC Press Taylor & Francis Group, New York*, 1-545.

Chernick, M. R. (1999). "Bootstrap Methods: A Practitioner's Guide." *John Wiley and Sons Inc, New York, NY.*

Chou, J. -S., Peng, M., Persad, K. R., and O'Connor, J. T. (2006) "Quantity-Based Approach to Preliminary Cost Estimates for Highway Projects." *Transportation Research Record: Journal of the Transportation Research Board,* No. 1946*,* 22-30.

Creese, R. C., and Li, L. (1995). "Cost Estimation of Timber Bridges Using Neural Networks." *Cost Engineering,* 34(5), 17-22.

Danielsson, P. –E. (1980). "Euclidean Distance Mapping." *Computer Graphics and Image Processing,* 14, 227-248.

Davison, A. C., and Hinkley D. V. (1997). "Bootstrap Methods and their application." *Cambridge series in statistical and probabilistic mathematics,* Cambridge University Press, UK.

Dillman, D.A., Smyth, J.D., and Christian, L.M. (2009). "Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method." *John Wiley & Sons Inc,* New Jersey.

Dupret, G., Koda, M. (2000). "Bootstrap re-sampling for unbalanced data in supervised learning." *European Journal of Operational Research,* 134, 141-156.

Efron B., and Tibshirani, R. J. (1993). "An Introduction to the Bootstrap." *Chapman and Hall*, New York, NY.

Elhag, T. M. S., and Boussabaine, A. H. (1998). "An artificial neural system for cost estimation of construction projects." *Association of Researchers in Construction Management,* 1, 219-26.

Emsley, M. W., Lowe, D. J, Duff, A. R., Harding, A., and Hickson, A. (2002) "Data modelling and the application of a neural network approach to the prediction of total construction costs." *Construction Management and Economics,* 20, 465-472.

Federal Highway Administration (FHWA). (2015). "Fact Sheets on Highway Provisions – Statewide Planning." *Statewide Planning,* <http://www.fhwa.dot.gov/safetealu/factsheets/statewide.htm> (Aug. 17th, 2015).

Federal Highway Administration (FHWA). (2007). "Major Project Program Cost Estimating Guidance."<https://www.fhwa.dot.gov/ipd/project_delivery/tools_programs/cost_estimating/guidance.aspx> (Oct. 14th, 2015).

Fink, A. (2009). "How to Conduct Surveys: A Step-by-Step Guide 4th edition." SAGE Publications, Inc., Thousand Oaks, CA, 1-125.

Flyvbjerg B., Skamris Holm, M., and Buhl, S. (2002). "Underestimating Costs in Public Works Projects: Error or Lie?" *Journal of the American Planning Association,* 68(3), 279-295.

Fowler, F. J. (2009). "Survey Research Methods 4th Edition." SAGE Publications Inc, Thousand Oaks, CA, 1-199.

Gransberg, D. D., Lopez del Puerto, C., and Humphrey, D. (2007). "Relating Cost Growth from the Initial Estimate to Design Fee for Transportation Projects." *Journal of Construction Engineering and Management,* KICEM, 133(6), 404-408.

Gransberg, D. D., Shane J. S., and Ahn J. (2011) "A Framework for Guaranteed Maximum Price and Contingency Development for Integrated Delivery of Transportation Projects." *Journal of Construction Engineering and Project Management,* 1(1), 1-10.

Gunduz, M., Ugur, L. O., and Ozturk, E., (2011). "Parametric cost estimation system for light rail transit and metro trackworks." *Expert Systems with Application,* 38, 2873-2877.

Gunaydin, H. M., and Dogan, S. Z., (2004). "A neural network approach for early cost estimation of structural systems of buildings." *International Journal of Project Management,* 22, 595-602.

Haleem, K., Gan, A., Lu, J. (2013). "Using multivariate adaptive regression splines (MARS) to develop crash modification factors for urban freeway interchange influence areas." *Accident Analysis and Prevention,* 55, 12-21.

Harbuck, R. H. (2007). "Are Accurate Estimates Achievable During the Planning of Transportation Projects?" *AACE International Transactions,* 16.1–16.6.

Hegazy, T., and Ayed, A., (1998). "Neural Network Model for Parametric Cost Estimation of Highway Projects." *Journal of Construction Engineering and Management,* 124(3), 210-218.

Jeong, H. S., and Woldesenbet, A. (2012). "Procedures and Models for Estimating Preconstruction Costs of Highway Projects." *Oklahoma Transportation Center,* OTCREOS10.1-19-F.

Kaplan, S., and Garrick, B. J. (1981). "On the Quantitative Definition of Risk." *Society for Risk Analysis*, 1(1), 11-27.

Kim, G. –H., An, S. –H., and Kang, K. –I., (2004). "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning." *Building and Environment,* 39, 1235-1242.

Kim, H., Seo, Y., and Hyun, C., (2012). "A hybrid conceptual cost estimating model for large building projects." *Automation in Construction,* 25, 72-81.

Lord, J., and Asante, M. A. (1999). "Estimating Uncertainty Ranges for Costs by the Bootstrap Procedure Combined with Probabilistic Sensitivity Analysis." *Health Economics,* 8, 323-333.

Lowe, D. J., Emsley, M. W., and Harding, A. (2006). "Predicting Construction Cost Using Multiple Regression Techniques." *Journal of Construction Engineering and Management,* 132(7), 750-758.

Mahamid, I. (2011). "Early Cost Estimating for Road Construction Projects Using Multiple Regression Techniques." *Australasian Journal of Construction Economics and Building*, 11(4), 87-101.

Minassian, V. K., and Jergeas, G. F. (2009). "A Prototype Risk Analysis for Determining Contingency Using Approximate Reasoning Method." *Cost Engineering,* 51(1), 26-33

Molenaar, K. R., (2005) "Programmatic Cost Risk Analysis for Highway Megaprojects." *Journal of Construction Engineering and Management,* 131(3), 343-353.

Montana Department of Transportation (MDT). *Cost Estimation Procedure for Highway Design Projects.* January 2007. <http://www.mdt.mt.gov/other/roaddesign/external/report_templates_guidance/costest_procedure_jan07rev.pdf> . (July 14th, 2015).

Moselhi, O., Hegazy, T., and Fazio, P., (1992). “Potential applications of neural networks in construction.” *Canadian Journal of Civil Engineering,* 19, 521-529.

Moselhi, O., and Siqueira, I., (1998). “Neural Networks for Cost Estimating of Structural Steel Buildings.” *AACE International Transactions,* 6.1-6.4.

Petroutsatou, K., Georgopoulos, E., Lambropoulos, E., and Pantouvakis J. –P. (2012). “Early cost estimating of road tunnel construction using neural networks.” *Journal of construction engineering and management,* 138(6), 679-687.

Petroutsatou, C., Lambropoulos, S., and Pantouvakis, J. -P. (2006). “Road Tunnel Early Cost Estimates Using Multiple Regression Analysis.” *Operational Research. An International Journal,* 6(3), 311-322.

Pewdum, W., Rujirayanyong, T., and Sooksatra, V., (2009). “Forecasting final budget and duration of highway construction projects.” *Engineering, Construction and Architectural Management,* 16(6), 544-557.

Rajkumar, T., and Bardina, J. (2003). “Training data requirement for a neural network to predict aerodynamic coefficients.” *Nasa Ames Research Center, California,* 1-12.

Sanders, S. R., Maxwell R. R., and Glagola C. R. (1992). “Preliminary Estimating Models for Infrastructure Projects.” *Cost Engineering,* 34(8), 7-13.

Schexnayder, C. J., Weber, S. L., and Fiori, C. (2003). “NCHRP Synthesis of Highway Practice: Project Cost Estimating.” *Transportation Research Board of the National Academics, Washington D.C.*

Setyawati, B. R., Sahirman, S., and Creese. R. C. (2002). “Neural Networks for Cost Estimation.” *AACE International Transactions,* 13.1-13.9.

Shane, J. S., Molenaar, K. R., Anderson, S., Schexnayder, C. (2009) “Construction Project Cost Escalation Factors.” *Journal of Management in Engineering,* 25(4), 221-229.

Sillars, D. N., and O’Connor, M. B. (2007). “Evolving Risk Analysis Techniques: Managing Project Development Risk from a Top-Down Perspective.” *Proc., Construction Research Congress*, ASCE, Bahamas, 914-924.

Smith, A. E., and Mason, A. K. (1997). “Cost Estimation Predictive Modeling: Regression versus Neural Network.” *The Engineering Economist,* 42(2), 137-167.

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

Sonmez, R. (2008). “Parametric Range Estimating of Building Costs Using Regression Models and Bootstrap,” *Journal of Construction Engineering and Management,* 134(12). 1011-1016.

Sonmez, R. (2011). “Range estimation of construction costs using neural networks with bootstrap prediction intervals.” *Expert systems with applications,* 38, 9913-9917.

Tatari, O., and Kucukvar, M. (2011). “Cost premium prediction of certified green buildings: A neural network approach. Building and Environment.” *Building and Environment,* 46, 1081-1086.

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

Turochy, R. E., Hoel, L. A., and Doty, R. S. (2001). “Highway Project Cost Estimating Methods used in the Planning Stage of Project Development VTRC 02-TAR3.” *Virginia Transportation Research Council,* 1-290.

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

Walczak, S. (2001). “An Empirical Analysis of Data Requirements for Financial Forecasting with Neural Networks.” *Journal of Management Information Systems,* 17(4), 203-222.

Walton, J.R., and Stevens, J.D. (1997). “Improving Conceptual Estimating Methods Using Historical Cost Data.” *Transportation Research Record: Journal of the Transportation Research Board,* No. 1575, 127-131.

APPENDIX A. INPUT VARIABLES

Date: 7/30/2015					Data statistics of the 189 projects and the input variables					
					Complexity Rating System			Binary rating		Other data input
	Suggested 29 inputs:	**Available:**	**Measures:**	**Data Source**	**High**	**Medium**	**Low**	**Yes**	**No**	
1	Urban or rural project	Y	Urban indicator	PPMS				41	148	
			District	PPMS						District 1: 57; District 2: 54; District 3: 42; District 4: 18; District 5: 18
2	Construction on Native American Reservations	Y	Binary Y/N indicator	PPMS, PFR				15	174	
3	Context sensitive design issues, controversy - level of environmental documentation									
4	Design AADT	Y	AADT at let year	GIS						AADT continuous range from 100 to 20667
			Highway functional classification	PFR						Collector: 38; Minor Arterial: 57; Principal Arterial (interstate): 34; Principal Arterial (non-interstate): 60
5	Design speed(s)	Y	Design Speed	PFR						Range from 30mph to 70mph
6	Site topography (steep, flat or undulating terrain)	Y	Terrain	PFR						Flat: 74; Rolling: 92; Mountainous: 23
7	Start and End Stations, Length and Width	Y	Length, width, area	TIS, PPMS, PFR						Length ranges from 0.6 miles to 26.84 miles
8	Existing surfacing conditions and depths									
9	Number of intersections in project									
10	Number of bridges in the project scope	Y	Number for deck treatment	PFR						Range from 0 to 9 bridge deck treatments for all projects
11	Intersection signalization and signage	Y	Signage and pavement marking complexity	PFR	114	57	18			
12	Letting Date	Y	Let quarter and year							Year 2009: 47; Year 2010: 50; Year 2011: 32; Year 2012: 39; Year 2013: 21
13	Horizontal and Vertical Alignment									
14	Extent of changes to the existing intersections									
15	Typical Section (depths of surfacing and aggregate)	Y	%mill	PFR						Proportion ranges from 0 to 1 on continuous scale
			%overlay	PFR						Proportion ranges from 0 to 1 on continuous scale
16	Curb & Gutter and Sidewalk	Y	ADA/sidewalk complexity	PFR	167	11	11			
17	Bridge type (steel or concrete) and complexity									
18	Volumes of excavation and embankment									
19	Geotechnical - subsurface & slope recommendations	Y	Geotechnical complexity	PFR	155	23	11			
20	Bridge deck area	Y	Area of deck treatments	PFR						0 square feet to 118,000 square feet on a continuous scale
21	Traffic Control - closures or detours	Y	WZSM	PFR						Level 1: 10; Level 2; 91; Level 3: 88
			Railroad complexity	PFR	168	21	0			
22	Environmental permitting requirements- wetlands									
23	Hydraulic recommendations and culverts									
24	Storm Sewer extents									
25	Bridge span lengths (between supports)									
26	Foundation complexity of the bridge									
27	Right-of-way acquisition and costs	Y	ROW complexity	PFR	186	3	0			
28	Extent of Utility relocations and costs	Y	Utility complexity	PFR	85	52	52			
29	Contract Time	Y	Contract time	PPMS						Range up to 260 days

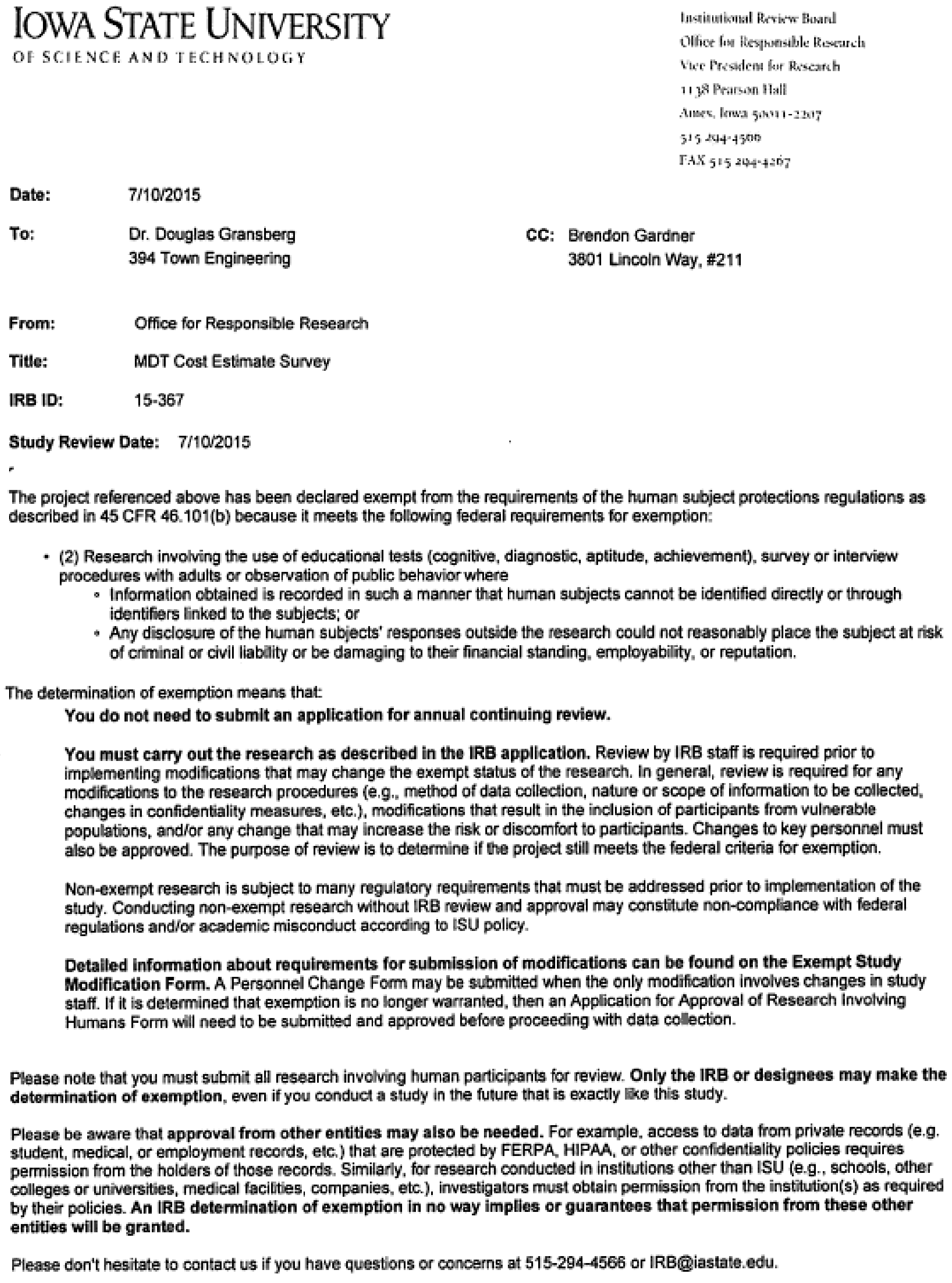
APPENDIX B. COMPLEXITY RATING CHART

Terrain/Topography	**Flat** Generally flat, fairly flat etc	**Rolling** Flat and rolling or gently rolling	**Mountainous** Gorges, steep terrain etc

	Low	**Medium**	**High**
Geotechnical Involvement	No digouts or other geotech	Roadway projects will require minor digouts Additional spot mill/fill in projects not receiving any mill (<3 intersections or bridge approaches or thick bridge mill in chipseal or overlay project)	Extensive sections of roadway digouts >3 spot mill/fill over and above the mainline works Relevel bridge approach slabs Multiple of the medium type works
Traffic signs and pavement markings	Standard pavement-marking replacement only (required on all projects) Or “traffic to assess reflectivity and upgrades required”	Standard pavement-marking replacement with any of the following two: - Replace or upgrade signs - Changes to pavement markings required/TWTL markings/lane changes - Significant pavement marking upgrades in urban area - Some sections of rumble-strip - Minor and singular safety sign: Weigh-In-Motion advance sign etc… intersection advance signs Or none of the above but rumble strips on the entire project.	As with medium rating plus any: - Flashing signs or traffic lights - Overhead signs - Lighting - Substantial upgrades to rumble-strips and any of the other medium works
Railroad Involvement	Low likelihood of requiring agreement >50ft from railroad	Possibly flagmen at times Project areas within 50ft of railroad and railway insurance required	Flagmen at times MRL agreement R/W acquisition and/or utility involvement with railroad
Utility Complexity	No utility involvement	Medium rating for any of the geotechnical, ADA/sidewalk or guardrail to reflect the possible utility identification or relocation No major utility relocations And/or Mill/Fill in urban area requiring ironwork to be raised and protected	High rating for any of geotechnical, ADA/sidewalk or guardrail or Significant utility disturbance is known
Environmental issues	Categorical Exclusion	Categorical Exclusion or Environmental Assessment	Environment Impact Study or complex Environmental Assessment required Studies of multiple alternatives

	Minimum interaction with environmental and permitting agencies Minor environmental impacts Do not involve cultural resources, hazardous waste, Section 4(f) evaluations or substantial flood plain encroachments	Cultural Resources (historical, archaeological etc), SHPO Wetland mitigation, 124 notification, 404 permit required Parkland involvement, hazardous waste, floodplain encroachments Water and air pollution mitigation Major coordination with Game or Fish and Boat commissions Endangered species Migratory Birds Cores required to test if AC is contaminated with asbestos	Continued public and elected officials involvement in analyzing and selecting alternates Other agencies (such as FHWA, COE, EPA, Fish, Wildlifte & Parks, DEQ, etc) are heavily involved to protect air; water; game; fish, threatened and endangered species; cultural resources (historical, archaeological, parks, wetlands, etc) etc Tribal involvement with resources
Guardrail (on bridge or highway)	No guardrail work	Either: - 1 rail upgrade or a few (1-3) bridges requiring end terminus upgrades - Awaiting recommendations from safety - Guardrail extensions on 1-bridge - Guardrail repairs - Minor guardrail replacement	Significant upgrades possibly involving: - >3 end terminus on guardrails - Guardrail extensions - Concrete bridge rails - Raising heights on >1 bridges or other guradrails - Entirely new guardrail installation - >1 rail upgrades
ADA and sidewalk	None	1 ADA intersection upgrade and/or minor sidewalk involvement or traffic furniture Detectable warning signs being added	More than 1 ADA upgrade and/or extensive sidewalk upgrades Curbing or traffic furniture upgrades.

APPENDIX C. SURVEY AND RESULTS



IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
1138 Pearson Hall
Ames, Iowa 50011-2207
515 294-4566
FAX 515 294-4267

Date: 7/10/2015

To: Dr. Douglas Gransberg
394 Town Engineering

CC: Brendon Gardner
3801 Lincoln Way, #211

From: Office for Responsible Research

Title: MDT Cost Estimate Survey

IRB ID: 15-367

Study Review Date: 7/10/2015

The project referenced above has been declared exempt from the requirements of the human subject protections regulations as described in 45 CFR 46.101(b) because it meets the following federal requirements for exemption:

- (2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey or interview procedures with adults or observation of public behavior where
  - Information obtained is recorded in such a manner that human subjects cannot be identified directly or through identifiers linked to the subjects; or
  - Any disclosure of the human subjects' responses outside the research could not reasonably place the subject at risk of criminal or civil liability or be damaging to their financial standing, employability, or reputation.

The determination of exemption means that:

- **You do not need to submit an application for annual continuing review.**
- **You must carry out the research as described in the IRB application.** Review by IRB staff is required prior to implementing modifications that may change the exempt status of the research. In general, review is required for any modifications to the research procedures (e.g., method of data collection, nature or scope of information to be collected, changes in confidentiality measures, etc.), modifications that result in the inclusion of participants from vulnerable populations, and/or any change that may increase the risk or discomfort to participants. Changes to key personnel must also be approved. The purpose of review is to determine if the project still meets the federal criteria for exemption.

  Non-exempt research is subject to many regulatory requirements that must be addressed prior to implementation of the study. Conducting non-exempt research without IRB review and approval may constitute non-compliance with federal regulations and/or academic misconduct according to ISU policy.

  Detailed information about requirements for submission of modifications can be found on the Exempt Study Modification Form. A Personnel Change Form may be submitted when the only modification involves changes in study staff. If it is determined that exemption is no longer warranted, then an Application for Approval of Research Involving Humans Form will need to be submitted and approved before proceeding with data collection.

Please note that you must submit all research involving human participants for review. **Only the IRB or designees may make the determination of exemption**, even if you conduct a study in the future that is exactly like this study.

Please be aware that **approval from other entities may also be needed.** For example, access to data from private records (e.g. student, medical, or employment records, etc.) that are protected by FERPA, HIPAA, or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **An IRB determination of exemption in no way implies or guarantees that permission from these other entities will be granted.**

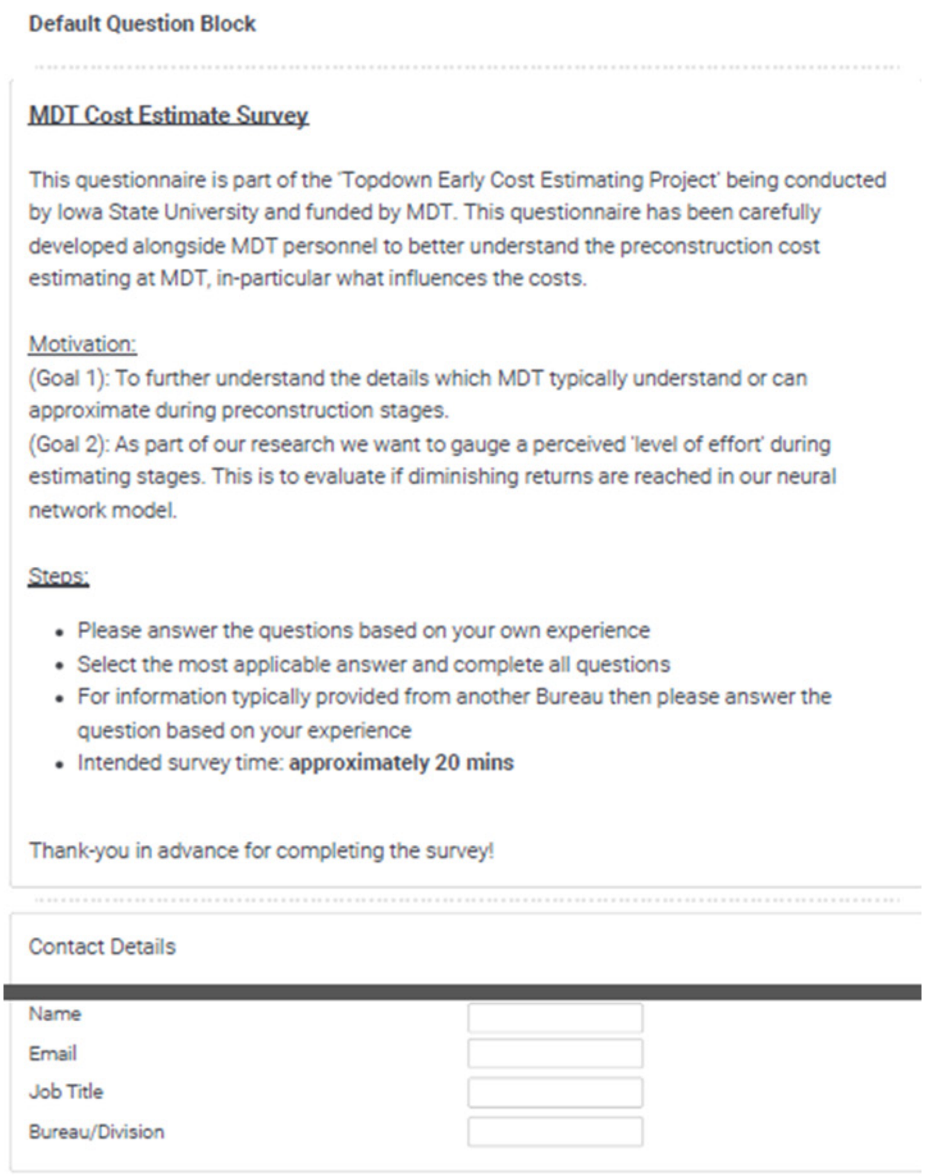
Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.

IRB #: 15-136 Exemption Approval

Review Date: 7/10/2015

MDT IOWA STATE UNIVERSITY Institute for Transportation

Confused by the questions? - feel free to contact
Brendon Gardner bgardner@iastate.edu (515) 708 011
Jeania Cereck jcereck@mt.gov (406) 454 589
Christie McOmber cmcomber@mt.gov (406) 454 590



Default Question Block

MDT Cost Estimate Survey

This questionnaire is part of the 'Topdown Early Cost Estimating Project' being conducted by Iowa State University and funded by MDT. This questionnaire has been carefully developed alongside MDT personnel to better understand the preconstruction cost estimating at MDT, in-particular what influences the costs.

Motivation:
(Goal 1): To further understand the details which MDT typically understand or can approximate during preconstruction stages.
(Goal 2): As part of our research we want to gauge a perceived 'level of effort' during estimating stages. This is to evaluate if diminishing returns are reached in our neural network model.

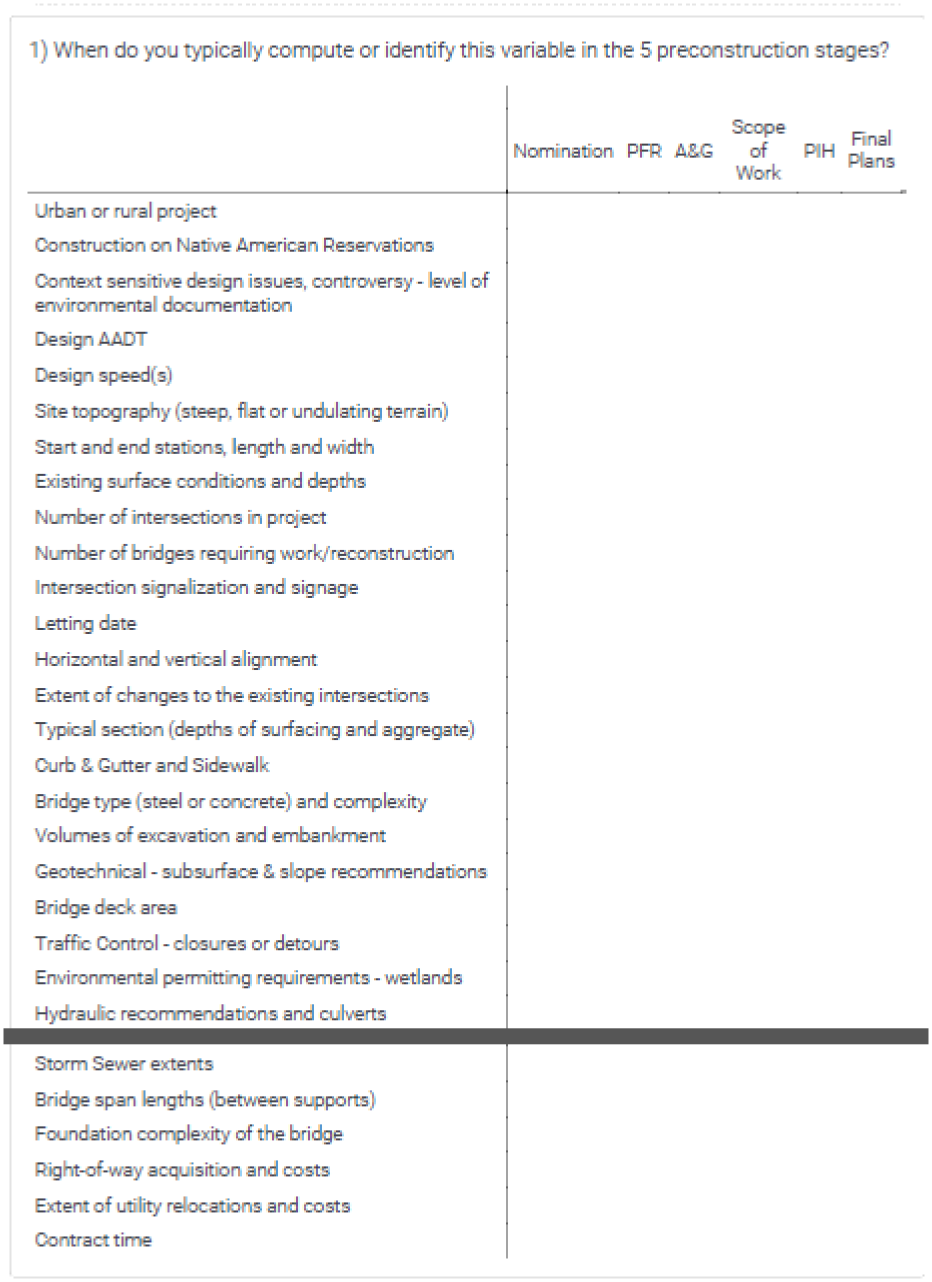
Steps:

- Please answer the questions based on your own experience
- Select the most applicable answer and complete all questions
- For information typically provided from another Bureau then please answer the question based on your experience
- Intended survey time: **approximately 20 mins**

Thank-you in advance for completing the survey!

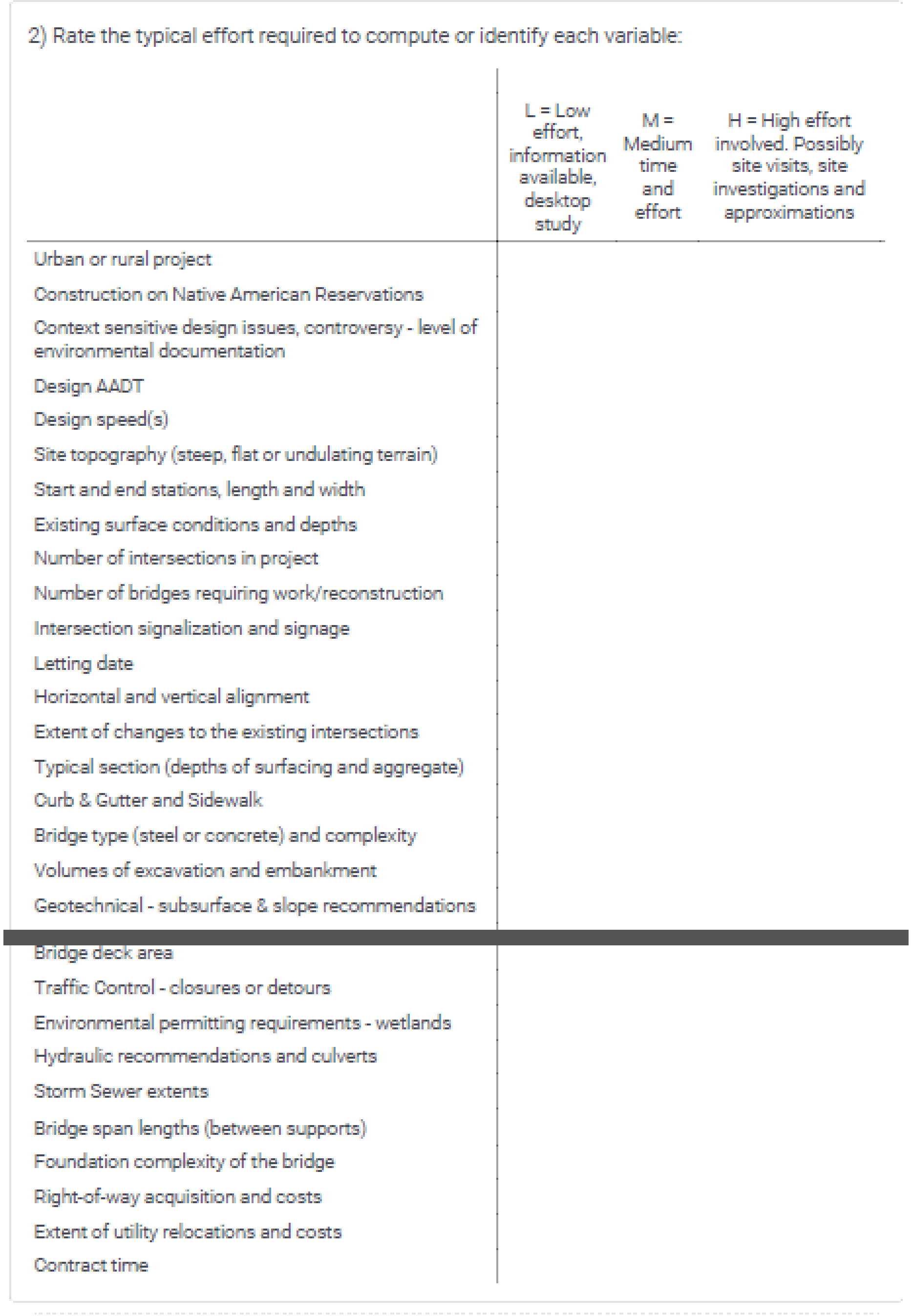
Contact Details

Name □
Email □
Job Title □
Bureau/Division □



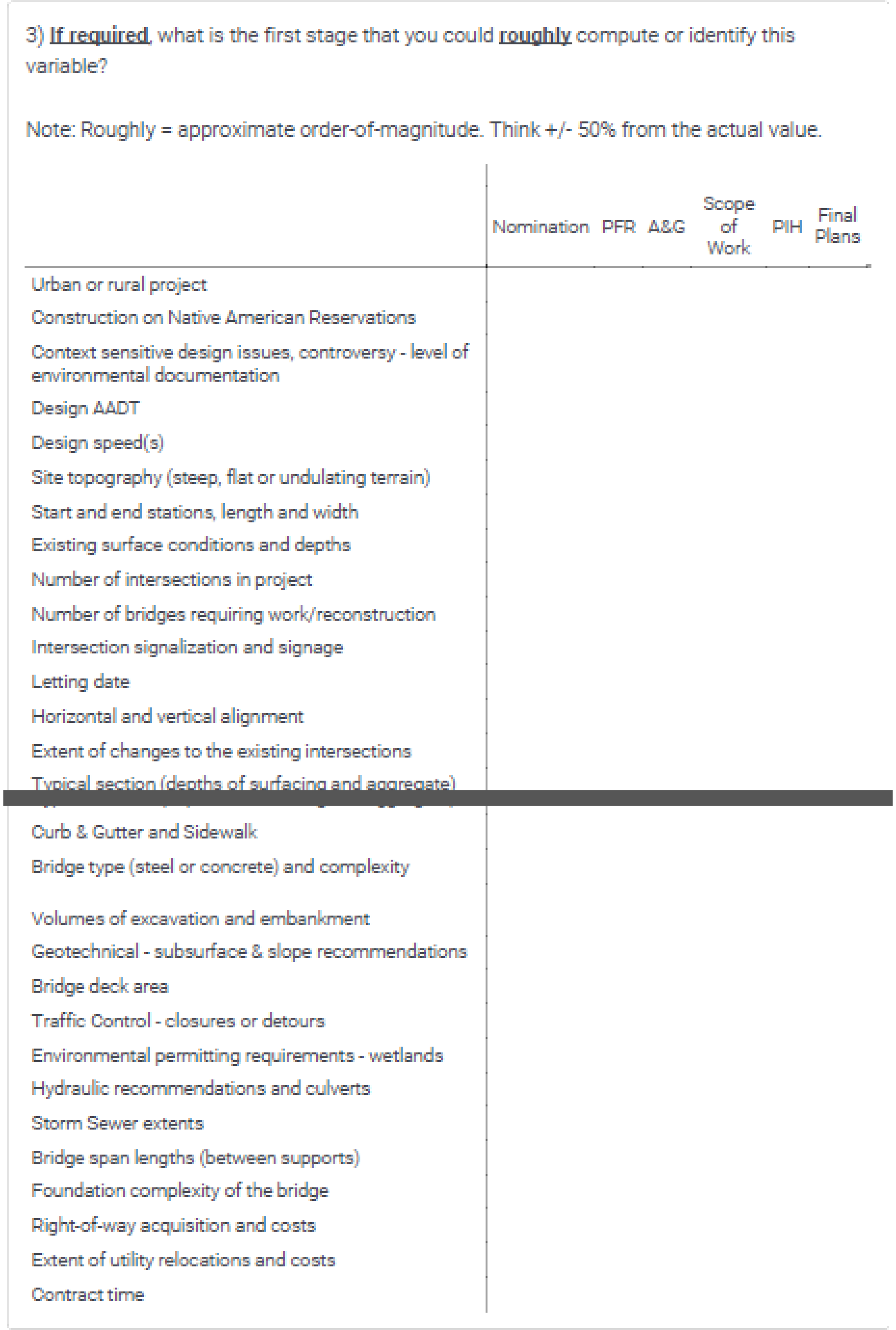
1) When do you typically compute or identify this variable in the 5 preconstruction stages?

	Nomination	PFR	A&G	Scope of Work	PIH	Final Plans
Urban or rural project						
Construction on Native American Reservations						
Context sensitive design issues, controversy - level of environmental documentation						
Design AADT						
Design speed(s)						
Site topography (steep, flat or undulating terrain)						
Start and end stations, length and width						
Existing surface conditions and depths						
Number of intersections in project						
Number of bridges requiring work/reconstruction						
Intersection signalization and signage						
Letting date						
Horizontal and vertical alignment						
Extent of changes to the existing intersections						
Typical section (depths of surfacing and aggregate)						
Curb & Gutter and Sidewalk						
Bridge type (steel or concrete) and complexity						
Volumes of excavation and embankment						
Geotechnical - subsurface & slope recommendations						
Bridge deck area						
Traffic Control - closures or detours						
Environmental permitting requirements - wetlands						
Hydraulic recommendations and culverts						
Storm Sewer extents						
Bridge span lengths (between supports)						
Foundation complexity of the bridge						
Right-of-way acquisition and costs						
Extent of utility relocations and costs						
Contract time						



2) Rate the typical effort required to compute or identify each variable:

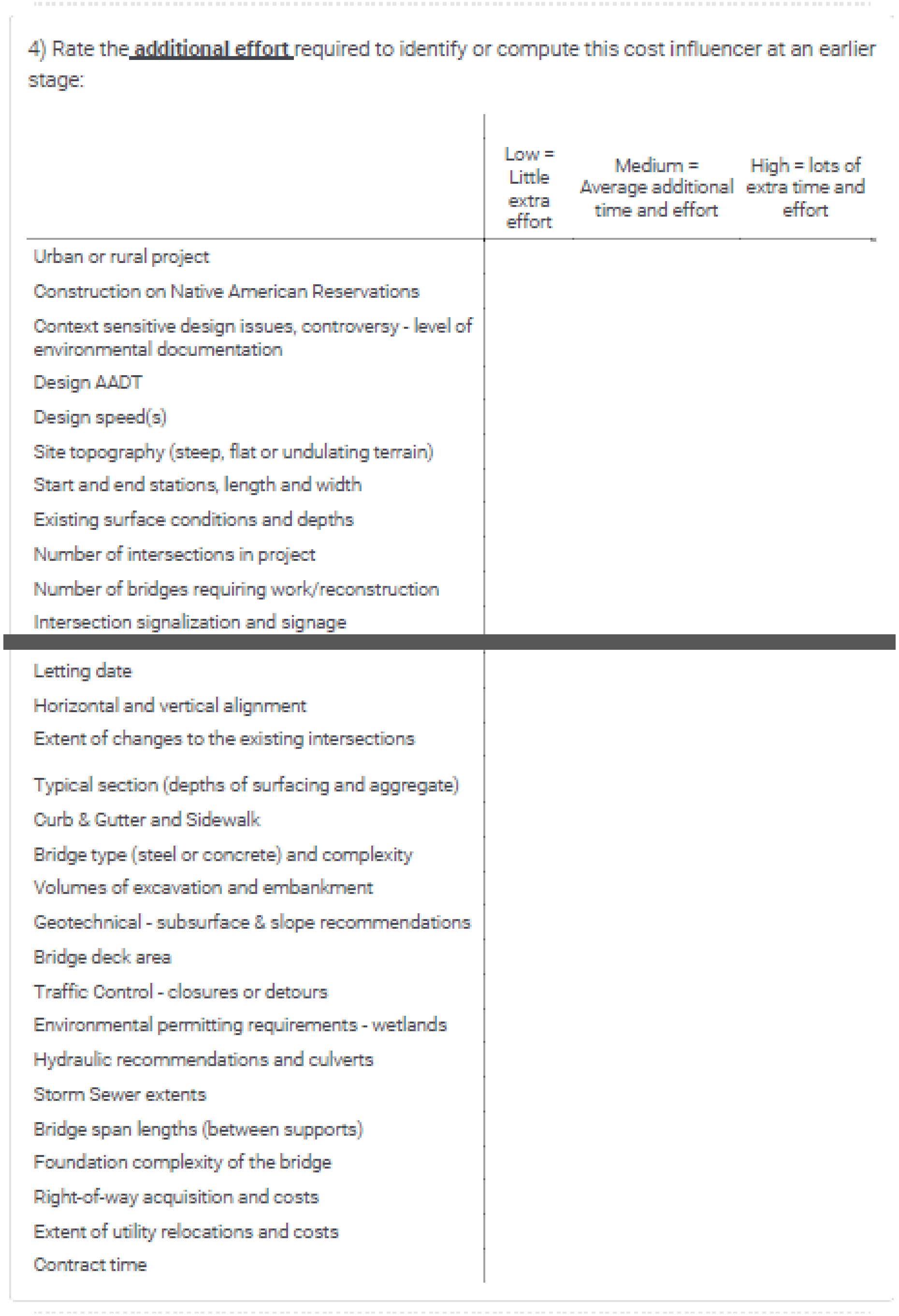
	L = Low effort, information available, desktop study	M = Medium time and effort	H = High effort involved. Possibly site visits, site investigations and approximations
Urban or rural project			
Construction on Native American Reservations			
Context sensitive design issues, controversy - level of environmental documentation			
Design AADT			
Design speed(s)			
Site topography (steep, flat or undulating terrain)			
Start and end stations, length and width			
Existing surface conditions and depths			
Number of intersections in project			
Number of bridges requiring work/reconstruction			
Intersection signalization and signage			
Letting date			
Horizontal and vertical alignment			
Extent of changes to the existing intersections			
Typical section (depths of surfacing and aggregate)			
Curb & Gutter and Sidewalk			
Bridge type (steel or concrete) and complexity			
Volumes of excavation and embankment			
Geotechnical - subsurface & slope recommendations			
Bridge deck area			
Traffic Control - closures or detours			
Environmental permitting requirements - wetlands			
Hydraulic recommendations and culverts			
Storm Sewer extents			
Bridge span lengths (between supports)			
Foundation complexity of the bridge			
Right-of-way acquisition and costs			
Extent of utility relocations and costs			
Contract time			



3) **If required**, what is the first stage that you could **roughly** compute or identify this variable?

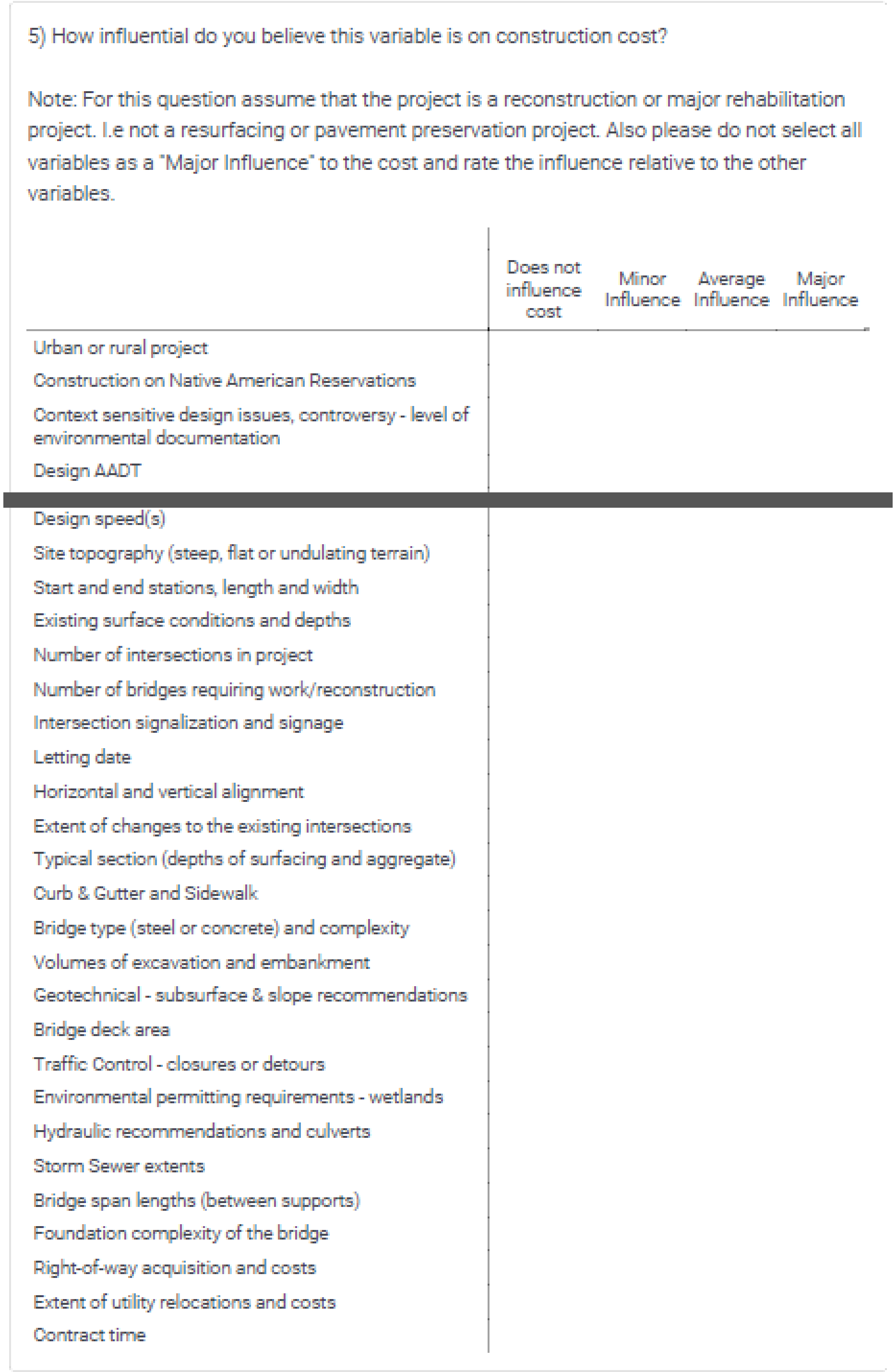
Note: Roughly = approximate order-of-magnitude. Think +/- 50% from the actual value.

	Nomination	PFR	A&G	Scope of Work	PIH	Final Plans
Urban or rural project						
Construction on Native American Reservations						
Context sensitive design issues, controversy - level of environmental documentation						
Design AADT						
Design speed(s)						
Site topography (steep, flat or undulating terrain)						
Start and end stations, length and width						
Existing surface conditions and depths						
Number of intersections in project						
Number of bridges requiring work/reconstruction						
Intersection signalization and signage						
Letting date						
Horizontal and vertical alignment						
Extent of changes to the existing intersections						
Typical section (depths of surfacing and aggregate)						
Curb & Gutter and Sidewalk						
Bridge type (steel or concrete) and complexity						
Volumes of excavation and embankment						
Geotechnical - subsurface & slope recommendations						
Bridge deck area						
Traffic Control - closures or detours						
Environmental permitting requirements - wetlands						
Hydraulic recommendations and culverts						
Storm Sewer extents						
Bridge span lengths (between supports)						
Foundation complexity of the bridge						
Right-of-way acquisition and costs						
Extent of utility relocations and costs						
Contract time						



4) Rate the **additional effort** required to identify or compute this cost influencer at an earlier stage:

	Low = Little extra effort	Medium = Average additional time and effort	High = lots of extra time and effort
Urban or rural project			
Construction on Native American Reservations			
Context sensitive design issues, controversy - level of environmental documentation			
Design AADT			
Design speed(s)			
Site topography (steep, flat or undulating terrain)			
Start and end stations, length and width			
Existing surface conditions and depths			
Number of intersections in project			
Number of bridges requiring work/reconstruction			
Intersection signalization and signage			
Letting date			
Horizontal and vertical alignment			
Extent of changes to the existing intersections			
Typical section (depths of surfacing and aggregate)			
Curb & Gutter and Sidewalk			
Bridge type (steel or concrete) and complexity			
Volumes of excavation and embankment			
Geotechnical - subsurface & slope recommendations			
Bridge deck area			
Traffic Control - closures or detours			
Environmental permitting requirements - wetlands			
Hydraulic recommendations and culverts			
Storm Sewer extents			
Bridge span lengths (between supports)			
Foundation complexity of the bridge			
Right-of-way acquisition and costs			
Extent of utility relocations and costs			
Contract time			



5) How influential do you believe this variable is on construction cost?

Note: For this question assume that the project is a reconstruction or major rehabilitation project. I.e not a resurfacing or pavement preservation project. Also please do not select all variables as a "Major Influence" to the cost and rate the influence relative to the other variables.

	Does not influence cost	Minor Influence	Average Influence	Major Influence
Urban or rural project				
Construction on Native American Reservations				
Context sensitive design issues, controversy - level of environmental documentation				
Design AADT				
Design speed(s)				
Site topography (steep, flat or undulating terrain)				
Start and end stations, length and width				
Existing surface conditions and depths				
Number of intersections in project				
Number of bridges requiring work/reconstruction				
Intersection signalization and signage				
Letting date				
Horizontal and vertical alignment				
Extent of changes to the existing intersections				
Typical section (depths of surfacing and aggregate)				
Curb & Gutter and Sidewalk				
Bridge type (steel or concrete) and complexity				
Volumes of excavation and embankment				
Geotechnical - subsurface & slope recommendations				
Bridge deck area				
Traffic Control - closures or detours				
Environmental permitting requirements - wetlands				
Hydraulic recommendations and culverts				
Storm Sewer extents				
Bridge span lengths (between supports)				
Foundation complexity of the bridge				
Right-of-way acquisition and costs				
Extent of utility relocations and costs				
Contract time				

Key to analyze the survey results:

Question 1) When do you typically compute or identify this variable in the preconstruction stages?						
Answer:	Nomination	PFR	A&G	SOW	PIH	Final Plans
Scale:	1	2	3	4	6	7

Question 2) Rate the typical effort required to compute or identify each variable:			
Rating:	L = Low effort, information available, desktop study	M = Medium time and effort	H = High effort involved. Possibly site visits, site investigations and approximations.
Scale:	1	2	3

Question 3) If required, what is the first stage that you could roughly compute or identify this variable?						
Answer:	Nomination	PFR	A&G	SOW	PIH	Final Plans
Scale:	1	2	3	4	6	7

Question 4) Rate the additional effort required to identify or compute this cost influencer at an earlier stage			
Rating:	L = Little extra effort	M = Average additional effort and time	H = Lots of extra effort and time
Scale:	1	2	3

Question 5) How influential do you believe this variable is on construction cost:				
Answer:	Does not influence cost	Minor influence	Average influence	Major influence
Scale:	1	2	3	4

Results using the key from above:

			1) When do you typically compute or identify this variable in the 5 / preconstruction stages?																												
Response ID	Role	Location	Urban or rural project	Construction on Native American Reservations	Context sensitive design issues, controversy - level of environmental documentation	Design AADT	Design speed(s)	Site topography (steep, flat or undulating terrain)	Start and End Stations, Length and Width	Existing surfacing conditions and depths	Number of intersections in project	Number of bridges in the project scope	Intersection signalization and signage	Letting Date	Horizontal and Vertical Alignment	Extent of changes to the existing intersections	Typical Section (depths of surfacing and aggregate)	Curb & Gutter and Sidewalk	Bridge type (steel or concrete) and complexity	Volumes of excavation and embankment	Geotechnical - subsurface & slope recommendations	Bridge deck area	Traffic Control - closures or detours	Environmental permitting requirements- wetlands	Hydraulic recommendations and culverts	Storm Sewer extents	Bridge span lengths (between supports)	Foundation complexity of the bridge	Right-of-way acquistion and costs	Extent of Utility relocations and costs	Contract Time
R_1eFy4CJrrMm3JqK			1	1	2	2	2	1	2	2	2	1	2	2	3	3	3	3	4	3	5	4	5	4	4	4	5	5	6	6	6
R_1DU5jJgtUbYMokT	Civil Engineering Specialist	Road Design	1	1	3	2	2	2	2	2	2	2	4	2	3	4	3	3	2	3	4	3	5	5	5	5	4	4	5	6	6
R_dg4ugYvLdQLZljb			2	2	2	2	2	2	3	2	2	2	2	2	3	3	3	3	3	3	3	5	5	4	3	4	3	4	5	5	6
R_1jfhj5wFrHvsyfF	Design Superviser	Highways/Road Design	2	2	4	2	2	2	3	4	2	2	2	4	3	4	3	3	4	3	4	4	4	4	4	5	4	4	5	5	6
R_1n1qlsafURopcxX	Project Design Manager - Butte District, Helena R	Highways Bureau/Engineering Division	1	1	2	3	2	2	2	2	1	1	2	1	3	4	3	4	4	3	4	3	5	3	4	5	3	3	6	6	6
R_3fDBxpI87M2jdqE			1	1	2	1	1	1	3	2	2	2	4	2	3	4	4	4	2	4	4	3	4	5	4	4	3	3	5	5	5
R_2zIazKl1jtMkHLI	Highways Engineer	Engineering/ Highways Bureau	1	1	2	1	1	1	2	2	2	1	3	3	2	4	3	2	4	3	3	1	3	2	4	4	5	5	5	3	5
R_2WIT1OSEFtocCBd	District Projects Engineer	Billings	1	1	4	2	2	1	1	1	1	1	4	1	3	4	4	4	4	5	4	5	4	4	4	4	4	5	5	5	5
R_bEfPyQKuDzHbbCZ			1	1	2	3	3	2	2	2	2	3	3	1	3	3	3	3	4	4	4	5	5	5	5	5	6	6	5	4	4
R_2YgHJsn5MvjQLY0	Road Design Supervisor	Highways	1	1	3	3	2	2	3	2	2	2	3	6	3	5	3	4	4	6	4	5	5	5	3	3	5	5	6	6	6
R_1g53A4uQG6YNjEd	Project Design Manager	Road Design	1	1	2	2	2	2	2	2	2	3	5	2	3	4	4	5	6	5	5	5	5	4	4	4	5	6	5	5	6
R_2zzHL8ADHeSyOQa	Design Supervisor	Missoula District	1	1	2	2	2	2	1	1	4	2	5	2	3	3	5	3	2	5	5	2	6	2	5	5	3	5	5	5	5
R_3Mlo4lu2WdLyI7x	CE Specialist IV	Highways Preconstruction	2	2	2	2	2	2	2	3	2	2	3	2	3	3	3	3	2	3	4	2	4	3	3	3	2	3	5	5	5
R_2wboZURMDofjOSB	Project Design Manager - GF District - Hlna	Road Design	1	1	3	2	2	2	1	2	2	2	2	1	3	5	3	5	4	5	4	4	5	5	5	5	5	5	6	6	6
R_YXicvrJ6nfdaSJz			2	2	2	2	2	2	3	2	2	4	4	3	3	3	3	3	4	3	4	3	5	5	3	3	4	4	6	5	6
R_Umyt4KDgJsM7Bpn	District Projects Engineer	Engineering	1	1	4	2	2	2	2	4	2	2	5	5	3	4	3	3	4	3	4	5	5	4	4	4	4	4	5	5	6
R_yEnCBh1sWKRS1MZ	Projects Engineer	Great Falls	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4	4	5	4	4	4	6	6	5
R_8B8kZEX8gknB4Pf	District Design Supervisor	Road Design	1	1	3	2	2	2	3	2	2	2	3	2	3	4	3	3	4	3	4	4	5	4	5	4	4	5	6	6	6
R_2tmqPgcYyXVtDBP	Project Design Engineer	Highways Bureau/Engineering	1	1	2	2	2	2	2	2	2	2	2	1	3	3	3	2	4	3	4	4	4	5	4	4	4	4	6	6	6
R_2uqQcZdnhzKODL0	Area Engineer	Bridge	1	1	4	2	2	2	3	3	3	2	5	1	3	4	3	2	3	3	5	3	3	4	4	4	3	5	5	6	6
R_2Eyt8bbkBpvebRB	District Preconstruction Engineer	Glendive District	1	1	2	2	2	2	1	2	2	2	4	2	3	2	3	2	4	3	3	4	5	5	3	3	5	5	5	5	5
R_1176Ah6vPzzTrWN	CE Spec IV	Highways/Preconstruction	2	1	2	4	2	2	3	2	1	2	4	1	3	5	2	4	1	5	3	5	5	2	3	3	2	5	5	5	6
R_2y3qTjT7q0ZCuRz	Bridge Area Engineer	Bridge/Engineering	1	1	2	3	3	2	3	3	2	1	5	2	3	3	3	3	4	3	4	1	2	2	4	4	4	4	5	5	4
R_8cCTmXt4zzGbHkF	Project Engineer	Consultant Design	1	1	2	3	2	2	1	3	2	2	5	2	3	5	3	3	3	5	3	3	4	4	5	5	3	5	6	5	6
R_3JmUpphTAMkMyOR	Missoula Dist Preconstruction Engineer	Missoula	1	1	5	2	2	2	2	3	2	1	4	4	3	4	3	2	3	3	4	3	4	4	4	4	4	4	1	1	6
R_2WGoJ6hIpTNOOWj	Project Fatilitation Specialist	Consultant Design	1	1	2	2	2	3	1	3	1	1	2	3	3	2	2	3	2	3	3	5	5	5	4	3	5	4		5	5
R_2wuClEWuMtVdyWD	Civil Engineer	Highways/Engineering	1	1	3	2	2	2	2	2	2	2	4	4	2	2	2	2	4	5	3	4	5	5	4	4	4	4	6	6	6
R_3KZzoTOR0GNcRrd	Project Design Engineer	Highways Bureau - Road Design	1	2	2	2	2	2	4	2	2	2	2	3	3	4	3	2	3	5	5	3	2	2	5	5	3	6	6	5	6
R_SI96EpxiXBc69Xz			2	2	2	2	3	2	3	3	2	2	3	3	3	4	3	3	3	5	5	3	5	4	4		3	5	5	5	6
R_24rqj7qqRRlMP4d	Butte DESS	Butte District	1	1	3	2	2	2	1	2	3	2	4	1	3	4	4	2	5	3	4	1	5	5	5	5	5	5	6	6	5
R_2diqSgVAgZaKy3u	District Projects Engineer	Missoula	1	1	2	2	2	2	1	2	2	1	3	2	3	3	3	2	3	3	3	3	2	2	3	3	4	4	3	4	5

2) Rate the typical effort required to compute or identify each / variable																													3) If required, what is the first stage that you could roughly compute or identify this variable?																												
Urban or rural project	Construction on Native American Reservations	Context sensitive design issues, controversy - level of environmental documentation	Design AADT	Design speed(s)	Site topography (steep, flat or undulating terrain)	Start and End Stations, Length and Width	Existing surfacing conditions and depths	Number of intersections in project	Number of bridges in the project scope	Intersection signalization and signage	Letting Date	Horizontal and Vertical Alignment	Extent of changes to the existing intersections	Typical Section (depths of surfacing and aggregate)	Curb & Gutter and Sidewalk	Bridge type (steel or concrete) and complexity	Volumes of excavation and embankment	Geotechnical - subsurface & slope recommendations	Bridge deck area	Traffic Control - closures or detours	Environmental permitting requirements- wetlands	Hydraulic recommendations and culverts	Storm Sewer extents	Bridge span lengths (between supports)	Foundation complexity of the bridge	Rght-of-way acquisition and costs	Extent of Utility relocations and costs	Contract Time	Urban or rural project	Construction on Native American Reservations	Context sensitive design issues, controversy - level of environmental documentation	Design AADT	Design speed(s)	Site topography (steep, flat or undulating terrain)	Start and End Stations, Length and Width	Existing surfacing conditions and depths	Number of intersections in project	Number of bridges in the project scope	Intersection signalization and signage	Letting Date	Horizontal and Vertical Alignment	Extent of changes to the existing intersections	Typical Section (depths of surfacing and aggregate)	Curb & Gutter and Sidewalk	Bridge type (steel or concrete) and complexity	Volumes of excavation and embankment	Geotechnical - subsurface & slope recommendations	Bridge deck area	Traffic Control - closures or detours	Environmental permitting requirements- wetlands	Hydraulic recommendations and culverts	Storm Sewer extents	Bridge span lengths (between supports)	Foundation complexity of the bridge	Rght-of-way acquisition and costs	Extent of Utility relocations and costs	Contract Time
1	1	1	2	1	1	2	3	2	2	3	1	2	3	2	2	3	2	2	2	2	1	2	2	3	3	3	3	2	1	1	3	2	2	1	2	2	3	1	3	2	3	3	3	3	3	3	4	4	4	4	4	4	4	4	6	6	4
1	1	2	1	2	1	2	2	1	2	2	1	3	1	2	2	2	3	2	2	1	2	3	2	2	3	2	2	1	1	1	1	2	1	1	2	2	2	2	3	2	3	3	2	3	3	4	4	4	4	3	4	4	3	4	5	6	4
3	3	3	1	1	1	2	1	1	2	2	1	2	2	2	2	2	2	2	1	1	1	2	2	1	2	3	3	1	1	2	2	2	2	2	3	2	2	2	3	2	3	3	3	3	3	3	4	3	5	5	3	3	3	4	5	5	6
1	1	3	1	2	2	2	1	1	1	1	1	3	3	1	3	1	3	2	1	3	2	2	2	1	1	3	3	2	1	1	3	1	1	2	3	2	1	2	2	2	2	2	2	3	2	3	4	2	3	2	3	3	2	3	3	3	4
1	1	2	2	1	1	2	2	1	3	2	1	3	3	3	2	3	2	3	2	2	2	3	3	2	3	3	3	1	1	1	2	1	1	1	2	1	1	1	3	1	3	3	3	4	3	3	3	1	2	2	3	3	3	3	4	4	4
1	1	2	1	1	1	2	1	1	3	2	1	2	2	2	2	2	2	3	2	3	3	2	2	3	3	3	3	2	1	1	3	1	1	1	2	2	1	3	4	1	2	3	3	4	3	3	3	3	3	4	3	3	3	3	4	4	3
1	1	1	2	1	1	1	1	2	1	3	1	3	2	2	1	2	3	3	1	1	2	3	3	3	3	3	2	2	1	1	1	1	1	1	1	1	1	1	1	2	1	3	2	1	3	2	3	1	2	1	3	3	3	4	3	2	3
1	1	1	1	1	1	1	2	1	1	2	1	3	3	2	3	2	3	2	2	2	2	2	2	2	2	3	3	2	1	1	2	2	2	1	2	2	2	2	4	2	3	4	4	4	4	5	4	4	4	4	4	4	4	4	5	5	5
1	1	2	1	1	1	2	2	2	2	2	2	3	3	2	2	2	3	3	2	2	3	3	2	3	3	2	2	2	1	1	3	2	2	2	1	3	2	2	3	4	3	3	3	3	3	3	3	4	3	5	5	4	6	6	5	5	5
1	1	2	1	1	1	2	2	1	1	1	1	3	2	2	3	2	3	2	1	2	3	3	3	2	2	3	3	2	1	1	2	2	2	1	2	2	2	2	2	5	5	5	3	5	4	5	5	4	4	5	5	5	5	5	5	6	6
1	1	1	1	1	2	2	1	1	1	2	2	3	3	2	2	2	3	3	2	2	2	3	3	3	3	3	3	2	1	1	2	2	2	2	1	2	2	1	2	2	3	3	2	2		2	2	2	2	3	3	3	3	3	2	2	3
1	1	2	1	1	1	1	1	1	2	2	1	3	3	2	2	3	3	3	2	2	2	3	2	3	3	3	3	1	1	1	2	1	1	1	1	1	1	2	4	4	3	3	3	3	4	3	4	4	4	4	4	4	4	4	5	5	4
1	1	1	2	1	1	1	2	1	2	2	1	2	2	2	2	2	1	2	1	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	2	3	3	3	3	2	3	3	2	4	4	3	3	2	2	5	5	4
1	1	2	1	1	1	1	2	1	3	2	1	3	3	2	2	2	2	3	1	2	3	3	3	2	2	3	3	1	1	1	2	1	1	1	1	1	1	1	2	1	3	4	3	4	4	4	4	4	4	4	4	4	4	4	5	5	5
1	2	2	1	1	1	2	1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	3	2	2	4	5	2	2	3	2	3	4	2	4	3	5	5	4	4	4	4	6	5	5
1	1	2	2	1	1	2	3	1	2	2	2	2	3	2	2	2	2	3	2	2	3	3	3	2	2	2	2	2	1	1	3	2	1	1	1	3	1	2	3	3	3	3	3	3	4	3	4	3	4	3	4	4	4	4	5	5	5
1	1	1	1	1	1	2	1	1	1	1	1	3	2	2	3	2	2	3	2	2	2	2	2	2	2	3	3	1	1	1	2	2	2	1	2	2	2	2	2	3	3	3	2	2	3	3	3	4	3	3	4	4	3	4	5	5	3
1	1	1	1	1	1	2	2	1	1	2	1	2	3	2	2	3	2	2	2	1	1	2	2	3	3	3	3	1	1	1	1	1	1	1	3	2	2	2	3	2	3	3	3	3	4	3	4	4	5	3	3	3	4	5	6	6	6
1	1	2	1	1	1	1	1	1	1	1	1	2	2	2	2	1	2	2	1	2	2	2	2	2	2	3	3	1	1	1	3	2	2	2	2	2	2	2	2	2	3	3	2	3	2	3	3	3	3	3	3	3	3	3	6	6	5
1	1	2	1	1	1	2	2	1	2	2	1	3	2	2	2	2	2	3	2	1	2	3	2	2	3	2	3	1	1	1	2	2	2	2	2	2	2	2	2	2	3	2	2	2	3	3	3	3	2	3	3	3	3	3	3	3	5
1	1	2	1	1	2	2	3	2	1	2	1	3	2	2	2	1	1	3	1	2	3	3	2	1	3	1	3	1	1	1	2	1	1	1	1	1	1	1	1	1	2	2	2	1	2	3	3	2	2	3	2	2	3	3	3	3	3
1	1	2	2	2	3	3	3	2	1	1	1	3	3	3	3	1	3	3	1	3	1	2	2	1	1	3	3	2	1	1	2	2	2	1	1	2	1	1	1	1	3	3	1	1	1	3	3	2	4	2	4	4	3	3	5	5	6
1	1	2	1	1	3	3	3	2	3	3	2	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3	2	4	4	3	2	2	2	2	2	2	1	2	2	3	3	3	3	3	3	3	1	2	2	3	3	3	3	4	4	4
1	1	2	1	1	1	2	3	1	1	2	2	2	3	2	2	1	2	3	1	2	1	2	2	1	2	2	2	2	1	1	2	1	1	1	2	3	1	1	4	2	3	3	3	4	3	3	3	2	4	5	5	5	4	4	5	5	5
2	2	2	1	1	1	1	1	1	2	2	1	2	2	2	1	2	1	2	1	2	2	2	2	2	2	2	2	1	1	1	2	1	1	1	1	1	1	1	4	3	3	2	3	1	1	3	3	1	1	1	1	1	3	3	1	1	1
1	1	2	2	2	3	1		1	1	1	1	3	2	2	2	2	2	3	1	2	2	3	3	3	3	3	3	2	1	1	2	2	2		1	3		1	1		3	3	3		1	3	3	3	3	2	3	4	4	3	4	4	4
1	1	2	1	1	1	1	2	1	1	2	2	3	2	2	1	2	3	3	2	2	2	3	2	2	3	3	3	2	1	1	2	1	1	2	2	1	1	1	2	5	2	2	1	1	1	6	2	2	5	6	4	3	3	3	6	6	6
1	1	2	2	1	1	1	2	1	1	1	1	3	3	2	2	2	2	3	1	2	1	3	3	2	3	3	3	2	1	1	2	2	2	1	1	2	1	1	1	3	2	2	2	1	2	3	3	2	2	2	2	3	3	3	6	3	3
1	1	2	1	2	1	2	2		2		2	3		2		3	3	3	2	2	2	3		3	3	3	2	2	1	1	2	2	2	2	3	3	2	2	3	3	3	3	3	3	3	3	4	3	3	3	4	4	3	3	3	3	4
1	1	3	2	1	1	1	1	1	2	3	2	3	3	2	2	3	3	3	1	3	3	3	3	3	3	3	3	2	1	1	3	2	2	2	1	1	1	2	2	1	3	3	3	2	4	3	4	3	4	3	4	4	4	4	5	5	5
1	1	3	1	1	1	2	3	1	1	2	1	3	3	2	3	3	3	3	2	2	2	3	3	3	3	3	3	1	1	1	3	1	1	1	1	2	1	1	3	2	3	3	2	3	2	3	3	3	3	2	3	3	3	3	4	4	3

4) Rate the additional effort required to identify or compute this cost influencer at an earlier																													5) How influential do you believe this variable is on construction / cost? / / Note: For this questi...																												
Urban or rural project	Construction on Native American Reservations	Context sensitive design issues, controversy - level of environmental documentation	Design AADT	Design speed(s)	Site topography (steep, flat or undulating terrain)	Start and End Stations, Length and Width	Existing surfacing conditions and depths	Number of intersections in project	Number of bridges in the project scope	Intersection signalization and signage	Letting Date	Horizontal and Vertical Alignment	Extent of changes to the existing intersections	Typical Section (depths of surfacing and aggregate)	Curb & Gutter and Sidewalk	Bridge type (steel or concrete) and complexity	Volumes of excavation and embankment	Geotechnical - subsurface & slope recommendations	Bridge deck area	Traffic Control - closures or detours	Environmental permitting requirements- wetlands	Hydraulic recommendations and culverts	Storm Sewer extents	Bridge span lengths (between supports)	Foundation complexity of the bridge	Right-of-way acquisition and costs	Extent of Utility relocations and costs	Contract Time	Urban or rural project	Construction on Native American Reservations	Context sensitive design issues, controversy - level of environmental documentation	Design AADT	Design speed(s)	Site topography (steep, flat or undulating terrain)	Start and End Stations, Length and Width	Existing surfacing conditions and depths	Number of intersections in project	Number of bridges in the project scope	Intersection signalization and signage	Letting Date	Horizontal and Vertical Alignment	Extent of changes to the existing intersections	Typical Section (depths of surfacing and aggregate)	Curb & Gutter and Sidewalk	Bridge type (steel or concrete) and complexity	Volumes of excavation and embankment	Geotechnical - subsurface & slope recommendations	Bridge deck area	Traffic Control - closures or detours	Environmental permitting requirements- wetlands	Hydraulic recommendations and culverts	Storm Sewer extents	Bridge span lengths (between supports)	Foundation complexity of the bridge	Right-of-way acquisition and costs	Extent of Utility relocations and costs	Contract Time
1	1	2	1	1	1	3	3	2	3	3	1	2	3	2	2	3	2	3	3	3	2	3	3	3	3	3	3	2	2	2	2	2	2	2	2	2	4	4	4	2	4	4	3	3	4	3	3	3	3	2	3	4	4	4	4	3	1
1	1	1	1	1	1	1	1	1	2	2	1	1	2	1	1	2	2	2	1	2	1	2	1	2	2	3	2	1	3	2	2	4	2	4	4	3	3	4	3	3	4	3	4	3	4	4	3	4	3	2	3	3	4	4	4	4	3
1	1	1	1	1	1	2	1	2	2	2	1	2	2	2	1	2	3	3	1	2	2	2	2	2	2	2	2	1	3	3	3	2	1	2	1	2	3	3	3	2	3	3	3	3	4	4	4	3	3	3	3	3	4	4	4	3	2
1	1	3	1	2	1	3	2	1	1	2	3	3	3	2	3	1	3	3	1	3	3	3	3	1	2	3	3	3	4	3	4	3	3	3	3	3	3	4	3	2	4	3	3	3	4	3	4	2	4	3	2	3	2	3	4	4	2
1	1	2	1	1	1	2	2	1	2	3	1	2	3	2	3	2	3	2	2	2	2	2	2	2	3	2	2	1	4	3	3	3	4	3	4	3	4	4	2	2	3	4	4	2	3	4	4	4	3	2	3	3	3	3	4	3	2
2	2	2	1	1	1	2	2	2	2	2	2	3	3	2	3	2	3	2	2	3	3	3	3	2	2	2	2	2	3	4	4	2	2	3	3	3	3	4	2	2	3	4	3	3	4	4	3	4	4	3	3	3	3	4	4	4	3
1	1	2	3	1	1	2	2	3	1	2	1	3	3	2	2	3	3	3	1	2	2	3	3	3	3	3	2	3	4	3	3	3	3	4	3	2	3	3	2	3	4	3	4	3	4	4	3	4	2	2	3	3	4	3	3	2	3
1	1	1	1	1	1	1	1	1	1	2	1	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	1	4	3	3	2	2	2	3	2	3	3	3	2	3	3	3	3	3	4	3	3	3	2	2	2	4	3	3	3	2
1	1	1	1	1	1	1	2	1	1	2	2	2	2	2	2	2	3	2	2	1	2	2	2	2	2	2	2	2	4	2	2	3	3	3	1	3	2		2	2	3	2	3	3	4	3	3	3	2	3	3	3	4	4	4	4	3
1	1	2	1	1	1	2	2	1	1	1	1	2	1	2	2	1	2	2	1	2	3	3	3	3	3	3	3	2	4	2	2	3	3	4	3	1	3	2	2	4	4	3	3	3	3	4	2	3	2	3	3	3	3	3	4	3	1
1	1	2	1	1	2	2	2	1	1	2	2	3	2	2	2	2	3	3	2	2	2	2	2	3	3	2	2	2	4	3	4	4	3	4	3	4	3	4	3	3	3	4	4	3	4	4	4	4	3	2	3	3	4	4	3	3	2
1	1	1	1	1	1	1	1	1	1	2	1	3	3	2	2	3	3	3	2	2	2	3	2	3	3	3	3	2	4	3	4	4	4	4	4	3	2	4	2	2	4	3	4	3	4	4	4	2	3	4	2	3	4	4	3	3	3
1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	1	2	2	1	2	2	2	2	1	1	3	3	2	2	2	2	2	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	1	2	2	2	2	1	2	2	2	1
1	1	2	1	1	1	1	1	1	1	1	1	3	3	1	3	2	3	3	2	2	3	3	2	2	2	3	3	1	4	3	3	4	4	4	2	3	3	3	2	4	4	2	4	2	3	4	4	3	3	3	2	2	3	4	2	4	4
1	1	2	1	1	1	2	1	1	2	2	2	2	2	2	2	3	2	2	2	3	3	2	2	3	3	3	3	1	3	3	4	4	4	4	4	4	3	4	3	3	4	4	3	2	4	4	4	3	3	3	3	3	4	4	3	3	2
1	1	1	1	1	1	1	2	1	2	2	1	2	2	2	2	2	2	3	2	1	2	3	2	3	3	2	2	1	4	4	3	3	3	4	3	3	2	4	3	3	3	4	3	3	3	4	3	3	3	3	3	3	4	4	3	4	3
1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	2	1	1	1	2	2	1	3	2	2	1	4	2	2	2	3	2	4	2	2	4	3	2	4	3	4	3	2	4	3	3	2	2	3	3	2	3	4	4	2
1	1	2	1	1	1	3	2	2	2	3	1	3	3	2	2	3	2	3	2	1	2	2	2	3	3	3	3	1	2	2	2	1	1	1	3	2	4	4	4	2	3	4	3	3	3	3	3	2	2	2	3	4	3	4	4	4	2
1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	4	4	4	3	3	4	3	3	3	4	4	3	4		4	4	4	4	3	3	4	3	4	4	4	4	3	3	2
1	1	2	1	1	1	1	2	1	2	2	1	3	2	2	1	2	2	2	2	1	3	3	2	2	2	2	2	2	2	2	3	2	2	2	4	2	3	4	2	2	4	2	3	2	4	4	4	4	3	2	3	3	3	3	3	2	2
1	1	2	1	1	1	1	1	1	1	2	2	3	3	2	3	2	3	2	2	2	2	2	2	2	2	2	3	2	4	2	3	2	3	3	4	4	2	4	2	2	3	3	4	4	4	4	3	3	2	1	3	3	3	3	1	1	3
1	1	2	3	2	1	3	3	1	1	2	1	3	3	3	2	1	3	2	1	2	2	2	3	2	1	2	2	2	4	2	4	4	2	4	3	2	2	3	3	2	4	3	2	3	3	3	3	2	3	3	3	3	3	2	3	4	2
3	3	3	1	1	2	2	1	1	1	2	1	2	3	2	2	3	3	3	2	2	2	2	2	3	3	3	3	2	4	4	4	2	2	4	3	2	3	4	2	3	4	3	2	2	4	3	4	4	3	2	3	2	4	4	4	4	4
1	1	2	1	1	1	2	2	1	1	2	2	2	3	2	2	2	2	3	1	2	2	2	2	1	2	2	2	2	4	2	4	2	1	3	2	3	4	4	4	2	3	4	3	3	4	3	4	2	4	3	3	3	3	3	4	4	2
1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	4	4	3	2	4	3	2	3	4	3	2	2	3	2	4	2	3	3	3	2	2	2	2	3	4	4	3	3
2	3	2	2	1	2	2	2	2	3	3	1	2	2	1	2	2	2	3	1	2	3	3	3	3	3	3	2	1	4	3	3	4	4	4	2	2	3	4	4	2	3	3	2	4	4	4	3	3	2	2	3	4	3	4	3	3	4
1	1	2	1	1	1	2	2	1	3	3	3	3	3	2	1	1	3	3	1	2	3	3	2	2	3	3	3	3	2	2	2	2	3	4	4	4	4	4	4	1	4	4	4	4	4	4	4	2	2	2	4	2	4	4	4	4	1
1	1	1	1	1	1	2	2	1	1	2	1	2	3	2	2	2	3	3	2	2	2	3	3	2	3	3	3	2	2	2	3	2		3	2	2	3	4	3	2	4	3	3	3	3	3	4	3	4	2	3	3	2	3	3	3	3
1	1	2	1	1	1	2	2	1	1	2	1	3	2	2	3	3	3	3	2	2	2	3	2	2	3	3	2	3	4	4	3	2	2	3	2	3	3	4	3	2	3	3	2	2	4	4	3	3	3	3	4	3	3	4	4	3	3
1	1	3	1	1	1	1	1	1	2	3	2	2	2	1	2	3	3	3	2	2	2	2	3	3	3	3	3	1	4	3	3	3	3	4	4	2	3	4	4	3	4	4	4	4	4	4	4	4	4	3	3	4	4	4	4	4	4
1	1	3	1	1	1	2	3	1	1	3	2	3	3	2	2	2	3	3	2	2	2	3	3	3	3	3	3	2	4	2	4	3	4	4	4	3	3	4	3	2	4	3	4	3	4	3	4	4	2	3	3	3	4	4	4	3	2