TOP-DOWN CONSTRUCTION COST ESTIMATING MODEL USING AN ARTIFICIAL NEURAL NETWORK

prepared for
THE STATE OF MONTANA
DEPARTMENT OF TRANSPORTATION

in cooperation with
THE U.S. DEPARTMENT OF TRANSPORTATION
FEDERAL HIGHWAY ADMINISTRATION

July 2017

prepared by
Douglas D. Gransberg
H. David Jeong
Ilker Karaca
Brendon Gardner

Iowa State University
Institute for Transportation
Ames, IA

RESEARCH PROGRAMS

MDT★

Top-down Construction Cost Estimating Model Using an Artificial Neural Network

Final Report

Prepared by: Douglas D. Gransberg, PhD, PE

H. David Jeong, PhD

Ilker Karaca

Brendon Gardner

IOWA STATE UNIVERSITY
Institute for Transportation

Technical Report Documentation Page

1. Report No. FHWA/MT-17-007/8227-001	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Top-down Construction Cost Estimating Model/Guide Using a Neural Network		5. Report Date July 2017	
		6. Performing Organization Code 4041762	
7. Author(s) Douglas D. Gransberg, PhD, PE., H. David Jeong, PhD, Ilker Karaca, Brendon Gardner		8. Performing Organization Report No.	
9. Performing Organization Name and Address Institute for Transportation Construction Management and Technology Program Iowa State University 2711 South Loop Drive, Suite 4700 Ames, IA 50010-8664		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No. 8227-001	
12. Sponsoring Organization Name and Address Research Programs Montana Department of Transportation (SPR) http://dx.doi.org/10.13039/100009209 2701 Prospect Avenue PO Box 201001 Helena MT 59620-1001		13. Type of Report and Period Covered Final Report December 2014 to July 2017t	
		14. Sponsoring Agency Code 5401	
15. Supplementary Notes Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration. This report can be found at http://www.mdt.mt.gov/research/projects/const/cost_est_neural_network.shtml.			
16. Abstract This report contains the information and background on top-down cost estimating using artificial neural networks (ANN)_to enhance the accuracy of MDT early estimates of construction costs. Upon conducting an extensive review of MDT's budgeting and cost estimating efforts, and following a survey of agency experts on the identification of the most salient project attributes with the dual-objectives of low effort and high accuracy, a rational method for top-down variable selection is proposed. Selected variables were further tested in their explanatory power of construction costs through the application of two cost estimating methodologies—multiple regression and artificial neural network methodologies. Both methods are shown to provide sizeable improvements over the agency's current levels of prediction accuracy for its construction costs. Potential accuracy gains are also demonstrated to depend on project work types. The comparison of mean absolute percentage errors across different estimating methods confirms that the potential benefits from the proposed methodologies are expected to rise as the project level complexity and uncertainty increase. New construction and bridge replacement projects, for instance, are expected to gain the most in estimating accuracy since these two groups seem to exhibit considerably higher levels of deviation from the MDT's preliminary cost estimates. To facilitate MDT's implementation of the suggested methodology described in this report, a cost estimation methodology was also presented in an Excel spreadsheet format. This achieves two goals. First, it provides an accessible tool to make top-down cost predictions for agency planners during the budgeting stage based on MDT's historical project data. Second, it furnishes a process through which the proposed model can be improved as new project information becomes available. Ultimately, the insights gained from this study are expected to contribute to a better formulation of the agency's early cost estimation and budgeting efforts.			
17. Key Words Construction, Cost estimating, Conceptual Estimates, Artificial Neural Network, Multiple Regression Analysis, Statistical Analysis		18. Distribution Statement No restrictions.	
19. Security Classification (of this report) Unclassified.	20. Security Classification (of this page) Unclassified.	21. No. of Pages 101	22. Price NA

IOWA STATE UNIVERSITY
Institute for Transportation

Acknowledgements

Disclaimer Statement

IOWA STATE UNIVERSITY
Institute for Transportation

TABLE OF CONTENTS

IOWA STATE UNIVERSITY
Institute for Transportation

TABLE OF FIGURES

TABLE OF TABLES

IOWA STATE UNIVERSITY
Institute for Transportation

EXECUTIVE SUMMARY

A better understanding of top-down estimating practices, and the resulting increases in the accuracy of budgeting efforts, may have significant contributions to public transportation agencies in their efforts to allocate agency funds more efficiently. This report thus provides an analysis and evaluation of top-down estimating methodologies to assist MDT in its early estimation of its construction costs. In so doing, the research effort applies an artificial neural network methodology, as well as a multiple regression estimation model, to compare prediction accuracy of proposed estimating approaches to those achieved under MDT's current practices. Four separate estimation equations are provided to predict agency costs under three broad project work types. Together these groups of work account for more than 80 percent of the agency's construction program.

Due to the critical nature of input selection for the cost estimation methodologies, the study allocated considerable effort to the proper identification of project variables that are often readily available at the early stages of an agency project. Upon conducting an extensive review of MDT's budgeting and cost estimating efforts, and following a survey of the agency experts on the identification of the most salient project attributes with the dual-objectives of low effort and high accuracy, the team was able to propose a rational method for top-down variable selection.

Selected variables were further tested in their explanatory power of construction costs through the application of two cost estimating methodologies—multiple regression and artificial neural network methodologies. Both methods are shown to provide sizeable improvements over the agency's current levels of prediction accuracy for its construction costs. Potential accuracy gains are also demonstrated to depend on project work types. The comparison of mean absolute percentage errors across different estimating methods confirms that the potential benefits from the proposed methodologies are expected to rise as the project level complexity and uncertainty increase. New construction and bridge replacement projects, for instance, are expected to gain the most in estimating accuracy since these two groups seem to exhibit considerably higher levels of deviation from the MDT's preliminary cost estimates.

To facilitate MDT's implementation of the suggested methodology described in this report, a cost estimation methodology was also presented in an Excel spreadsheet format. This achieves two goals. First, it provides an accessible tool to make top-down cost predictions for agency planners during the budgeting stage based on MDT's historical project data. Second, it furnishes a process through which the proposed model can be improved as new project information becomes available. Ultimately, the insights gained from this study are expected to contribute to a better formulation of the agency's early cost estimation and budgeting efforts.

Finally, the report outlines potential research areas for future work. The integration of early project-level data with actual construction costs and tailoring MDT systems with early estimating in mind remain logical next steps to fully attain the efficiencies suggested through the analysis provided in this report.

IOWA STATE UNIVERSITY
Institute for Transportation

CHAPTER 1. INTRODUCTION

Background Summary

The issue of accurate estimating is essentially tied to the efficient use of available public capital (Janacek 2006). Early estimates conducted during the planning phase often turn into project budgets before the final scope of project work is adequately quantified (Anderson et al. 2007; Alshanbari 2010). Additionally, since preconstruction costs are by definition a small portion of the total project delivery cost, they are typically estimated as a standard percentage of estimated construction costs. Hence, if the capital project is underestimated, preconstruction costs will be similarly underestimated. A 2002 study involving 258 transportation projects collectively valued at $90.0 billion (Flyvbjerg et al. 2002) found that 86% experienced actual costs that were on average 28% higher than initially estimated. That study concludes that "*underestimation of costs at the time of decision to build is the rule* rather than the exception for transportation infrastructure projects" (Flyvbjerg et al. 2002, italics added). Using Flyvbjerg's cost growth would mean that the agencies delivering these projects would be short $1.4 billion in the preconstruction phases of project development. The fact that project scope and quality is defined during the planning and design phases of the project development process means that an accurate estimate of construction costs will furnish sufficient funding for the early phase planning and design activities.

MDT Project FHWA/MT-08-007/8189 *Highway Project Cost Estimating and Management* furnished a bottom-up conceptual estimating procedure that appears to be risk-adjusted but utilizes extremely small sample populations. The study found that MDT sees a 46% growth in construction cost from programming to construction completion. Montana's small population and its huge area makes it imperative that MDT squeeze every last penny out of its federal and state highway funding to provide as much service as it can afford. So, reducing cost growth from the early estimate is a priority. To do so, requires that cost certainty be increased and that means better conceptual estimates.

NCHRP Report 574: *Guidance for Cost Estimation and Management for Highway Projects During Planning, Programming, and Preconstruction* expresses the motivation for this MDT research project in this manner:

> "Over the time span between the initiation of a project and the completion of construction, many factors influence a project's final costs. This time span is normally several years, but for highly complex and technologically challenging projects the time span can easily exceed a decade. Over that period, numerous changes to the project scope and schedule will occur." (Anderson et. al 2007).

Problem Statement

States with small populations and large amounts of highway lane miles to service must use every penny appropriated for design, construction and maintenance as wisely as possible. To do so, requires that early estimates of project costs are not overly inflated, potentially preventing precious federal-aid funding from being obligated on other projects. Worse yet, if the budget overage is found late in the fiscal year, the incremental overage can be lost to the state by FHWA year-end reapportioning (Anderson et al. 2006). The other side of the coin is a concept called "optimism bias" where engineers unintentionally underestimate project cost and keep the project "alive" by

making unrealistically optimistic assumptions in the project's estimate and schedule (Jennings 2012; Flyvbjerg et al. 2002).

The issue is exacerbated by the fact that early costs estimates made with the least amount of design detail often become the final project budget. The result is to create a bias either toward over-estimating actual costs to provide financial room for the scope to grow as the project development process proceeds (Jennings 2012) or toward underestimating costs because of misplaced optimism (Flyvbjerg et al. 2002). Research has shown that the unintended consequence of over estimating is that project managers will attempt to use all the funding available within a given project's authorization to avoid losing it rather than return the overage as soon as it is identified (Anderson et al. 2006). Thus, the efficient use of available capital is compromised (Janacek 2006). The solution is to develop a system where early estimates can be developed using rationally derived contingencies for the unknowns at the time of the estimate (Anderson et al. 2006).

This research seeks to leverage the work completed in NCHRP 15-51: *Guide for Estimating Preconstruction Services Costs*, by extending the parametric models developed for preliminary engineering to the estimating of construction project cost at the earliest stages of project planning and development. It will deliver a top-down cost estimating model that uses "stock" spreadsheet and database software without the need to purchase special software or hardware (i.e. MS Office products only).

Research Objectives

The final product will permit MDT's project managers to prepare cost estimates for design and construction of typical MDT capital improvement and rehabilitation projects. Hence, the technical objectives of the study are as follows:

- To develop a framework for building a database from MDT bid tabulations.
- To develop a parametric estimating model that can be fed from the database using a neural network to assemble CERs and produce top-down estimates at early stages of project development.
- To develop a stochastic cost modeling system that will assist MDT engineers in calculating rational contingencies for early estimates.

IOWA STATE UNIVERSITY
Institute for Transportation

CHAPTER 2. METHODOLOGY

This chapter firstly describes how the research team divided up the scope-of-works for each of the different models to be constructed. Secondly, it describes how the 'global database' for the cost estimating model was created through combining multiple data sources. Finally, the Chapter explains the methodology to create and use the artificial neural network for predicting construction costs.

Model Scope

MDT currently divides projects into the work-types shown in Table 1. It was determined that no single data-driven cost estimating model could encompass these work-types together due to the vast difference in scopes. Therefore, the research team analyzed work-types which MDT perform and the frequency of each.

Table 1. Break down of the work-types at MDT

Construction	**Spot Improvement**
110 New Construction	234 Bridge Protection
120 Relocation	310 Roadway and Roadside Safety Improvements
130 Reconstruction – with added capacity	311 Railroad/Highway Crossing Safety Improvements
140 Reconstruction – without added capacity	312 Structure Safety
141 Reconstruction – remove and replace culverts	**Miscellaneous**
150 Major Rehabilitation-with added capacity	313 Pedestrian and Bicycle Safety
151 Major Rehabilitation-without added capacity	410 Traffic Signals and Lighting
222 Bridge Replacement with a culvert with no added capacity	411 Signing, Pavement Markings, Chevrons, etc.
223 Bridge Replacement with a Culvert while adding capacity	412 Miscellaneous Electronic Monitoring or Information Services
Pavement Preservation	510 Environmental
160 Minor Rehabilitation	520 Landscaping, Beautification
170 Restoration and Rehab – PCCP	610 Maintenance Stockpiles
172 Restoration and Rehab - Facilities	620 Bicycle and Pedestrian Facilities
180 Resurfacing – Asphalt (thin lift<=60.00mm) (including safety improvements) (Pavement Preservation)	660 Historic Preservation
181 Resurfacing – Asphalt (thin lift<=60.00mm) (Scheduled Maintenance)	710 Pedestrian and Bicycle Facilities CTEP
182 Resurfacing – PCCP	**Facilities**
183 Resurfacing – Seal and Cover	111 New Construction – Facilities
184 Resurfacing – Gravel	
185 Resurfacing – Crack Sealing	
Bridge	
210 New Bridge	
220 Bridge Replacement with added capacity	
221 Bridge Replacement with no added capacity	
230 Bridge Rehabilitation with added capacity	
231 Major Bridge Rehabilitation without added capacity	
232 Minor Bridge Rehabilitation	
233 Bridge Preservation	

To suitably assign work-types to each of the models, basic statistical tools were used to analyze which work-types occur most frequently and which account for the most significant proportion of cost to MDT. Statistical analysis was based on 1,012 different projects provided to the research team for projects constructed between 2007 and 2014. Figure 1 shows an analysis of the major work-types broken down by cost and frequency. It is observed that resurfacing is the most common work-type (431 of 1,012 projects studied). However, by total expenditure, construction projects represent the greatest cost to MDT. Spot improvement projects are more common than bridge

IOWA STATE UNIVERSITY
Institute for Transportation

projects, however, the total cost of the spot improvement projects is significantly lower than that of bridge projects.
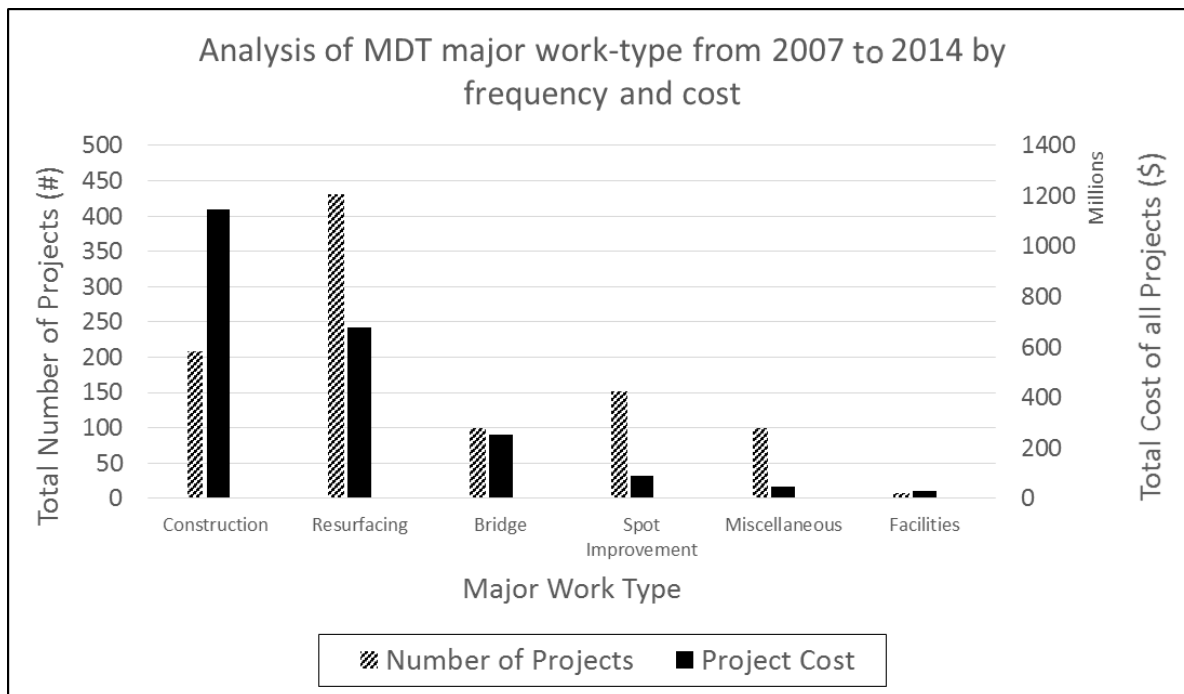
Figure 1. Analysis of MDT work-types by frequency and cost

Figure 2 shows that the total cost of spot improvement, miscellaneous and facility type projects is very small relative to the total expenditure. Based on the analysis of Figures 1 and 2, it was decided to initially trial three different estimating models only and not complete models for the other minor work types as these only account for a small proportion of costs. The three estimating models will be:

- Construction
- Resurfacing
- Bridge

These models will encompass most projects (by cost) at MDT.

IOWA STATE UNIVERSITY
Institute for Transportation
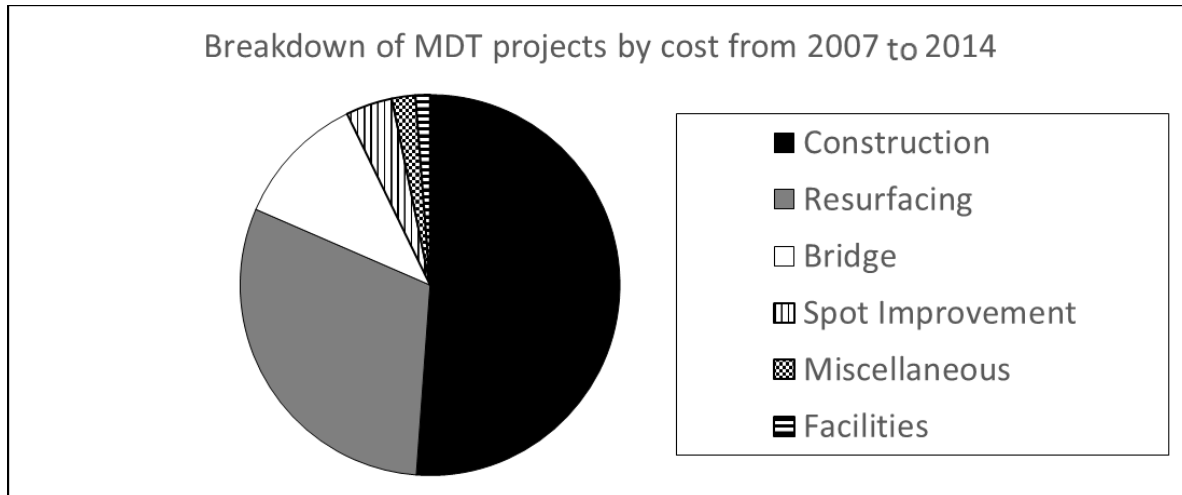
Figure 2. Breakdown of MDT work-types by cost of the projects

Although construction projects are the largest by cost at MDT, the research team commenced by building a model for pavement perseveration work-types only. Through meetings at MDT, it was determined that these pavement preservation projects would be the most predictable work-type.

Input Variables

Previous authors of publications involving data-driven artificial neural networks to predict construction costs have recognized the importance of selecting the correct input variables for the estimating model. From a review of relevant publications and interviews at MDT, 29 possible input variables have been identified. These are shown in Table 2.

IOWA STATE UNIVERSITY
Institute for Transportation

Table 2. Input variables that were recognized at MDT through interviews

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
10	Volumes of excavation and embankment	27	Letting Date
11	Geotechnical - subsurface and slope recommendations	28	Context sensitive design issues, controversy
12	Bridge deck area	29	Contract time
13	Hydraulic recommendations and culverts		
14	Storm drain 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		

Typically, when artificial neural network models are created then different combinations of the inputs to that model are included through trial-and-error. The combination of inputs which results in the lowest estimating error is then selected as the final inputs. To aid the selection of inputs for this project, the research team conducted a 'Cost Estimating Survey' at MDT to help understand each of the proposed input variables. The survey and results are briefly described in the following section. The final recommended input variables are described in Chapter 4.

Survey and results

A total of 31 preconstruction engineers at MDT answered five questions on the 29 potential input variables shown in Table 2. These 5 questions were:

1. When do you typically compute or identify this variable in the 5 preconstruction stages?
2. Rate the typical effort required to compute or identify this variable
3. If required, what is the first stage that you could roughly compute or identify this variable?

IOWA STATE UNIVERSITY
Institute for Transportation

4. Rate the additional effort required to identify or compute this cost influencer at an earlier stage
5. How influential do you believe this variable is on construction cost? (assume a major reconstruction or major rehabilitation)

The cost estimating survey was developed with assistance from MDT employees to ensure that the terminology and questions made sense to participants. The survey was distributed online with a link at MDT by Highways Bureau Chief to obtain the best possible response rate. Appendix C contains the full survey questions and summarized results. Two interesting findings, which were used by the research team to select the input variables, are described below.

Firstly, the research team recorded when MDT typically knows each of the 29 input variables (Question 1); and secondly, what is the earliest possible stage that MDT could know that input variable (Question 3). This insight is useful so that the estimating model, and hence the 'global database', can contain input variables that can be calculated or computed at the conceptual estimating stage. Figure 3 summarizes the average responses to Questions 1 and 3 from the survey. It can be observed in Figure 3 that the first 10 of the 29 inputs are known prior to, or shortly after, the preliminary field review (PFR) cost estimating stage. As an example, respondents generally perceived that once a project is nominated, the urban/rural indicator is known immediately, so too is the project's location with respect to a Native American reservation. This makes sense because these inputs are determined as soon as a project has been selected and the general proximity recognized.

It is apparent in Figure 3 that some of the input variables are not known or computed until well after the PFR cost estimating stage. These inputs include traffic closures, storm drain extents and right-of-way acquisition costs. This finding was important for the research team, but it did not necessarily mean that those inputs required exclusion from the cost estimating model. Instead the research team realized that these variables would not be known to a high degree of confidence or precision, therefore top-down approximations of that value would need to be made.
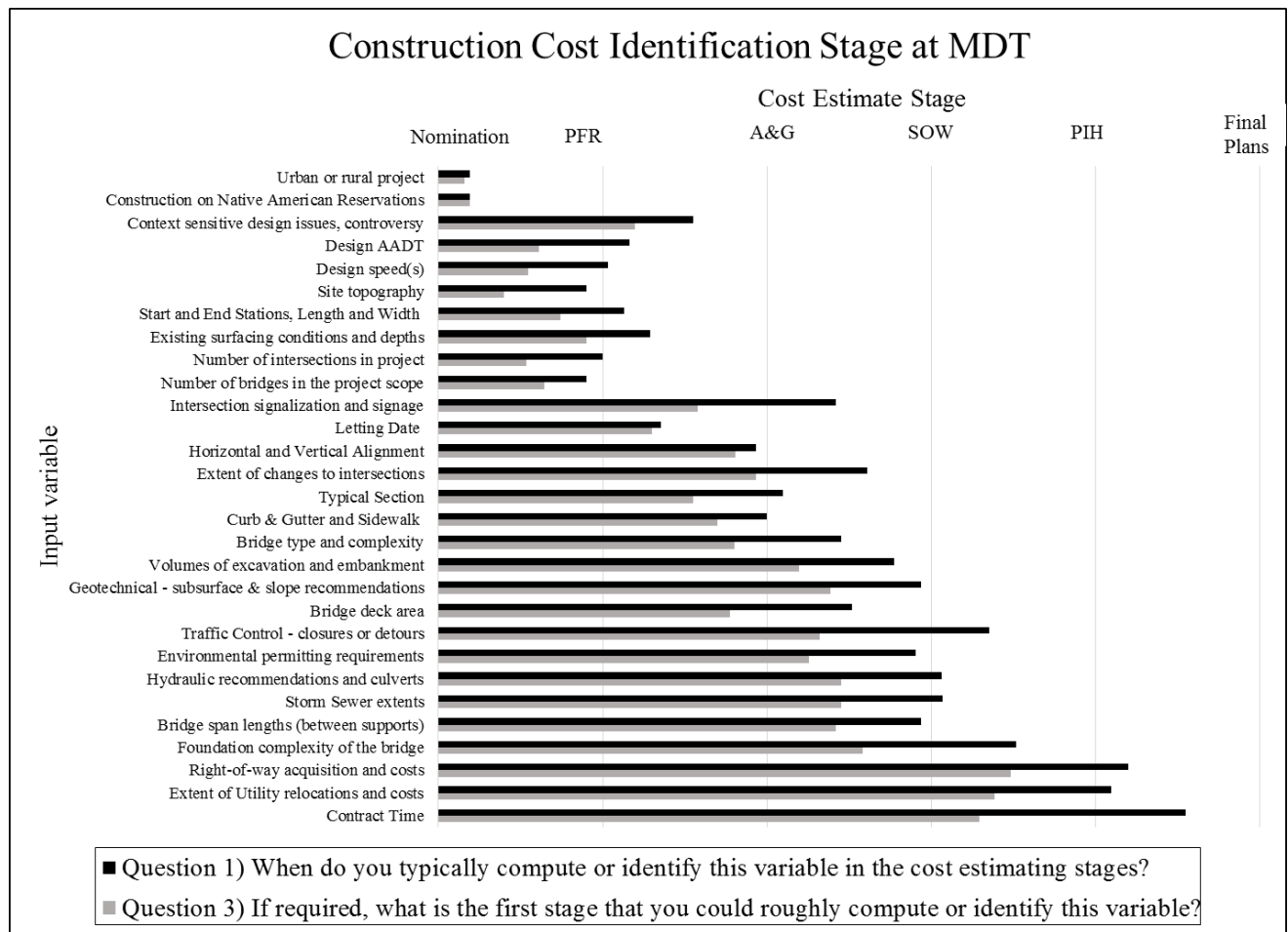
Figure 3. Understanding when MDT are knowledgeable on each of the 29 input variables (Questions 1 and 3 from the survey)

Secondly, the results of the survey were used to create a rational input variable selection method. An entire academic paper has been written to explain this process (Appendix D). Essentially, the research team investigated which variables MDT perceive as having a high influence on the construction cost, but require a low level-of-effort to compute or identify. It was proposed that these high-impact and low-effort variables should be added to the cost estimating model first to minimize the efforts required by MDT to conduct the conceptual estimate yet produce reasonable results. The survey results, with the proposed high-impact and low-effort variables, are shown in Figure 4. Note that the data-labels from 1 to 29 correspond to the input variables in Table 2 shown earlier in the report.
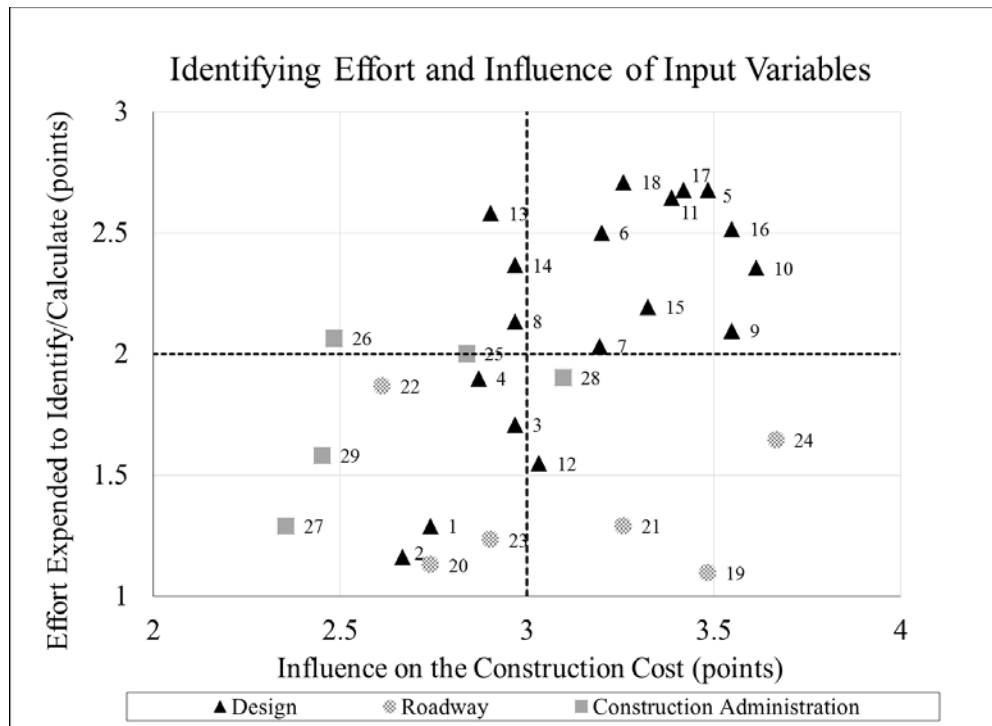
Figure 4. Understanding the level-of-effort and influence on the construction cost for each of the 29 input variables (Questions 2 and 5 from the survey)

The final recommendation of input variables for the pavement preservation model was based on the survey results in this section, interviews at MDT, literature reviews and trial and error using the model. Those recommended input variables are shown in Chapter 4. Long-term, once the final inputs are agreed upon, MDT could minimize the inputs stored in the 'global database' used for the cost estimating model and reduce its data collecting efforts.

Global Database Construction

To create the database (Task 3) for the cost estimate, multiple databases from MDT had to be integrated. Figure 5 schematically shows the multiple MDT databases that were combined to form the 'global database' for the cost estimating prediction model.

IOWA STATE UNIVERSITY
Institute for Transportation
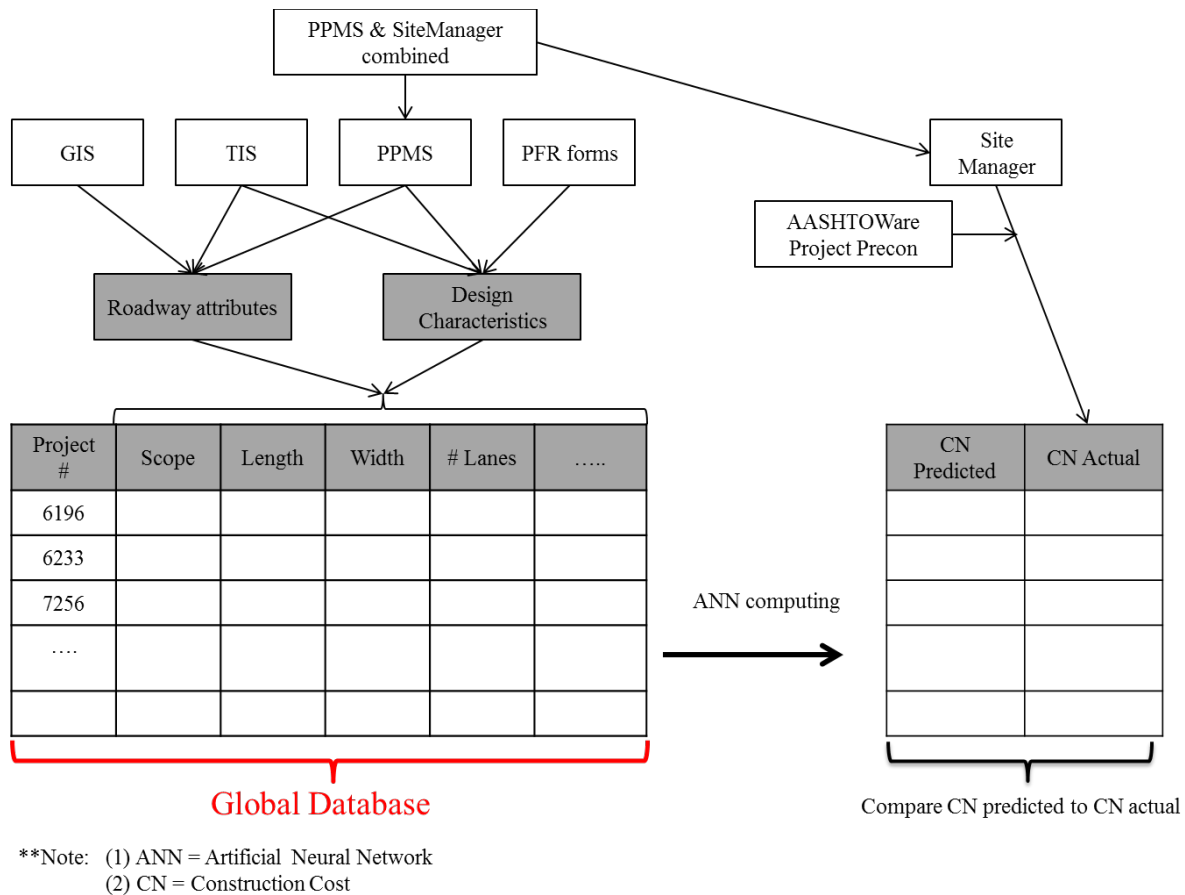
Figure 5. Incorporating databases to form a 'global database' for predicting construction cost with the model

Combining the databases was a resource intensive task for the research team due to:

- contradictions amongst the multiple databases,
- understanding the many data attribute fields,
- varying degrees of accuracy from one data source to another,
- ensuring that all of the same projects were aligned,
- finding data points that match those from the 29 suggested by MDT, and
- manual text mining of PFR reports to convert into a useable data format.

Each of the databases shown in Figure 6 are described in detail below.

PPMS and SiteManager® databases

The PPMS database stores information during the preconstruction stages of the projects, while the SiteManager® database is data collected during the construction stages of the project. This original database was provided to the research team and contained information from 1,012 projects of the different work-types.

This database was narrowed to include only pavement preservation work-types commencing construction between 2009 and 2013 (5 years of data). These years were selected because:

1. data in earlier years is sparse, and

IOWA STATE UNIVERSITY
Institute for Transportation

    2. data for projects completed in later years (2014 or 2015 start dates) do not contain all the final construction cost information.

SiteManager® data included final actual construction cost for each of the projects. To account for inflation 3% was applied to all projects in the 'global database' to bring the cost up to a base year.

PFR data extraction

The Preliminary Field Review (PFR) reports are completed by the district design project managers at the conceptual estimate stage. The reports, which are typically 6-10 pages long and include a cost estimate, are sent to the transportation commission from MDT for funding approval. The reports detail project information such as the expected scope-of-works, traffic impacts and highway dimensions at the conceptual estimating stage. During Task 3 the ISU graduate research assistant visited the Helena office and extracted all PFR reports from the Document Management System (DMS). This was completed using the methodology outlined in Appendix A. Note however that this procedure is only recorded for reference and no further reports need extracting for any work-type.

During the research meetings held with MDT on 26th and 27th February 2015, the research team realized that information contained in the PFR reports contained valuable attributes to include in the cost estimate model. Project attributes known at the time of the PFR stage are exactly the types of information available to conduct the cost estimate, thus are important cost predictors. The challenge with this project information contained in the PFR reports is that very little of the information is transferred to the PPMS database. This information in the PFR reports is textual information that needs to be converted into a more useable data format.

GIS meta-data

GIS information is specific to the roadway and not tied or linked to a project. As such, the GIS meta-data does not contain start and stop locations of projects. Data is instead recorded at set distances along the highway as key highway features change. Despite this disconnect to project information, the GIS information still contained relevant corridor attributes which the research team could utilize, for example the widths of roadways. GIS meta-data was manually searched in project locations to extract relevant corridor information.

TIS database

The TIS database contains final design information. This database is constructed by extracting information from final plans. From this database, basic project information (length, width, start RP and end RP) were extracted. Length and width information from the multiple databases could then be compared for consistency.

Completed global database

A summary of the completed 'global database' for pavement preservation projects is shown in Figure 6. Because the 29 possible input variables for the 'global database' were decided prior to collecting the data, it was not possible to directly find input variables that matched those specified. As a result, the research team identified 'measures' to best represent each of the 29 input variables. These measures and a summary of the inputs are both shown in Figure 6. Appendix B provides further information on the complexity rating system referenced in Figure 6.

IOWA STATE UNIVERSITY
Institute for Transportation

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	
		Urban indicator	PPMS				41	148	
1 Urban or rural project	Y	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									
		AADT at let year	GIS						AADT continuous range from 100 to 20667
4 Design AADT	Y	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									

Figure 6. The 29 suggested input variables, designated measures for that input variable and the data source(s) of each measure

IOWA STATE UNIVERSITY
Institute for Transportation

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	
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		WZSM	PFR						Level 1: 10; Level 2: 91; Level 3: 88
		Y	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

Figure 6. - continued

CHAPTER 3. RESULTS AND DISCUSSION

Due to the cost factors, unique to different types of construction projects, a separate top-down estimation equation was developed for each of the leading agency work types. The three work types identified for this purpose included resurfacing (which was further grouped under two categories as seal and cover, and rehabilitation projects due the considerable difference in their average project sizes), new construction and bridge replacement projects. Table 3 presents a summary of the projects included in fitting regression equations for the top-down estimation of construction costs.

Table 3. Summary of project types included in derivation of top-down cost estimating

Work Type	No. of projects	Low range ($)	High range ($)	Average project size ($)	Standard deviation
Resurfacing					
Seal and cover	81	66,000	1,400,000	455,000	328,000
Rehabilitation	97	529,000	5,000,000	1,904,000	911,000
Reconstruction	89	2,140,000	12,750,000	6,720,000	3,001,000
Bridge replacements	38	336,000	2,481,000	1,071,000	632,000

Rehabilitation Work Types: Seal and Cover; Work Type 183

Rehabilitation Work Types: (Minor rehab, thin lift resurfacing; Work Types 180, 181, 160)

Together, these three groups of projects account for more than 82% of the agency's construction program (included work types shown in bold font in Table 4) based on the 996 projects awarded between 2006 and 2015, with over $2.2 billion construction costs.

Table 4. Relative share of project work types (based on MDT project data 2006-2015)

Work Type	Average project size	Total project volume	Share of total
140 - RECONSTRUCTION - WITHOUT ADDED CAPACITY	**4,798,059**	**532,584,588**	**23.7%**
130 - RECONSTRUCTION - WITH ADDED CAPACITY	**7,845,479**	**321,664,649**	**14.3%**
180 - RESURFACING-ASPHALT (THIN LIFT<=60.OOMM)	**2,156,764**	**297,633,386**	**13.3%**
151 - MAJOR REHABILITATION-WITHOUT ADDED CAPACITY	**7,058,970**	**162,356,317**	**7.2%**
221 - BRIDGE REPLACEMENT WITH NO ADDED CAPACITY	**2,622,763**	**141,629,190**	**6.3%**
160 - MINOR REHABILITATION	**2,441,174**	**117,176,370**	**5.2%**
181 - RESURFACING-ASPHALT (THIN LIFT<=60.00MM)	**1,790,586**	**105,644,568**	**4.7%**
183 - RESURFACING - SEAL and COVER	**575,975**	**89,852,039**	**4.0%**
310 - ROADWAY and ROADSIDE SAFETY IMPROVEMENTS	585,887	84,953,677	3.8%
110 - NEW CONSTRUCTION	6,002,891	84,040,473	3.7%
170 - RESTORATION and REHAB - PCCP	3,677,677	58,842,839	2.6%
232 - MINOR BRIDGE REHABILITATION	2,094,922	46,088,275	2.1%
231 - MAJOR BRIDGE REHABILITATION NO ADDED CAP.	**2,486,638**	**34,812,928**	**1.6%**
220 - BRIDGE REPLACEMENT WITH ADDED CAPACITY	**3,018,530**	**30,185,304**	**1.3%**
111 - NEW CONSTRUCTION - FACILITIES	3,930,298	27,512,088	1.2%
150 - MAJOR REHABILITATION-WITH ADDED CAPACITY	3,317,361	19,904,165	0.9%
710 - CTEP PEDESTRIAN AND BICYCLE FACILITIES	3,708,217	14,832,866	0.7%
141 - RECONSTRUCTION - REMOVE and REPLACE CULVERTS	1,749,727	13,997,818	0.6%
410 - TRAFFIC SIGNALS and LIGHTING	273,403	13,396,736	0.6%
620 - BICYCLE and PEDESTRIAN FACILITIES	628,052	8,164,676	0.4%
120 - RELOCATION	4,044,517	8,089,034	0.4%
172 - RESTORATION and REHAB - FACILITIES	878,199	7,025,589	0.3%
650 - MISCELLANEOUS STUDY PROGRAMS	2,277,560	4,555,120	0.2%
411 - SIGNING, PAVEMENT MARKINGS, CHEVRONS, ETC.	225,093	4,051,682	0.2%
510 - ENVIRONMENTAL	313,510	3,135,104	0.1%
222 - BRIDGE REPL. WITH A CULVERT NO ADDED CAPACITY	**970,093**	**2,910,280**	**0.1%**
185 - RESURFACING - CRACK SEALING	567,664	2,838,320	0.1%
312 - STRUCTURE SAFETY	937,726	2,813,179	0.1%
311 - RAILROAD/HIGHWAY CROSSING SAFETY IMPR.	432,556	1,297,668	0.1%
313 - PEDESTRIAN and BICYCLE SAFETY	935,935	935,935	0.0%
182 - RESURFACING - PCCP	477,597	477,597	0.0%
412 - MISC. ELECTRONIC MONITORING OR INFO. SERV.	157,084	314,168	0.0%
GRAND TOTAL	**2,236,636**	**2,243,716,629**	**100.0%**

IOWA STATE UNIVERSITY
Institute for Transportation

Upon identification of the project work types, several linear regression equations were tested to achieve the best model fit to the agency project construction cost data. The final model specifications are shown in Table 5. In finalizing model specifications, variables were included first based on theoretical justification, and various combinations of project attributes were examined through stepwise regression to improve model fit and the normality of residuals. Due to the emphasis placed on the potential explanatory power of top-down project attributes in the early stages of cost estimation efforts, the proposed model variables were chosen only if they were believed to be readily available to the agency personnel at the project's inception. Hence, with each additional variable added to the estimation equation, any improvements in the resulting R-squared and mean absolute percent error (MAPE) values was weighed against increasing the complexity of the model specification.

Table 5. Summary of model specifications used for top-down estimation of construction costs

Project attributes	Seal and Cover	Rehabilitation	Reconstruction	Bridge Replacement
Area	X	X	X	
Length	X	X	X	X
Width	X	X	X	X
Highway functional classification	X			
Urban Area	X	X	X	X
Geographical complexity	X	X		
Added capacity			X	
No of bridges		X		
Expected contract time		X	X	
Indian reservation				X
Resurfacing variables				
Milling volume		X		
Overlay volume		X		
PFR Milling scope		X		
PFR Overlay Scope		X		
Bridge types				
Concrete				X
Pre-stressed concrete				X
Steel				X

MDT's preliminary field review (PFR) reports have been the primary source of top-down project attributes used to generate construction cost estimates. Although most of the project attributes used as prediction variables were also available on MDT's project management data files, and can be readily used to predict construction costs for future projects, some additional variables, such as

IOWA STATE UNIVERSITY
Institute for Transportation

milling and overlay scope parameters were not, are thus recommended to be tracked for ongoing model implementation and refinement. Other such variables included the design type for bridge replacement projects, which were captured from a reconciliation of the MDT's National Bridge Inventory database with the bridge replacement projects included in the analysis.

As expected, almost all models include some key project attributes, such as roadway surface area, length, width, and urban area indicator. Yet several other key attributes, such as geographical complexity, were included when their presence improved the models' prediction performance. In addition, some project attributes, such as resurfacing scope parameters and bridge design types, were included only in their respective work type estimation equations.

Table 6 and Figure 7 provide a comparison of the prediction accuracy of multiple regression, neural network and MDT preliminary construction cost estimates for the broad work type groups mentioned above. The best multiple regression model fit, as indicated by the R-squared value of 0.86, was achieved for seal and cover work type projects. Not only did these projects tend to be small in size (with average construction costs of $455,000), their limited scope of work activities arguably lead to higher levels of accuracy when measured in mean absolute errors. Despite the relative success of fitting cover and seal projects, however, the proposed multiple regression and neural network estimation methods fall short of providing meaningful improvements over the agency's early construction cost estimates for pavement preservation project in general. As mentioned earlier, the relatively lower levels of uncertainty inherent in pavement preservation projects, also seem to help MDT to reach its highest levels of estimating accuracy for this work type.

On the other hand, for the remaining two leading work types, the two proposed prediction methodologies do provide considerable improvements over the agency estimates. For example, the regression-based prediction model results in a 22% MAPE level, compared to the agency's 30% MAPE value (Table 6) for the 89 projects included under the reconstruction work type. Further, an even higher reduction in MDT's average prediction error for early cost estimates for bridge replacement projects was achieved. The regression and neural network methods achieved 30% and 25% MAPE values respectively, whereas the MDT prediction accuracy for these types of projects was significantly higher at 43%.

Since these two groups of work types together make up approximately 50% of the MDT construction volume, the implications of the suggested enhancements in MDT's prediction accuracy for its early construction estimates could lead to sizeable improvements in its budgeting efforts.

Table 6. Summary of model specifications and prediction accuracy for the top-down estimating of construction costs for leading project types

Work Type	No. of projects	Average construction cost ($)	R Squared	PFR MAPE	Regression MAPE	Neural Network MAPE
Rehabilitation	81	455,000	0.86	0.21	0.21	0.20
Rehabilitation	97	1,904,000	0.63	0.21	0.25	0.23
Reconstruction	89	6,720,000	0.74	0.33	0.22	0.22
Bridge Replacements	38	1,071,000	0.73	0.43	0.30	0.25

IOWA STATE UNIVERSITY
Institute for Transportation
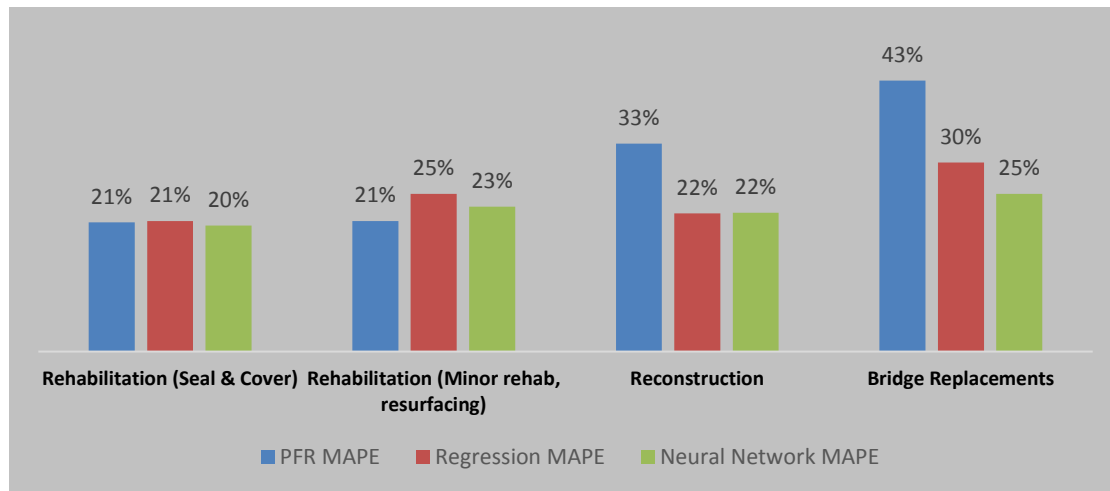
Figure 7. Prediction accuracy of multiple regression, neural network cost estimating methods, and MDT estimates (PFR)

As a result of the analysis discussed above, Figures 8, 9, and 10 show flow charts that describe the process to compare current agency cost estimating accuracy to proposed multiple regression and neural network cost estimation models. The output of the analysis will determine whether the proposed statistical models improve estimating accuracy over the current system. It will also furnish a quantitative measure of the amount of improvement for each separate dataset. The findings of the top-down cost estimation analysis presented here are also made available in an Excel spreadsheet format to enable MDT to estimate construction costs during its project funding stages. Each of the furnished spreadsheets, customized based on the unique estimation equations for different project work types, is a self-contained file that includes a user's manual/instructions, project data driving the regression estimations, and a methodology to measure the accuracy of model estimates in the form of mean absolute percentage errors.
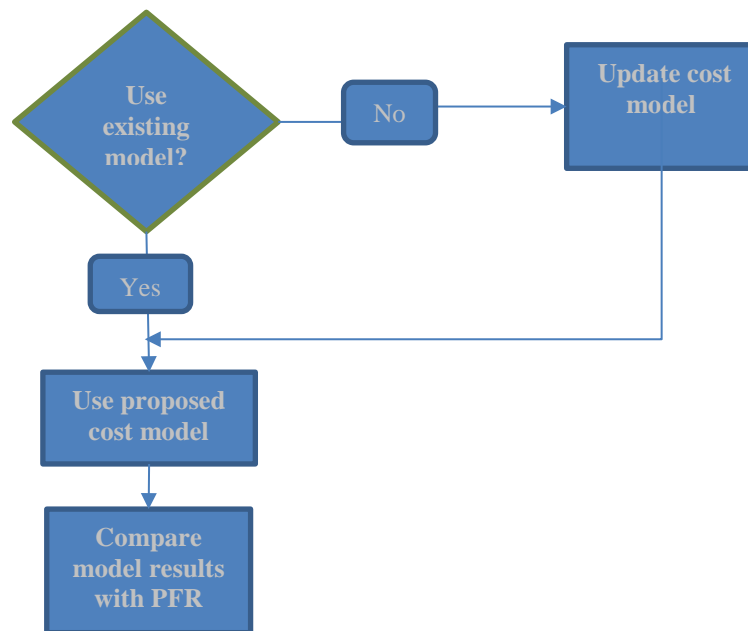
**Figure 8. Proposed cost prediction process for top-down estimation
of MDT construction costs**
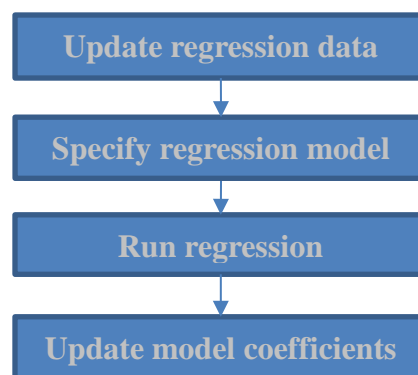
Figure 9. **Proposed update process for cost prediction model**
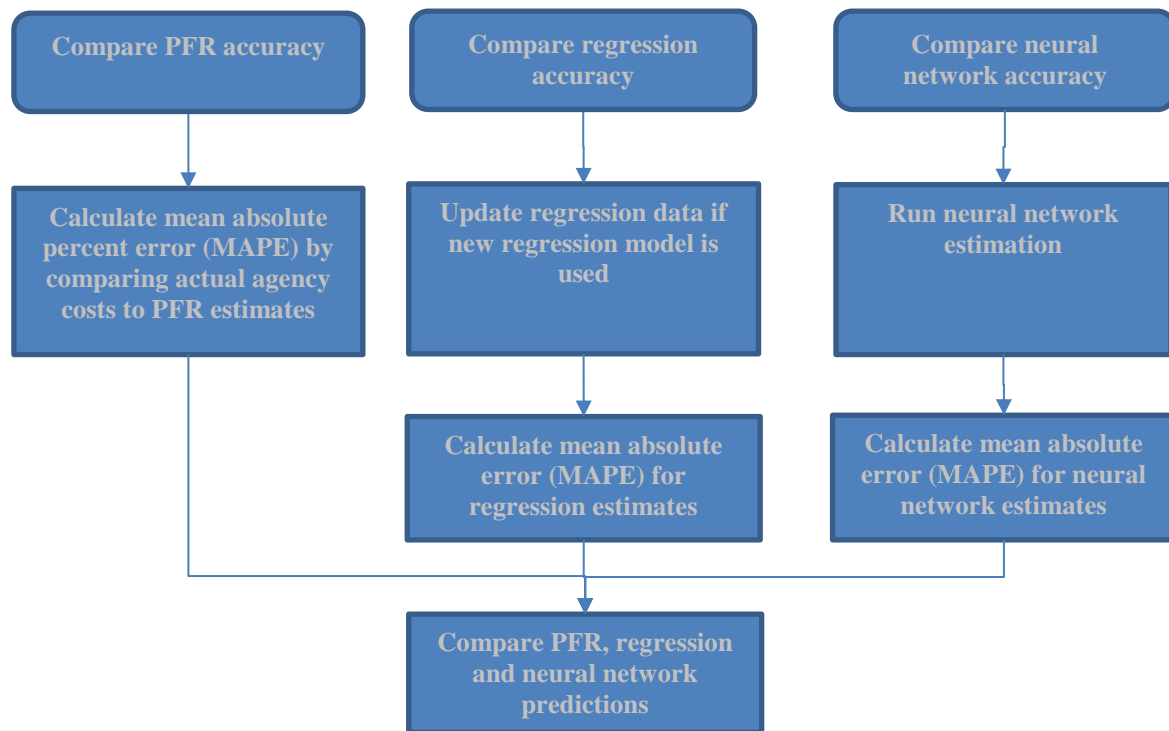
Figure 10. Proposed process for comparison of agency cost estimating accuracy to proposed multiple regression and neural network cost estimation models

The details of the results of statistical analyses shown in Figure 10 are found in Appendix A. The interpretations of the research results are discussed in detail in three peer-reviewed journal articles. Rather than repeat them in the body of this report, they are contained in the Appendices D, E, and F to the report for the convenience of the reader. The next section summarizes the highlights of each paper in terms of each papers' major conclusions and contributions.

Additional Research Findings

In addition to the developed estimating models and research findings for this project, the research team has utilized data from this project to further extend the highway cost estimating body-of-knowledge. Three research papers have been completed for submission to academic construction journals and are summarized below. The implications of their findings to the MDT project have been detailed.

Paper 1 Summary: "Quantifying Efforts in Data-Driven Conceptual Cost Estimating Models for Highway Projects"

This paper, included in Appendix D, investigated how to establish the fewest number of input variables required at the conceptual cost estimating stage to develop a suitable estimate. Typically, it is perceived that more project detail enhances cost estimating accuracy, however this paper challenged that theory by investigating the minimum number of variables that needed be included.

IOWA STATE UNIVERSITY
Institute for Transportation

A rational method was proposed and then validated to select the most suitable input variables for MDT. This involved the Cost Estimating survey conducted at MDT to identify those input variables which have the largest influence on the construction cost but require the least amount of knowledge to identify at the conceptual stage. The paper found that after around 6-8 input variables, then additional input variables no longer enhanced the estimate accuracy. As a result, it is suggested that MDT prioritize collecting these highly influential variables in the 'global database' which is to be used by the cost estimating model.

Paper 2 Summary: "Stochastic Conceptual Cost Estimating of Highway Projects to Communicate Uncertainty using Bootstrap Sampling"

This paper, included in Appendix E, investigated how artificial neural networks developed for estimating a construction cost could be extended to produce a stochastic cost (as opposed to a point estimate). This research paper recognized that not all cost estimates at the initial stage are known to the same degree of confidence. For example, MDT will generally be very confident on the cost of a chip-seal project compared to a highway reconstruction due to the level of unknowns at the early stages. This research therefore attempted to investigate if the artificial neural network could better communicate the conceptual cost estimate through a range of construction costs.

This paper realized substantial benefits for MDT to express the cost estimate as a range of costs at the initial stage (as opposed to a single number). The confidence intervals produced as part of this research could be used to communicate the level of certainty to transportation committees or the public. Alternatively, the range of cost estimates developed could aid a highway agency to rationally assign contingency at the conceptual stage to individual projects.

Paper 3 Summary: "Rationally selecting data for Highway Construction Cost Estimating at the Conceptual Stage"

This paper, included in Appendix F, investigates the sizes of databases used by previous authors of conceptual cost estimating models. It challenges practical application of those publications achieving such high performance but using so few projects in their database to power their model. The most relevant finding of this paper to the MDT project is that larger databases powering an estimate model only increase the accuracy and reliability of an estimating model. This is in-line with other literature findings.

IOWA STATE UNIVERSITY
Institute for Transportation

CHAPTER 4. CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER RESEARCH

Early cost estimating methodologies provided in this report build on an important finding also developed through the course of this work, which is described in detail in Research Paper 1. While the desirability of more data to improve estimating accuracy is accepted as a truism, this study demonstrates that not all project attributes have similar information content. In fact, a handful of variables can reach accuracy levels that cannot be necessarily improved with the addition of further variables.

The outcomes of this research study thus suggest a top-down estimating approach can help achieve considerable efficiencies in improving MDT's budgeting and project funding processes. The study investigates the applicability of multiple regression and neural network estimation approaches, in particular, and finds that their employment in the early estimation of agency construction costs can increase the accuracy of agency estimates.

The findings confirm the importance of project work types in the accuracy of early cost estimates. New construction and bridge replacement projects are found to have relatively higher potential to gain from further process improvements. Even when the accuracy of the proposed methodologies seem to be limited for some work types, it should be noted that the inclusion of these methods into agency practices can still be desirable given the relative ease of generating early estimates. As with any estimation model that relies on historical cost data, however, the performance of the estimation equation depends on the relevance and accuracy of the data used to predict model coefficients. It is therefore important to regularly monitor and calibrate model specification and functional form provided in the estimating spreadsheet.

The main findings of the three research papers written under the study are summarized as follows:

Research Paper 1

This paper, included in Appendix D, shows that a carefully constructed input selection method may achieve optimal model specification by eliminating the often ad hoc stepwise regression practices common in early estimation practices. Such a methodology indeed promises considerable efficiencies to agency budgeting processes due to the simplicity and relative availability of project attributes during a project's inception stages.

Multiple regression and neural network estimation models both reached the goal with the dual-objectives of low effort and high accuracy, suggesting that top-down estimating practices are capable of matching or exceeding accuracy levels that are typically reached with considerably more input intensive estimation equations. In the case of conceptual estimation of MDT construction costs, both the multiple regression and artificial neural network approaches showed that incremental data variables detail to the model reached a point of diminishing returns at roughly six to eight high impact/low effort variables.

In fact, adding further input variables using either model technique resulted in diminishing returns of the model performance. This finding has positive implications for practitioners willing to employ data-driven conceptual cost estimating techniques.

Research Paper 2

As the second research paper, Appendix E, emphasizes, point estimates are single numbers with no indication of the level of confidence with which they have been developed. In later estimating stages, when quantities are known, highway agencies can be more confident and can express the estimate in that form.

However, for the earlier estimate stages, where project scope is less developed, the estimate should be expressed in a manner that describes the estimator's confidence; providing a range does just that. The communication of estimate confidence through a range could help remove optimism and bias inherent with conceptual cost estimates.

Additionally, the power of developing an empirical distribution for individual projects highlights a method that highway agencies can use to assign contingency. The findings of 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.

Research Paper 3

Despite the widely held belief that more data increases the accuracy and reliability of data-driven CCE models, a content analysis of 20 data-driven construction cost estimating models revealed that some models had a very low prediction error despite using few projects to train the model (Appendix F).

To help improve the accuracy of construction cost estimating methods, 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 only a subset of project data could reasonably improve model prediction accuracy as long as key attributes were captured in the sample data.

The vast improvement in computing technologies over the past 30 years, including artificial neural network estimation techniques, holds considerable potential in improving the performance of construction cost estimating practices. The DOTs, in particular, could benefit from computational advances in their budgeting efforts.

Recommendations for further research

Cost estimating equations in this report were developed through the consolidation of high-level project information that is available during the project inception phase with the projects' final construction costs based on their contract award information. Due to the constantly evolving nature of project scopes during the project development stage, the ease of updates to early cost estimates as scope changes occur will be critical for the efficient implementation of the proposed methodologies. As such, the integration and timely update of early project information on MDT project management systems is a logical next step in further improving the initial model specifications provided here. Further, tailoring MDT project management systems with an emphasis on capturing project information essential to the accuracy of early estimating practices is expected to increase the confidence levels of agency's budgeting efforts notably. Finally,

identifying those projects that experienced considerable variances from funding to the award stage and the analysis of such unexpected deviations from baseline budgets will ensure the calibration of the estimation equations as MDT's dynamic planning needs continue to evolve.

APPENDIX A. EXTRACTING PFR REPORTS AND REGRESSION AND NEURAL NETWORK RESULT COMPARISON

Extracting PFR reports

This appendix is a summary of method used to extract the PFR reports that were used to create the 'global database'
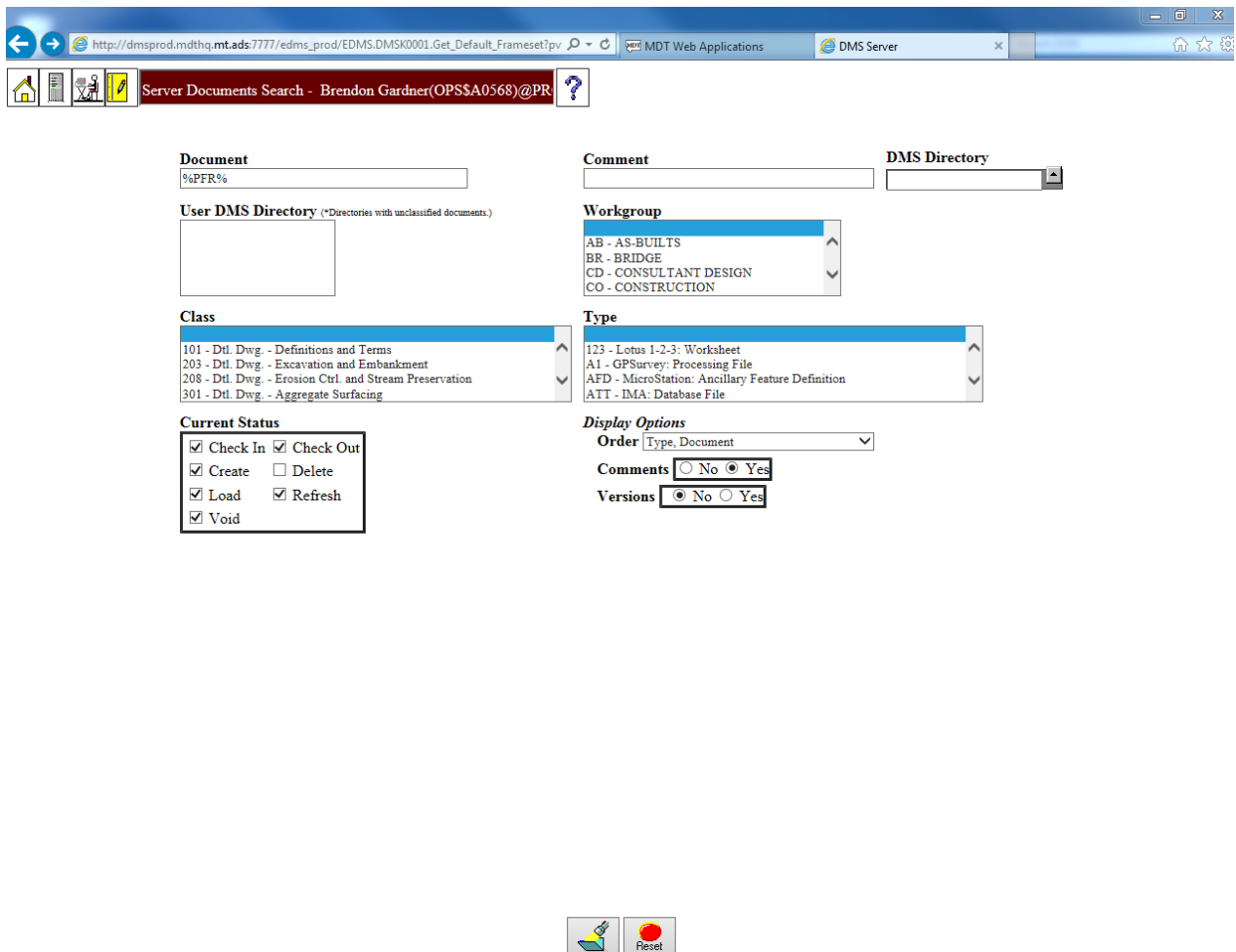
>> Log-in to DMS: MDT intranet >Resources>Web-applications>Document Management System (DMS) located under the Highways and Engineering section.
>> Go to home (top-left button)
>> Type %PFR% into document.
>> Highlight class, workgroup and type in the blank area
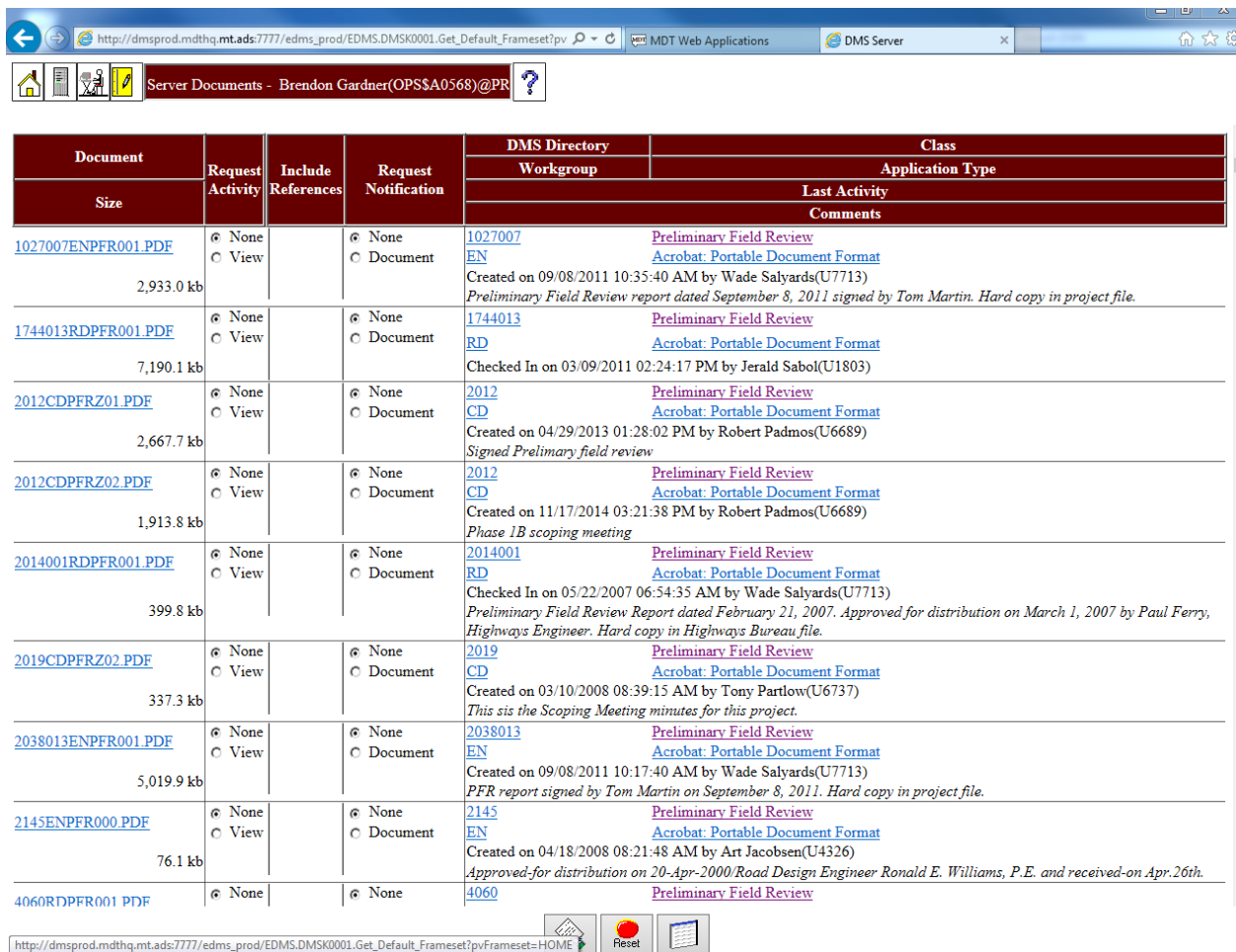
>> Click

… then it will take a-while to find as searching for all PFR's

>> On the ones that you want click view (can do many views at once).

>> Then click (in the middle bottom)

>> DMS Processes to find all the documents you have clicked view on
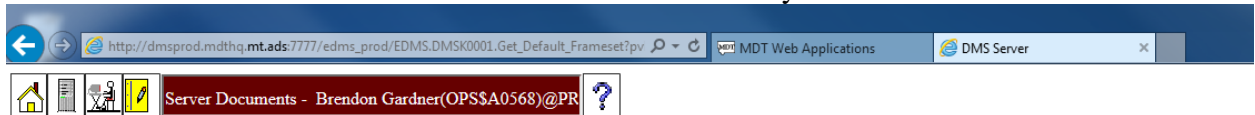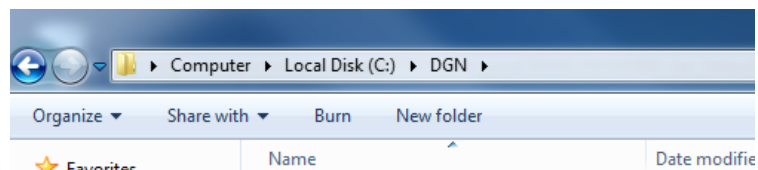
>> The files you view will be in

>> Can just copy them out.

IOWA STATE UNIVERSITY
Institute for Transportation

Statistical Output of Multiple Regression Analysis

Multiple regression output for Pavement Preservation (Seal and Cover) Work Types

Regression Statistics	
Multiple R	0.93
R Square	0.86
Adjusted R Square	0.85
Standard Error	125,802
Observations	81

	df	SS	MS	F	Significance F
Regression	6	7.41E+12	1.24E+12	78.08	0.000
Residual	74	1.17E+12	1.58E+10		
Total	80	8.59E+12			

	Coefficients	Standard Error	t Stat	P-value
Intercept	108,644	85,320	1.27	0.21
Area	0.20	0.06	3.43	0.00
Width	-2,102	2,218	-0.95	0.35
Length	297	11,146	0.03	0.98
Hwy functional class.	46,627	26,649	1.75	0.08
Geographical complex.	78,023	44,806	1.74	0.09
Urban area	21,260	42,815	0.50	0.62

IOWA STATE UNIVERSITY
Institute for Transportation

Multiple regression output for Pavement Preservation (Rehabilitation) Work Types

Regression Statistics	
Multiple R	0.79
R Square	0.63
Adjusted R Square	0.57
Standard Error	595,407
Observations	97

	df	SS	MS	F	Significance F
Regression	12	5.00E+13	4.16E+12	11.75	0.000
Residual	84	2.98E+13	3.55E+11		
Total	96	7.98E+13			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-1,774,122	521,349	-3.40	0.00
Area	-0.18	0.40	-0.46	0.65
No. of bridges	36,062	22,356	1.61	0.11
Width	24,853	9,344	2.66	0.01
Length	211,304	45,654	4.63	0.00
Hwy functional class.	245,378	100,640	2.44	0.02
Geographical complex.	23,380	103,893	0.23	0.82
Urban area	-465,181	215,998	-2.15	0.03
Exp. contract time	9,308	2,766	3.37	0.00
PFR: Milling	-458,937	256,998	-1.79	0.08
PFR: Overlay	1,536,393	283,527	5.42	0.00
Milling volume	1.27	0.96	1.32	0.19
Overlay volume	-2.50	1.45	-1.72	0.09

IOWA STATE UNIVERSITY
Institute for Transportation

Multiple regression output for New Construction Work Types

Regression Statistics	
Multiple R	0.86
R Square	0.74
Adjusted R Square	0.72
Standard Error	1,602,916
Observations	89

	df	SS	MS	F	Significance F
Regression	6	5.86E+14	9.77E+13	38.01336	0.000
Residual	82	2.107E+14	2.57E+12		
Total	88	7.967E+14			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-808,955	979,192	-0.83	0.41
Area	3.09	1.36	2.27	0.03
Length	-100,172	196,739	-0.51	0.61
Width	-15,271	23,782	-0.64	0.52
Exp. contract time	39,241	4,094	9.59	0.00
Urban area	196,097	473,943	0.41	0.68
Added capacity	1,127,040	424,866	2.65	0.01

IOWA STATE UNIVERSITY
Institute for Transportation

Multiple regression output for Bridge Replacement Work Type

Regression Statistics	
Multiple R	0.85
R Square	0.73
Adjusted R Square	0.66
Standard Error	366,464
Observations	38

	df	SS	MS	F	Significance F
Regression	7	1.074E+13	1.53E+12	11.42134	0.000
Residual	30	4.029E+12	1.34E+11		
Total	37	1.477E+13			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-622,604	403,809	-1.54	0.13
Width	51,601	11,273	4.58	0.00
Length	5,177	760	6.81	0.00
Concrete	-437,182	318,033	-1.37	0.18
Pre-stressed conc.	-650,399	290,474	-2.24	0.03
Steel	-474,285	340,202	-1.39	0.17
Urban area	204,121	276,859	0.74	0.47
Indian reservation	679,729	191,413	3.55	0.00

IOWA STATE UNIVERSITY
Institute for Transportation

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 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	Categorical Exclusion or Environmental Assessment 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	Environment Impact Study or complex Environmental Assessment required Studies of multiple alternatives Continued public and elected officials involvement in analyzing and selecting alternates Other agencies (such as FHWA, COE, EPA, Fish, Wildlife and 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 guardrails - 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.

IOWA STATE UNIVERSITY
Institute for Transportation

APPENDIX C. SURVEY AND RESULTS

MDT ★   IOWA STATE UNIVERSITY
                 Institute for Transportation

Confused by the questions? - feel free to contac
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			

IOWA STATE UNIVERSITY
Institute for Transportation

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			

IOWA STATE UNIVERSITY
Institute for Transportation

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				

IOWA STATE UNIVERSITY
Institute for Transportation

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 and 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 and 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 acquisition 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	4	2	3	4	3	3	2	3	4	3	5	5	5	4	4	5	6	6	6	
R_dg4ugYvLdQLZljb		Road Design	2	2	2	2	2	3	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	3	4	2	2	2	4	3	4	3	3	4	4	4	4	4	5	4	4	5	5	5	6		
R_1n1qIsafURopcxX	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_3fDBxpl87M2jdqE			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	2	2	2	1	3	3	2	4	3	2	4	3	3	1	3	2	4	4	5	5	5	3	5	
R_2WlT1OSEFtocCBd	District Projects Engineer	Billings	1	1	4	2	2	1	1	1	1	1	4	1	3	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	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	3	2	2	3	4	4	4	5	6	5	5	5	4	4	4	5	6	5	5	6		
R_2zzHL8ADHeSyOQa	Design Supervisor	Missoula District	1	1	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_3Mlo4lu2WdLyl7x	CE Specialist IV	Highways Preconstruction	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	1	3	5	5	4	5	4	4	5	5	5	5	5	5	6	6	6		
R_YXicvrJ6nfdaSJz			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	4	2	2	5	5	3	4	3	4	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	5	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	1	3	3	3	2	4	3	4	4	4	4	5	4	4	6	6	6		
R_2uqQcZdnhzKODL0	Area Engineer	Bridge	1	1	4	2	2	3	3	3	2	5	1	3	4	3	2	3	3	5	3	4	4	4	3	5	5	6	6		
R_2Eyt8bbkBpvebRB	District Preconstruction Engineer	Glendive District	1	1	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		
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	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	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_2WGoJ6hlpTNOOWj	Project Fatilitation Specialist	Consultant Design	1	1	2	2	3	1	3	1	1	2	3	3	2	2	3	2	3	3	5	5	4	3	5	4		5	5		
R_2wuClEWuMtVdyWD	Civil Engineer	Highways/Engineering	1	1	3	2	2	2	2	2	4	4	2	2	2	2	4	5	3	4	5	5	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	3	3	4	3	2	3	5	5	3	2	2	5	5	3	6	6	5		
R_Sl96EpxiXBc69Xz			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

The table has two main sections:

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?

Each section contains the following column variables (repeated for both sections):

- 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

IOWA STATE UNIVERSITY
Institute for Transportation

	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

<div align="center">

APPENDIX D. RESEARCH PAPER 1

</div>

Quantifying Efforts in Data-Driven Conceptual Cost Estimating Models for Highway Projects

<div align="center">

Brendon Gardner, Douglas D. Gransberg, and H. David Jeong. *Paper submitted to the ASCE Journal of Construction Engineering and Management.*

</div>

A modern dilemma has emerged in light of ever improving technological advances, explosive 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 extra 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 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.

Keywords: Conceptual cost estimating, highway infrastructure, artificial neural networks, multiple-regression analysis

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 Statewide Transportation Improvement Program (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).

IOWA STATE UNIVERSITY
Institute for Transportation

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 selected 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 in Highway Agencies

Conceptual cost estimating (CCE) is the first construction cost estimate completed for a project, as shown in Figure 1. At the conceptual stage there is little information known about a project and the detailed design has not yet begun. As the importance of the information required for the conceptual estimate increases, so too does the need to expend additional design and planning, this in-turn extends the project planning period. Further design and planning details can be included in later, more confident, estimates.
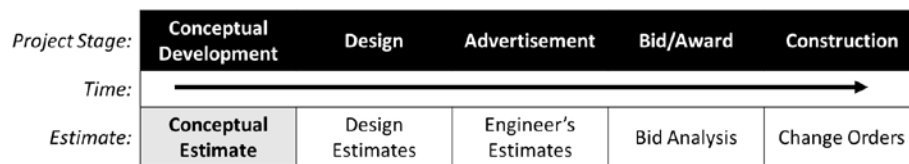
Fig. 1. Construction cost estimating timeline (adapted from Schexnayder et al. 2003)

Highway agencies cannot afford to over-invest their design time and effort 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. 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).

CCE techniques currently used by highway agencies vary by state. Byrnes (2002) and Turochy et al. (2001) have both completed surveys on cost estimating at the planning stage. These studies found that CCE approaches utilized by highway agencies are generally classed into one the following three categories:

1. "cost-per-mile" of typical sections of highway or bridge,
2. estimating approximate quantities of major work items, or
3. no documented or uniform method, instead using experience and engineering judgement.

Despite these techniques developed, CCE at highway agencies still requires improvement. Flyvbjerg et al. (2002) investigated 258 public transportation projects and found that 86% of those

IOWA STATE UNIVERSITY
Institute for Transportation

projects had experienced cost growths since the initial estimate, on average they were 28% higher than the initial estimate. Further, in 2003, Schexnayder et al. stated that recent publicity has called into question the "ability of departments of transportation to forecast accurately and to control the final cost of their projects", this was stated in the *NCHRP Synthesis of Highway Practice - Project Cost Estimating.*

Data-driven techniques using Artificial Neural Networks (ANNs) and Multiple Regression Analysis (MRA) have been frequently suggested in the literature for CCE and show equal, if not superior results than those currently used by highway agencies (Bell and Ghanzanfer 1987; Hegazy and Ayed 1998; Moselhi and Hegazy 1992). The performance of these models are well within the acceptable estimate range for the planning stage suggested by the American Association of State Highway and Transportation Officials (AASHTO) in the *Cost Estimating Guidelines* (AASHTO 2013). In 2002 Byrnes found that no state highway agency is yet employing mathematical models, NCHRP report 574 (Anderson et al. 2007) reached the same conclusion.

Both MRA and ANNs link a historical database of project attributes to the actual construction cost of each project. These relationships identified within the data can then be used to forecast the construction cost of future projects. MRA links the information with a linear equation to the construction cost (Turochy et al. 2001). Each attribute is assigned a weight when the linear equation is developed such that the error in forecasting the construction cost is minimized. ANNs on the other hand use artificial intelligence to find patterns to describe the construction cost from a historical database of project attributes (Pewdum et al. 2009). Historical data is used to train the ANN model and recognize relationships within the database. This trained model is then used to forecast future construction costs by looking for similar patterns.

No matter the CCE technique employed by highway agencies or suggested in the literature, a particular level of project definition (or design effort) is required in order to conduct the 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 sooner that the initial estimate is developed, the smaller the level of project definition required for CCE with commensurately lower cost and effort. This then means that estimators and designers can focus their efforts on projects which are beyond the planning stage and are likely to reach construction.

Data-Driven CCE Models – prior studies

CCE techniques reviewed in this research include ANN and MRA models, these are both commonly suggested in the literature and will be referred to as data-driven CCE 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;

IOWA STATE UNIVERSITY
Institute for Transportation

Petroutsatou et al. 2012). Bell and Ghanzanfer (1987) selected their final MRA model by building many models through trial and error and then selecting the model which produced the least error. This technique was used in at least four other studies (Creese and Li 1995; Hegazy and Ayed 1998; Gunduz et al. 2011 and Petroutsatou et al. 2012).

Input variables selected have a large effect on the prediction capability of the CCE model. Bell and Ghanzanfer (1987) concluded this using MRA to predict the construction cost of highway projects. 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). Bell and Ghanzanfer investigated the accuracy of the input variables in their research. Gunyadin and Dogan made reference to the selection of the input variable types, and Setyawati et al. referred to optimizing the number of input variables to achieve better prediction accuracy. Model creators usually only have a one-time commitment to collecting the cost predictors (Smith and Mason 1997). If model creators select cost predictors which require a large amount of data collection and processing effort then it will burden the usefulness of the model as a CCE tool.

Despite the amount of previous research in data-driven CCE models, none of the work reported in the construction literature studied quantifies the effort required to conduct the cost estimate. 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.

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. The literature analysis was a starting point of this research to see if additional inputs improve estimating accuracy.

A total of 16 publications were studied with results from data-driven CCE models. Publications were selected that involved either ANN or MRA prediction algorithms to output the construction cost of the project using input variables at the early design stage. From each of the publications both the performance and the number of input variables used to produce their best performing model was collected. The results of this research is shown in Table 1. Some publications tested both the ANN and MRA techniques which resulted in a total of 20 models for comparison shown graphically in Figure 2.

Performance of both the ANN and MRA models were measured using the Mean Average Percentage Error (MAPE). This method is commonly used by authors of data-driven CCE models (Petroutsatou et al. 2012; Gunduz et al. 2011; Mahamid 2011; Hegazy and Ayed 1998). Calculation of the MAPE is furnished using Equation 1 (Mahamid 2011).

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

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 collected for the ith project

IOWA STATE UNIVERSITY
Institute for Transportation

Table 1. Construction cost estimating models studied

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%	–	RC 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-Tahtabai 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
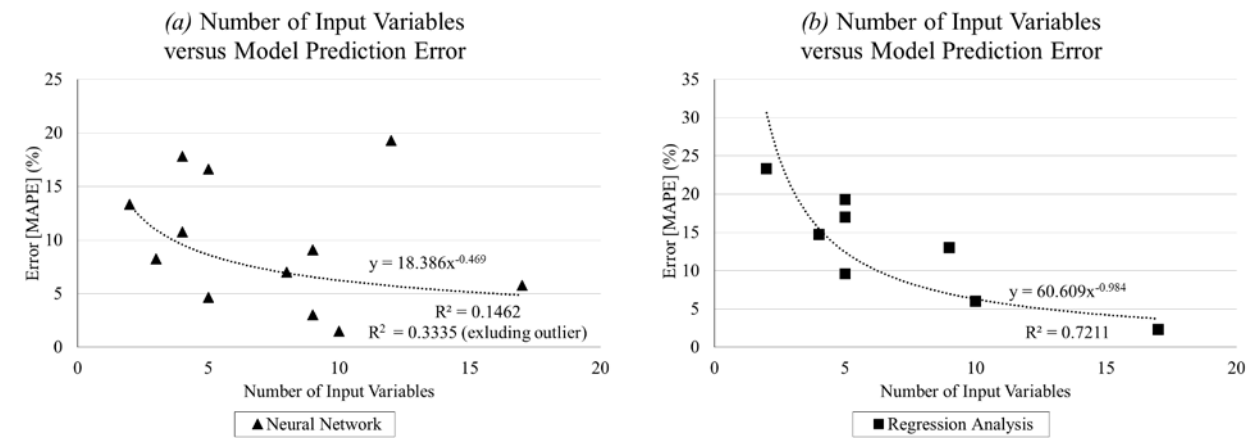
Fig. 2. Literature analysis of inputs versus error

The results from the literature analysis in Figures 2a and 2b show that previous publications are achieving lower error through more input variables. Both plots in Figure 2a and 2b 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 in the literature with the power curve R^2 value being 0.7211. When the obvious outlier in Figure 2a is removed then the R^2 value in that plot increases to 0.335.

A weakness of this conclusion is that the literature is for projects of many different scopes. Additionally, none of these past studies have converted their input variables into perceived effort, this means that effort and performance cannot be directly compared. Instead an assumption of this literature analysis is that each input variable requires equal estimating effort. The results of this study specifically quantify input variable effort and suggest that not all input variables require the same level of effort to compute.

The requirement to minimize estimating effort for CCE 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 3, to judge 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.
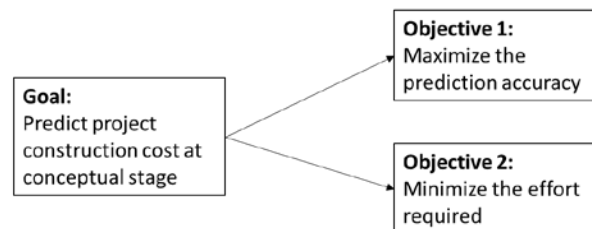
Fig. 3. Proposed dual-objective hierarchy tree for conceptual cost estimates

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.

IOWA STATE UNIVERSITY
Institute for Transportation

Data will be furnished from one state highway agency, Montana Department of Transportation (MDT) to conduct the research.

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 4 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 (3a) is compared to completing the task in reverse order (3b) in order to validate framework effectiveness.
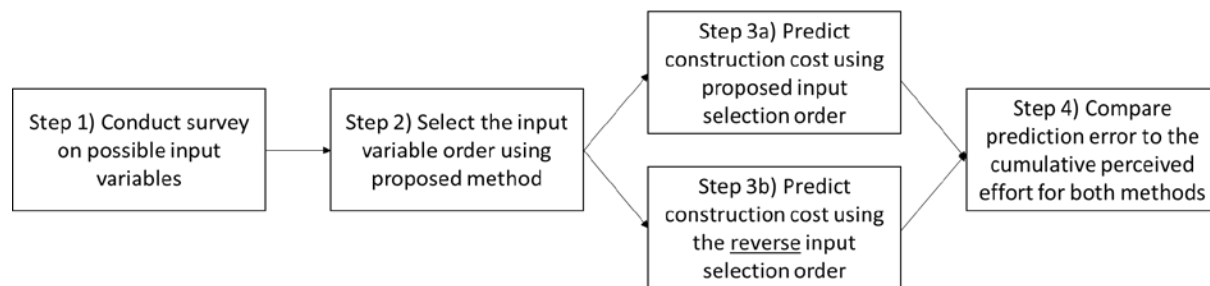
Fig. 4. Research steps

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. 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 2. 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.

IOWA STATE UNIVERSITY
Institute for Transportation

Table 2. 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		
8	Curb, gutter and sidewalk		**Construction Administration attribute**
9	Bridge type and complexity	25	Traffic Control - closures or detours
10	Volumes of excavation and embankment	26	Environmental permitting requirements- wetlands
11	Geotechnical - subsurface and slope recommendations	27	Letting Date
12	Bridge deck area	28	Context sensitive design issues, controversy
13	Hydraulic recommendations and culverts	29	Contract time
14	Storm drain 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		

Respondents of the survey 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 questions were designed with an ordinal (categorical) scale where respondents are required to select the most suitable answer as shown in Figure 5 (Fink 2009; Fowler 2009).

IOWA STATE UNIVERSITY
Institute for Transportation

| 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

Fig. 5. 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 researchers were satisfied that the 37% response rate reflected 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 6 with the input variables suggested to be selected in the bottom right hand quadrant.
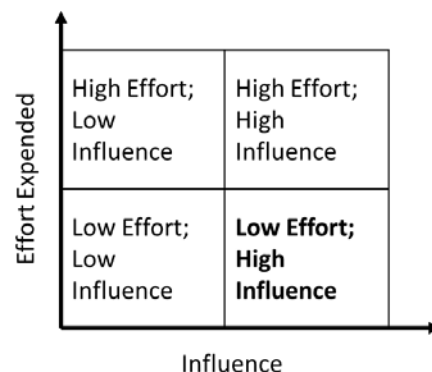
Fig. 6. Selecting input variables to meet the dual-objectives of CCE

To validate this selection process the research team combined the perceptive 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 databases provided by MDT 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 6 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 6, this is explained in

IOWA STATE UNIVERSITY
Institute for Transportation

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 7, the numbers relate to the 29 attributes from Table 2. 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 5. 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 5, they are referred to as "points" from here on.
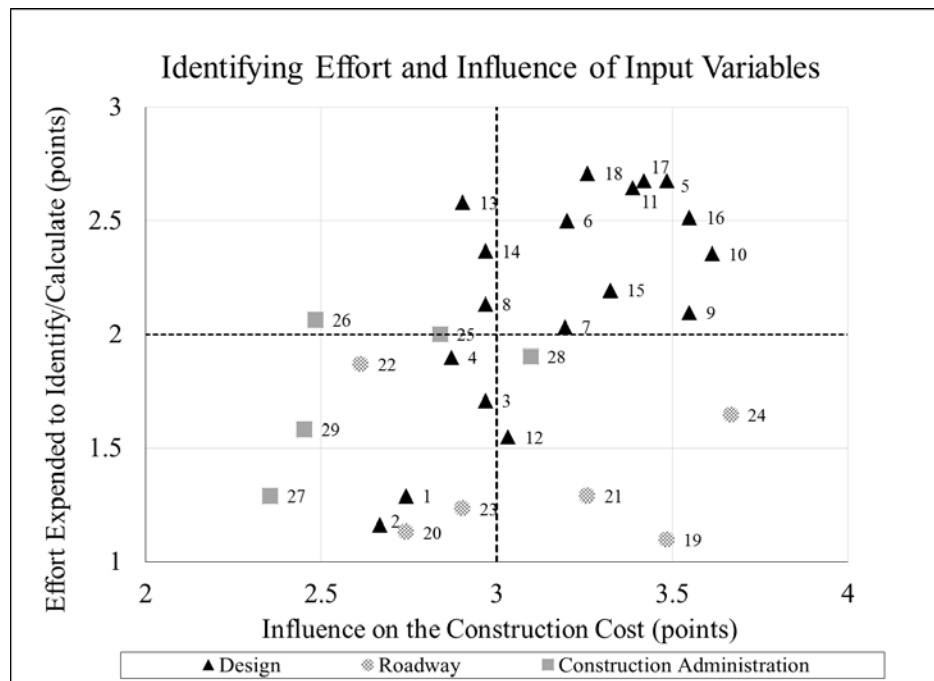
Fig. 7 Results of MDT cost estimating survey

Visually, there are a number of interesting results which can be observed in Figure 7. Firstly, only 5 of the 29 attributes shown in Table 2 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. There was only one design factor identified in the bottom right quadrant.

Secondly, all the attributes in the top right quadrant of the Figure 7 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

IOWA STATE UNIVERSITY
Institute for Transportation

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 attribute (as shown in Figure 8).The formula to calculate each distance was based on the Euclidean distance, and shown in Equation 1 (Danielsson 1980).

$$Distance\ to\ ideal\ input\ (points) = \sqrt{(x_i - A)^2 + (y_i - B)^2} \tag{1}$$

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 Figure 8
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 Figure 8
i = the input attribute being measured, ranges from 1 to 29
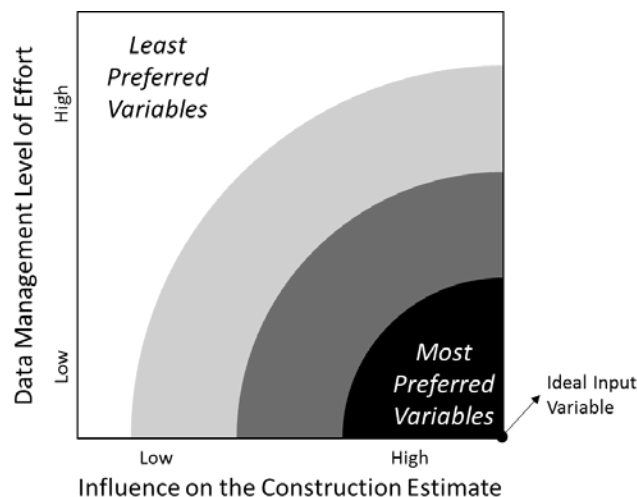
Fig. 8. Preference for selecting input variables

A total of 189 pavement preservation projects were provided to the research team from MDT. The projects were made available from existing databases and conceptual project reports completed during the planning phase of each project. The research team then compiled a database which included as many of the 29 input variables for the ANN and MRA model as possible. Because the survey was created for generic project types, some of the project attributes were not relevant to pavement preservation projects. For the purposes of this case study 13 input variables relevant to pavement preservation were selected. These were selected based on guidance from MDT

IOWA STATE UNIVERSITY
Institute for Transportation

personnel and are shown in Table 3. An example is the exclusion of right-of-way acquisition costs, it was determined at MDT to be very unlikely that pavement preservation projects would involve such occurrence.

Table 3. 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 1
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 and 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 and 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 3 along with the calculated distance to the 'ideal input variable' shown in Figure 8. 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 152 projects were used to train each model and the same 38 projects were used to test the model and calculate the MAPE. The randomly selected 38 test projects accounted for 20% of the database, this proportion of testing to training data was based on previous literature (Petroutsatou et al. 2012; Moselhi and Siqueira 1998).

IOWA STATE UNIVERSITY
Institute for Transportation

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 8 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 9 illustrates the results of each approach.
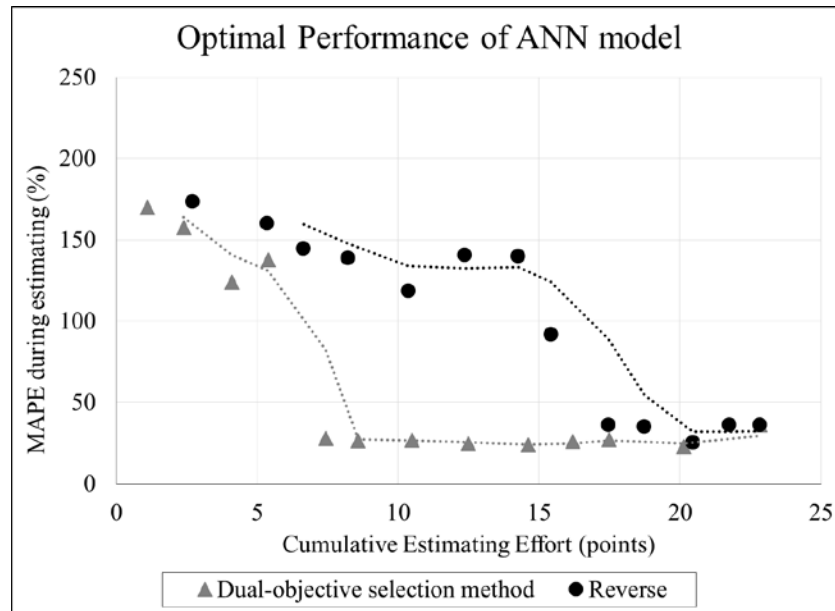
Fig. 9. ANN performance and effort expended

Figure 9 shows that when input variables are added in the order suggested by this research the model can more quickly reach reasonable accuracy with less effort. It also minimizes the number of input variables required to achieve the lowest possible MAPE. The corresponding model reached around 25% 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.

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.

IOWA STATE UNIVERSITY
Institute for Transportation

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 used with commercial software for MRA. 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 10. This helped to validate the selection process. 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 *Guideline 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 research tested the input variable selection framework with both ANN and MRA.
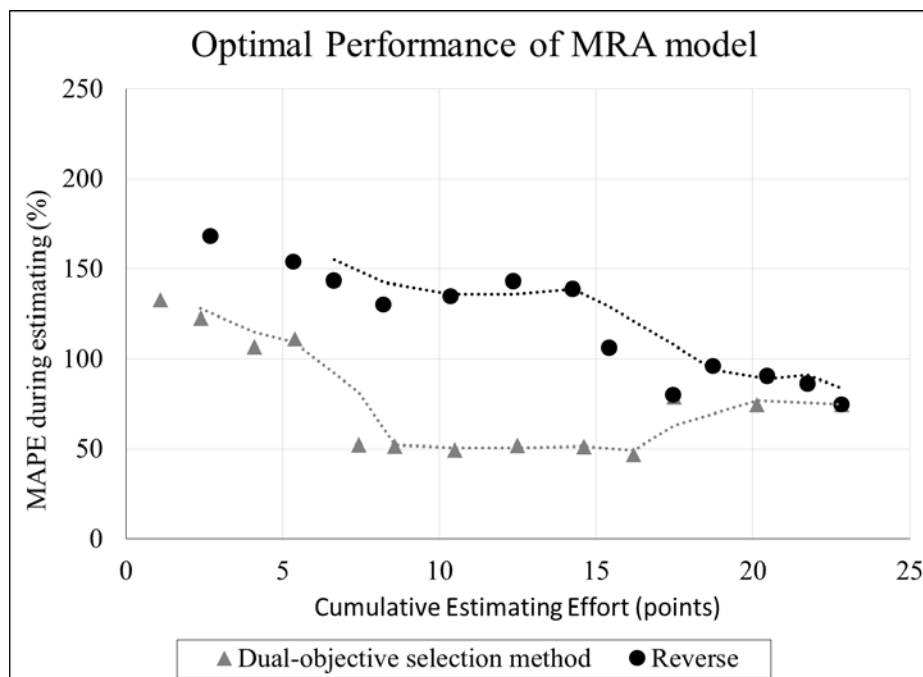
Fig. 10. MRA performance and effort expended

IOWA STATE UNIVERSITY
Institute for Transportation

It is interesting to note that with the MRA model using the reverse order of input variables never reaches the optimal prediction accuracy of around 50%. Also the regression analysis actually performs superior with less input variables and after a point the prediction error 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 obsolete. 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 less effort can achieve satisfactory accuracy at the conceptual estimating stage. 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 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.

IOWA STATE UNIVERSITY
Institute for Transportation

APPENDIX E. RESEARCH PAPER 2

Stochastic Conceptual Cost Estimating of Highway Projects to Communicate Uncertainty using Bootstrap Sampling

Brendon J. Gardner, Douglas D. Gransberg and Jorge A. Rueda. *To be submitted to the ASCE-ASME Journal of Risk and Uncertainty Part A.*

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, the estimate usually deviates substantially from the actual construction cost. When expressed as a deterministic value, an estimate often leads to a false inference of accuracy by those not familiar with the vagaries of conceptual cost estimating, making it difficult for an agency to explain cost growth. Communicating the conceptual estimate stochastically allows an agency to produce a probability distribution of the likely construction costs 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. However, instead of fitting available distributions, this research generates an empirical distribution to express a cost estimate range. Creating empirical distributions eliminates assumptions required for selecting named distributions. The stochastic data-driven model developed in this paper combines artificial neural networks and bootstrap sampling using 189 highway projects to train and test the estimating model.

Keywords: Cost estimating, risk, range estimating, confidence level, stochastic, bootstrap sampling, artificial neural network, cost transparency

Introduction

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 stage, these cost estimates are required by public agencies for statewide fiscal funding requirements (Anderson et al. 2007, FHWA 2015). It is known that many highway agencies experience substantial cost growth from this initial estimate, shown as the conceptual estimate in Figure 1, to the final construction cost (Flyvbjerg et al. 2002; Schexnayder et al. 2003; Chou et al. 2006).
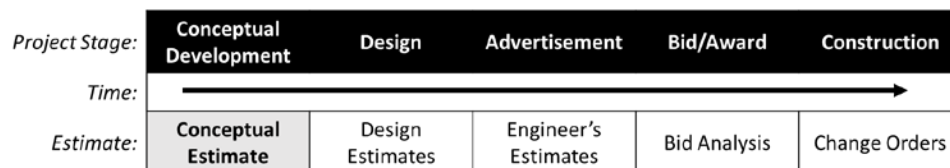
Fig. 1. Construction cost estimating timeline (adapted from Schexnayder et al. 2003)

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. That AASHTO publication also acknowledges the difficulty in quantifying uncertainty associated the cost at the conceptual stage cost. Typically the uncertainty at the conceptual stage is assigned as a percent of the construction costs (Molenaar 2005, Byrnes 2002, Turochy et al. 2001). Byrnes (2002) reported that state highway agencies add a contingency ranging from 5-45% depending on project type and uncertainty; similar contingency factors were also reported by Turochy et al. (2001).

Reflecting the construction cost as a point estimate (i.e. a specific 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. A stochastic cost estimate is a range of costs with probability levels associated with each cost actually occurring.

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 is "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.

Recent literature demonstrates that 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.

Flyvbjerg et al. (2002) 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. included 258 transportation infrastructure projects from different historical periods, geographical regions and project types, with a combined value of $90B. Three main reasons for the statistical significance were investigated; these were: 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 construction cost estimates and increase the level of cost transparency. Using historical project data to forecast costs

IOWA STATE UNIVERSITY
Institute for Transportation

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). 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 assumed 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 development 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,
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.

The research team worked with data provided by the Montana Department of Transportation (MDT) to create a database for estimating the construction cost of pavement preservation projects. Real construction costs from completed projects at a highway agency are used to test and validate the method presented.

IOWA STATE UNIVERSITY
Institute for Transportation

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 for 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 also a holistic approach, however it is directly proportional to the confidence in the conceptual cost estimate.

Due to improved data storage technology, agencies now have the luxury of data from previously constructed projects to assist with assigning the contingencies. This can be done by looking at other similar projects in a 'top-down' method and recognizing the types of projects that exhibit less confidence in the initial estimate; these are higher risk and the contingency assigned should be accordingly reflective. The power of this data has already been realized in its ability to estimate the construction costs of projects. Data-driven models have been created by numerous researchers to conduct construction cost estimates using Artificial Neural Networks (ANN) or Multiple-Regression Analysis (MRA) (Petroutsatou et al. 2012; Kim et al. 2004; Creese and Li 1995). More specifically some data-driven models have been created for highway construction cost estimating at the conceptual stage and suggested for use by highway agencies (Bell and Ghanzanfer 1987; Hegazy and Ayed 1998; Mahamid 2011). These published models use a 'top-down' estimating approach to calculate a point estimate of construction costs.

This research paper leverages the 'top-down' cost estimating approach developed for calculating not only the construction cost, but also an associated contingency based on the risk profile of the

IOWA STATE UNIVERSITY
Institute for Transportation

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.

Data-driven conceptual cost estimating

Two data-driven cost estimating methods, ANN and MRA, have been commonly published in the literature with proven results for estimating the conceptual costs of highway projects (Bell and Ghanzanfer 1987; Hegazy and Ayed 1998; Mahamid 2011). ANNs use pattern recognition to forecast future costs based on the historical database of previously constructed projects (Kim et al. 2004). Multiple-Regression Analysis creates a linear regression equation by assigning weights to particular project attributes through the method of least error (Turochy et al. 2001). Future construction costs are estimated using the assigned weights from the training set of data to complete the equation. Multiple researchers have proven the ability of ANN to produce superior results to MRA in the field of construction cost estimating (Petroutsatou et al. 2012; Kim et al. 2004; Moselhi 1998), some researchers have proven the contrary (Gunduz et al. 2011; Setyawati et al. 2002). This research solely focuses on the ability of ANNs, although the concepts presented could easily be extended to produce stochastic estimates with MRA or for areas outside of highway conceptual cost estimating.

The data-driven models presented in the literature for estimating construction costs generally make a point estimate (Sonmez 2008). As discussed above, and further emphasized by Sonmez, the point estimate provides no information about the level of uncertainty associated with that estimate. Contingency is especially critical at the conceptual stage as there is generally more uncertainty associated with that estimate, when compared to later design estimates (Sonmez 2008, AASHTO 2013). As a result the data-driven model developed as part of research is to be combined with bootstrap sampling to turn the point estimate output into a stochastic range.

Bootstrap sampling method

The bootstrap sampling method provides a simple process to randomly resample an 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 randomly sampling an original data-set, shown in Figure 2. There a two methods to sample the original data-set which is the process labelled '*(a) sampling*' shown in Figure 2 (Efron and Tibshirani 1993; Davison and Hinkley 1997). These two methods are:

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

IOWA STATE UNIVERSITY
Institute for Transportation
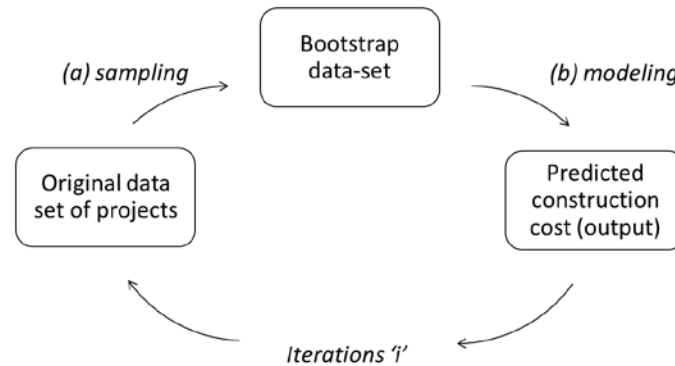
Fig. 2. Bootstrap sampling process (developed from Efron and Tibshirani (1993))

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. In sampling WOR 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 randomly selected 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 contrary. If sampling WR is used then provided that the bootstrap sample is much smaller than the population size (n << N) then the probability of sample repetitions will be small anyway (Efron and Tibshirani 1993). This research tests the sampling WOR method.

Once the bootstrap sample of projects is created, the construction cost (output) can be calculated by modeling, labelled '*(b) modeling*' in Figure 2. Two modelling methods presented above were ANN or MRA to predict the construction cost. Because ANN and MRA are data-driven estimating techniques the output will vary with the input of projects selected in the randomly selected bootstrap sample. 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 2. Iterating the bootstrap process 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

IOWA STATE UNIVERSITY
Institute for Transportation

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 study makes use of the bootstrap approach to create a stochastic cost estimate, the details of which are covered in the methodology 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 in this paper instead focuses on typical projects for highway agencies and should not require the employment of a specialist to manage.

Sonmez (2008) used bootstrap sampling WR to calculate a probabilistic conceptual cost estimate of a building project. The number of projects used to train the MRA model was 19 (N=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 (n=20). Because the bootstrap sample was larger than the number of training projects available (n>N) then sampling WR was used. Sonmez stated that further studies should include larger data-sets, this paper contributes to the limitation outlined by Sonmez through the use of 151 projects in the training database as opposed to 19.

In other fields, researchers used the bootstrap procedure to represent uncertainty for incremental cost-effectiveness ratios for endoscopy clinical procedures (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 to simulate outcomes to produce probability in a commercial spreadsheet. In 2004 Sonmez used this approach to create a range estimate using normal

IOWA STATE UNIVERSITY
Institute for Transportation

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 the stochastic output with a point estimate, then both methods of estimating construction costs were completed using a database of 189 projects. Development of the ANN model is described in more detail later in this paper (Results I). The differences in the two estimating models are shown in Table 1.

Table 1. ANN model arrangement for the 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

For the point estimate, shown in Table 1, an ANN model was developed using 80% of all possible 189 projects for the training data-set (151 training projects). Then, this model was tested using the remaining 38 projects (38 testing projects). The same training and testing data-sets were next used to create the stochastic estimating model. No adjustment to the ANN model architecture, input attributes or modeling software were made in order to create the stochastic output; the only exception being the projects used to train the ANN model, these projects were randomly selected through the bootstrap sampling process introduced in the background section. As shown in Table 1, a total of 121 projects were randomly selected for each of the 100 bootstrap samples. The three main steps taken to create the point estimate and stochastic estimate output are summarized:

1. ANN data-driven model created to predict construction cost as a point estimate for 38 test projects.
2. Stochastic estimating model created using the base estimating model from Step 1, bootstrap samples of 121 randomly selected projects were used instead of the entire training set of data. Bootstrap sampling WOR was completed with f=0.8 (i.e. 80% of the database randomly selected in each bootstrap sample). A total of 100 iterations were completed producing 100 point estimates from the bootstrap samples. The combination of these formed the stochastic estimate.

IOWA STATE UNIVERSITY
Institute for Transportation

3. Point estimate and stochastic estimates were compared.

The error in the point estimate was calculated using the Mean Average Percentage Error (MAPE). This method is traditionally used by authors of data-driven conceptual estimating models (Petroutsatou et al. 2012; Gunduz et al. 2011; Mahamid 2011; Hegazy and Ayed 1998). Calculation of the MAPE is furnished using Equation 1 (Mahamid 2011).

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

where:
n = Number of data-points in the testing data-set
P_i = Predicted construction cost for the ith project
A_i = Actual construction cost for the ith project

The performance of the range estimate could not be measured using the MAPE (Eq. 1) 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 of the confidence interval, this validation technique was summarized in Table 1.

The performance of each estimating method was measured by comparing the actual construction cost to the predicted. However, comparing the MAPE from the point estimate model and results from the range estimate model were difficult. As such the research team qualitatively assessed the ability to communicate the individual uncertainty associated with each project from the point estimate output to that of the range estimate output, this was Step 3.

Data Analysis and Results

The results section is divided into two parts. In the first part (Results I) the base ANN model is developed and results surrounding the point estimate shared. That section also includes the method used to select the input variables for the model and how the database was created. In the second part, the point estimate is further developed into a stochastic estimating model (Results II).

Results I: Point estimating model

A total of 189 projects were made available to the research team from MDT for analysis in various databases and project report formats. The databases and project documents included all highway projects completed from 2009 until 2013. The projects were pavement preservation with a predominant work-type of chip seal, thin lift overlay or mill and fill.

The authors conducted two days of interviews at MDT to establish the main cost attributes (inputs) which could best predict the construction cost at the conceptual stage. Studying previous literature on data-driven conceptual cost estimating 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. From a review of inputs in those literatures and interviews at MDT a total of 17 cost influencing attributes were deemed most relevant to pavement preservation projects. These attributes are shown in Table 2.

IOWA STATE UNIVERSITY
Institute for Transportation

Table 2. Input variables used in the database

Input Variables trialed in ANN model	
Urban/Rural Indicator*	Letting date
Construction on Native American Reservations*	Typical Section (depths of surfacing and aggregate) *
Design AADT*	Curb and Gutter and Sidewalk*
Design speed	Geotechnical recommendations*
Site Topography*	Bridge deck areas*
Start and End Stations, Length and Width*	Traffic Control – closures or detours*
Number of bridges in scope*	Right-of-way acquisition and costs
Intersection signalization and signage*	Extent of Utility relocations and costs*
Contract time*	

*denotes input variables which were included in final model through trial and error

The database of 189 projects, with the input variables from Table 2, were split into two groups; training and testing. In the published literature typically 20-30% of the data is used to test and validate the model (Petroutsatou et al. 2012; Moselhi 1998). For this research 20% of projects, which accounted to 38 projects, were retained for testing the performance of the prediction capability. The split between the number of projects used to train the ANN model and the testing projects was highlighted earlier in Table 1.

Actual project construction costs required inflation to a base year to reflect the rising construction costs. The data was collected for projects over a construction period of 5 years (2009-2013). An inflation factor of 3% per annum was applied to all projects from the expected mid-point of construction to align with the year 2014 (base-reference). The 3% inflation rate was selected based on the historical average inflation rate for projects at MDT and advice from meetings.

The database was organized in a commercial spreadsheet with 17 input variables shown in Table 2. A base ANN was created using a common add-on to that software. Trial and error was used to determine which combination of input variables most accurately predicted the construction cost. This technique is commonly used in literature (Bell and Ghanzanfer 1987; Creese and Li 1995; Hegazy and Ayed 1998; Gunduz et al. 2011; Petroutsatou et al. 2012). Table 2 denotes the final 14 input variables used in the ANN prediction model.

The 14 input attributes from the 151 training 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 38 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 3. It should be noted that a straight line with a slope of

IOWA STATE UNIVERSITY
Institute for Transportation

1 (45 degree angle) passing thorugh the origin represents the point where the predicted construction costs is exactly equal to the actual construction cost.
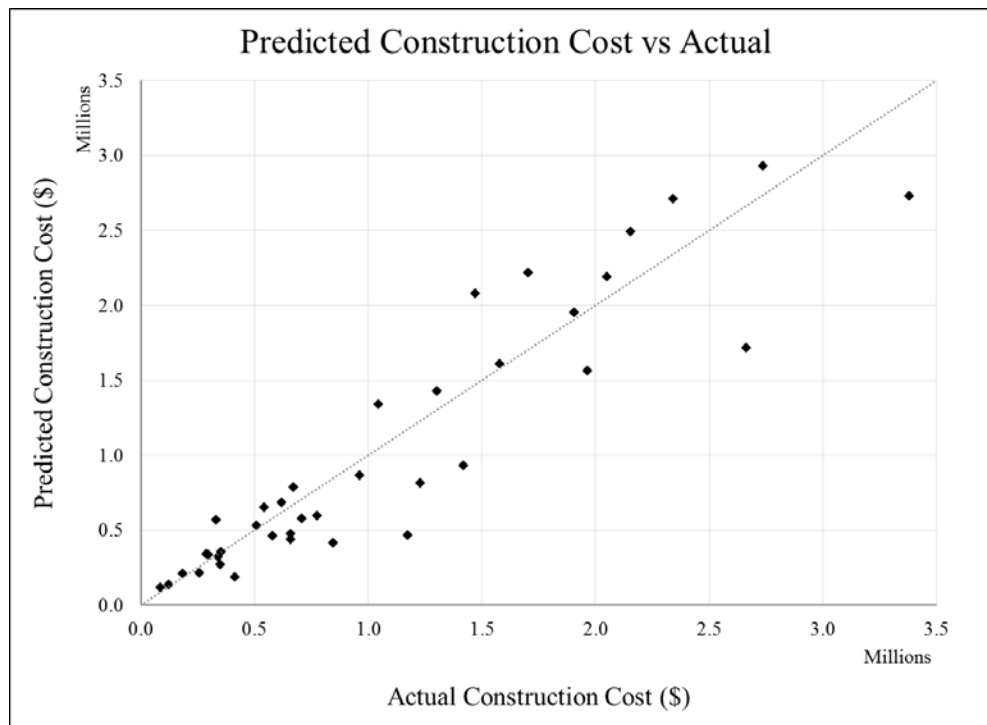
Fig. 3. Visual representation of the ANN prediction modeling tool

The performance of the ANN model was calculated using MAPE, shown in Eq. 1. This was calculated to be 23% and shown in Table 3, well within the recommended performance in the AASHTO *Practical Guide to Cost Estimating* (2013) at the conceptual stage. The error from each of the individual 38 projects are shown in Table 3, these errors are averaged to calculate the MAPE. 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. However, 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 3. Hence, this finding shows that the MAPE does not reflect the confidence of each individual project. The model produced in this paper much more confidently predicts the construction costs of some projects when compared to others. The stochastic estimating method produced in the following section creates individual contingencies for each project based on the confidence in that project and associated data.

Table 3. Point estimate from the model versus the 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

IOWA STATE UNIVERSITY
Institute for Transportation

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 using Equation 1)		22.9%	

Results II: Stochastic estimating model

Range estimate results for all 38 test projects are shown in Table 4. These 38 test projects are the same projects used to test the point estimate model (Results I). The minimum and maximum values

IOWA STATE UNIVERSITY
Institute for Transportation

were the two extreme values predicted during the 100 iterations of this estimating approach. The probability levels 5%, 15%, 85% and 95% indicate probabilities that costs will be below that value.

To highlight the ability to use confidence intervals an example is project 7907 shown in Table 4, our model predicts a 90% confidence interval that the cost will range between $1,406,550 and $2,870,946. Those values are the 5% and 95% probability level costs, thus the subtraction of the two probability levels indicates a 90% chance range. Similarly, the 15% and 85% probability level costs would provide a 70% confidence interval of construction costs. For project 7907 the actual construction cost of $2,049,786 lies within the both the narrower 70% and wider 90% confidence interval predicted by the model.

Table 4. 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

IOWA STATE UNIVERSITY
Institute for Transportation

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

From the stochastic estimating results in Table 4 there are some interesting outcomes:
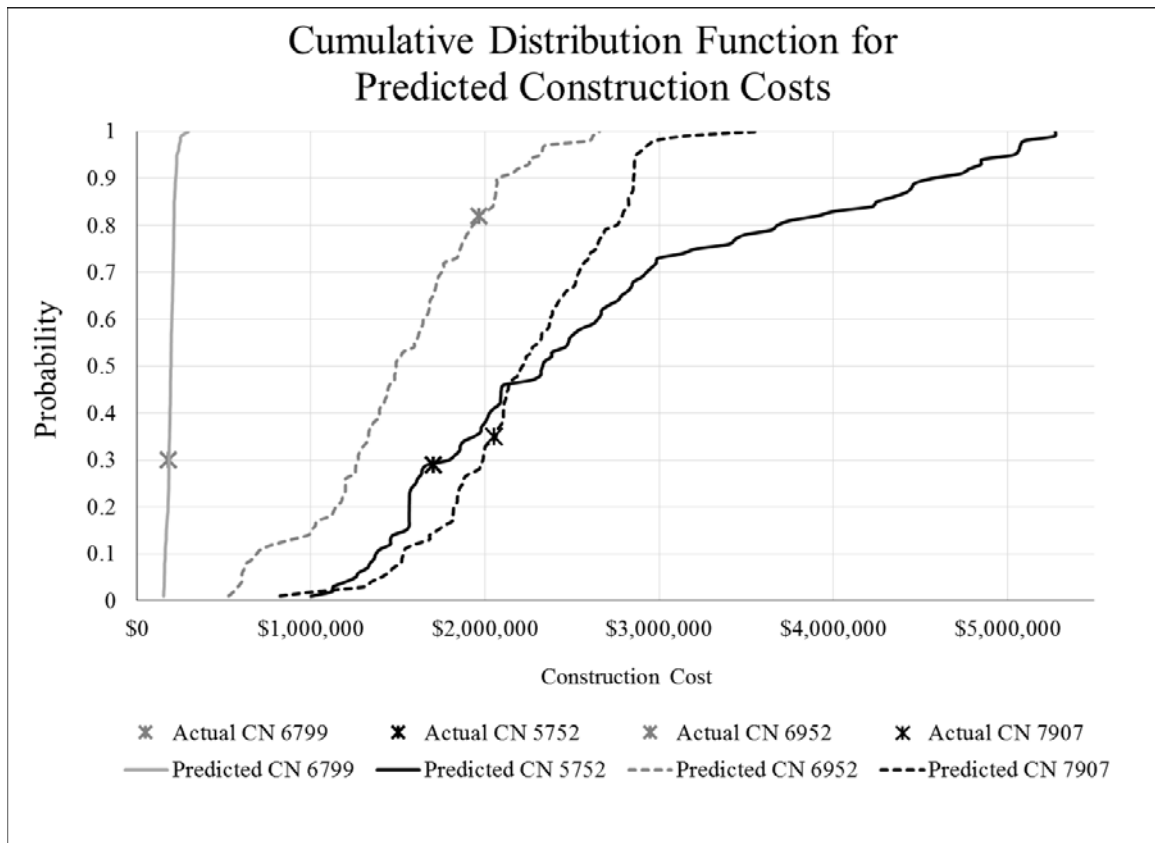
- 35 of the 38 test projects fall within the minimum and maximum expected extremes predicted throughout the 100 bootstrap samples.
- 27 of the 38 test projects fall within the 5% and 95% expected cost.
- 18 of the 38 test projects fall within the 15% and 85% expected cost.

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

Figure 4 displays the stochastic estimate for four selected projects. Project 6799 is a chip-seal project and is known to a very high degree of certainty. This is shown in Figure 4 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 these two projects have parallel confidence intervals, the cost of project 7907 is higher in both the predicted estimate and actual cost 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.

IOWA STATE UNIVERSITY
Institute for Transportation

Fig. 4. Visual representation of estimate confidence for four of the 38 test projects

The actual construction costs for each project, shown in Figure 4, fall within the confidence intervals for their respective ranges predicted with the model. The four plots in Figure 4 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 disadvantage of 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.

IOWA STATE UNIVERSITY
Institute for Transportation

- 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 named distribution could not easily be fitted to all projects to assess their confidence levels.

This research presented here is an example of how a highway agency could embrace 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 in this study 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. All bootstrap samples were randomly selected. Future studies should increase the number of iterations and the size of the data-base for higher confidence in the results. Additionally, investigation into optimal bootstrap sampling techniques could be conducted, this includes the sampling fraction used and comparison of sampling WOR to WR could be investigated.

Conclusion

Point estimates are single numbers with no indication of the level of confidence with which they have been developed. In later estimating stages, when quantities are known, highway agencies can be more confident and can express the estimate in that form. However, for the earlier estimate stages, where project scope is less developed, the estimate should be expressed in a manner that describes the estimator's confidence; providing a range does just that. The communication of estimate confidence through a range could help remove optimism and bias inherent with conceptual cost estimates. Additionally, the power of developing an empirical distribution for individual projects highlights a method that highway agencies can use to assign contingency. The findings of 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.

IOWA STATE UNIVERSITY
Institute for Transportation

APPENDIX F. RESEARCH PAPER 3

Rationally Selecting data for Highway Construction Cost Estimating at the Conceptual Stage

Brendon J. Gardner, Douglas D. Gransberg and H. David Jeong. *To be submitted to the ASCE Journal of Computing*

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 Departments of Transportation (DOT). This paper investigates how data point selection limits the practical performance of these models and why this is a reason they 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 needs 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.

IOWA STATE UNIVERSITY
Institute for Transportation

Introduction

Estimating construction costs at the conceptual stages of project development is critical for decision-makers to determine a reasonable project budget and to make decisions regarding the project's ultimate feasibility (Harbuck 2007; Lowe 2006; AASHTO 2013). In addition, DOTs need reasonable accuracy in estimating conceptual construction costs to ensure that tentative construction programs optimize available fiscal year funding. Conceptual cost estimating (CCE) is defined in this research as the construction estimate during early scoping when little project definition is available (AASHTO 2013).

Under-estimating 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 Cost Estimating Procedures 2007).

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's work observed that up to a point, the greater the design budget the lower the construction cost growth from its initial estimate. Thus, underfunding the design budget yields the potential for construction budget overruns. The design budget, a major portion of the preconstruction budget, is typically established as a percentage of estimated construction costs (Jeong and Woldensenbet 2012). 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.

A problem during the CCE stage is the "limited information" known about the project during the planning stage (AASHTO 2003; AASHTO 2013). Importantly, it is at the CCE stage where designers have the most influence on the end project cost. This introduces the "cost estimating dilemma" suggested by Becker in 1990 (Figure 1). Confidence in CCE enables designers to alter designs and realize savings when they have the ability to influence the cost of the project. The cost of construction is "impacted significantly by decisions made at the design stage" (Gunaydin 2004).
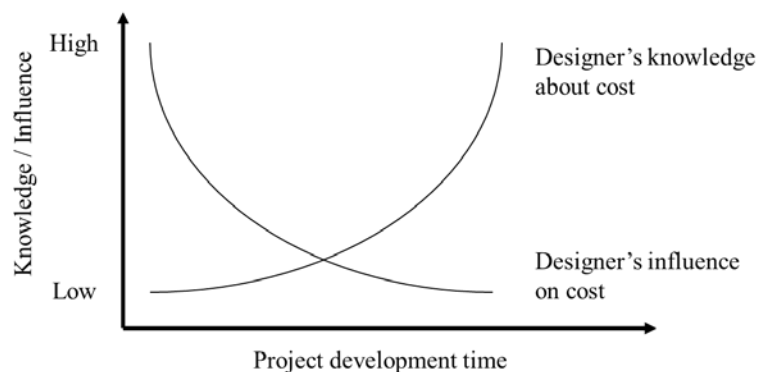
FIGURE 1 Cost estimating dilemma (adapted from Bode 2000).

Background

CCE techniques recommended for use by DOTs (AASHTO 2013) are calculated through statistical relationships between project definition and historic costs. A survey into the current practices of CCE methods at DOTs was conducted by Turochy et al (2001) for Virginia DOT. This was in response to "attention from news media and elected officials" due to major increases in highway project cost estimates since the planning stage. The responses from nine DOTs found the methods generally fell into three categories

1. "cost-per-mile" of typical sections
2. estimating "rough" quantities of the major work items, and
3. no documented or uniform method at all such as the use of experience and engineering judgement.

The same practices were discovered by Byrnes (2002) when he surveyed all 50 DOTs. It was found in both studies that no State DOTs are employing sophisticated mathematical models suggested in the literature.

The advancement in digital technology and data storage capacity has meant that DOTs have an abundance of data available from past projects for analysis and use for estimating future projects costs. Two data-driven techniques popular in CCE literature are artificial neural networks (ANNs) and multiple regression analyses (MRA) of which there have been numerous publications over the past two decades (Petroutsatou 2012; Gunduz 2011). This research specifically focuses on these two techniques referred to as data-driven CCE models from here on throughout this paper.

MRA is the development of a linear equation to link independent project variables to the cost (Turochy 2001). The equation assigns weights to each of the independent data-points to best link the contribution of each variable to the construction cost with the least amount of error. Future construction costs can be estimated using the same equation weights but with the new independent variables.

ANNs do not require knowledge of the link between the construction cost and the variables (Kim 2004). The model uses artificial intelligence to find patterns within the data-base 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 the patterns, these patterns can then be recognized in the new data for forecasting the dependent variable.

Bell and Ghanzanfer published one of the first MRA models for predicting the cost of highway construction maintenance projects in 1987 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 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).

Since Bell and Ghanzanfer 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 published an MRA model with only a 6% error on test projects. Creese in 1995 published an ANN model with 8.24% estimating error for the construction costs of timber bridges.

IOWA STATE UNIVERSITY
Institute for Transportation

In 1998 Hegazy 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 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 data-base, 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.

Despite these 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 2006; Byrnes 2002; Walton 1997). Turochy et al (2001) concluded that DOTs are not employing computer model techniques to improve 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 computer estimation is the ability to remove bias and possible pressure to keep estimates under published budget ceilings, a challenge estimators regularly face (NCHRP report 574). Flyvbjerb et al (2002) discovered that to enable construction to proceed, underestimation is the rule rather than exception for transport infrastructure projects. 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, rather than emotion.

Literature supports the case that more data in the prediction model results in improved reliability and accuracy. When Bell and Ghazanfer created a highway data-driven CCE model with MRA using 174 projects their research concluded in 1987 - "larger data sets tend to reinforce the reliability of the model". This judgement is supported by many authors of data-driven CCE models (Setwayati 2002, Gunaydin 2004, Tatari 2010 and Gunduz 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 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 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 in 2004 and Lowe in 2006 used 530 and 286 historical projects respectively for their data-bases.

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 used 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

IOWA STATE UNIVERSITY
Institute for Transportation

has on model accuracy. More specifically, 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 organize the publications 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 accuracy 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 1 and outline the brief scope for the types of projects selected. 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 advantages and disadvantages of the two different CCE modelling techniques is not discussed here. However, it is noted that some developers of data-driven CCE models concluded that ANN performs superior when compared to MRA for accuracy (Petroutsatou 2012; Kim 2004; Moselhi 1998), others determined the contrary (Gunduz 2011; Setyawati 2002). All authors were however unanimous in agreeing that ANN and MRA techniques are promising to complete CCE going forward in in the construction industry.

The error in the data-driven CCE models collected for comparison was calculated using the same method, Mean Average Percentage Error (MAPE) of the testing data. This method is commonly used by authors of data-driven CCE models (Petroutsatou 2012; Gunduz 2011; Mahamid 2011; Lowe 2006; Kim 2004; Gunaydin 2004; Emsley 2002; Setyawati 2002; Al-Tahbai 1999; Elhag 1998; Hegazy 1998). Calculation of the MAPE is furnished using Equation 1 (Mahamid 2011). Where authors had not used this method our research team recalculated the error to enable direct comparison.

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

where:

IOWA STATE UNIVERSITY
Institute for Transportation

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

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

A_i = Actual construction cost from the historical records collected

**TABLE 7 Construction cost estimating models studied**

CCE literature	Year published	Data points	ANN estimating error	MRA estimating error	Brief Project Scope
Petroutsatou et al. (*18*)	2012	149	4.65%	–	Tunnels in Greece
Mahamid (*29*)	2011	131	–	13.0%	Highway (various sizes)
Gunduz et al. (*17*)	2011	16	5.76%	2.32%	Light rail track works in Turkey
Lowe et al. (*3*)	2006	286	–	19.30%	Buildings in UK
Petroutsatou et al. (*31*)	2006	149	–	9.6%	Tunnels in Greece
Kim et al. (*12*)	2004	530	3.0%	7.0%	Residential Buildings in Seoul, Korea
Gunaydin and Dogan (*9*)	2004	30	7.0%	–	RC 4-8 story residential buildings in Turkey (limited to structural skeleton)
Emsley et al. (*26*)	2002	288	16.6%	–	Buildings
Setyawati et al. (*23*)	2002	41	13.4%	9.2%	Education Building Construction
Al-Tahtabai et al. (*30*)	1999	40	9.1%	–	Highway Construction
Hegazy and Ayed (*16*)	1998	18	19.33%	–	Highway Construction in Newfoundland, Canada
Elhag and Boussabaine (*25*)	1998	30	17.80%	–	School Construction
Moselhi and Siqueira (*28*)	1998	34	10.77%	14.76%	'Typical' steel framed low-rise buildings
Creese and Li (*15*)	1995	12	8.24%	–	Timber Bridges
Sanders et al. (*14*)	1992	11	–	6.0%	Urban Highway Bridge widening in Alabama
Bell and Ghazanfer (*13*)	1987	174	–	17.0%	Highway Construction Maintenance projects

– = data not applicable to that publication

Since Bell and Ghazanfer 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

**IOWA STATE UNIVERSITY**
Institute for Transportation

if not more training data to be used in the data-driven CCE models for reliability and confidence. Figure 2 shows that only three authors in the study population used more than the 174 historical construction projects that Bell and Ghanzanfer 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 Ghanzanfer in 1987, yet there are still many published models using very few historical construction projects in their ANN or MRA analysis.
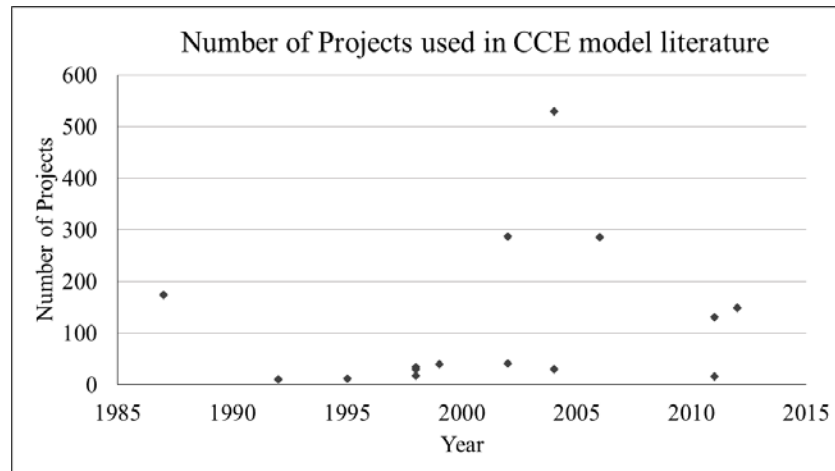
FIGURE 2 Timeline showing the number of historical projects used in data-driven CCE models.

Literature from data-driven CCE models support the hypothesis that lack of data will result in unreliable CCE (Bell and Ghanzanfer 1987; Elhag 1998; Setwayati 2002; Gunaydin 2004; Tatari 2010; Gunduz 2011) and could therefore be a reason for limited industry use. However, findings from the quantitative 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 3 there is little to no trend. The arrow shows the direction of the trend expected from literature findings. 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.
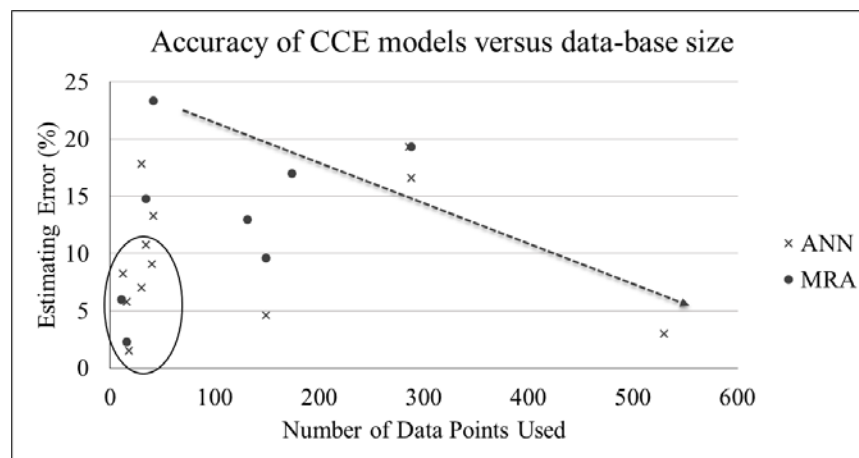
FIGURE 3 Accuracy of data-driven CCE models published and the number of data points used.

Literature suggests that increasing the data-base within the CCE models will result in improved reliability and accuracy (Bell and Ghanzanfer 1987; Elhag 1998; Setwayati 2002; Gunaydin 2004; Tatari 2010; Gunduz 2011). This is conflicting with results from the quantitative content analysis shown in Figure 3. An explanation for this could be that these data-driven CCE models have been built for projects of very narrow scope.

Creese (1995) created a model specifically for timber bridges using only 12 projects. Sanders (1992) limited scope of their data-driven CCE model to bridge widening only, using 11 projects. Sanders 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 (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 light rail projects only and did not account for other infrastructure in the project (Gunduz 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 Ghanzanfer 1987; Setwayati 2002; Gunaydin 2004; Tatari 2010; Gunduz 2011). However a quantitative 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.

IOWA STATE UNIVERSITY
**Institute for Transportation**

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 applicability 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 4. 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; Fink and Fowler).

| 1. Define the population | ⇨ | 2. Distribute key input attributes | ⇨ | 3. Represent the population in the sample |
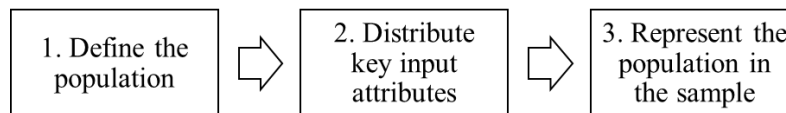
FIGURE 4 Proposed rational sampling steps.

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 Montana DOT for analysis. This data-base 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 and fill projects, the three main major work-types, all less than $5M in value. A total of 226 projects remained for analysis – this was our research population of data.

Step 2: Distribute key input attributes: The database was organized in a commercial spreadsheet with nine input variables shown in Table 2. A base ANN was created using a common add-on to that software. The ANN software was used to test which of the nine input variables were key influencers of the construction costs. This was completed by constructing an ANN model to predict the construction cost using the entire population of 226 projects. A randomly selected 20% of the data was used as a test set, in the published literature this is usually selected as between 20-30% (Petroutsatou 2012; Moselhi 1998). Prediction of the construction cost was found for this database to be most sensitive to two main input attributes: the highway classification and the length of the project. The distribution of these two attributes across all 226 projects was analyzed visually so that when samples were taken within the population they could be appropriately represented.

TABLE 2 Input variables used

| Highway Classification*	Surface Type
Let Quarter	Urban/Rural Indicator
Contract Time	Length*
Location (District)	Roadway Width
Scope	Total Construction Cost [Output]

*denotes attributes analyzed

Step 3: Represent the population in the sample: A test sample was collected first to separate it from any data used to train the model, not doing so would undermine testing results of the data-driven CCE model. A test set of 57 projects was nominated in this example (25% of the total population), leaving a possible 169 projects to train a model. The 57 projects were selected and removed for testing by iteratively selecting projects until distribution of the two attributes aligned with the distribution in the entire population. Selecting test data reflective of the population will test the true performance of the cost estimating model against its intended end-use. Selecting test data with this consideration has been previously ignored in data-driven CCE models studied in this research.

Next, a control sample of 85 projects were selected from the remaining 169 available projects in the training data-set. This was completed with the same method of selection as the test data, by iteratively selecting projects until the distribution of highway classification and length

IOWA STATE UNIVERSITY
**Institute for Transportation**

matched that of the population. The distribution of projects representing each attribute for the population, control sample and test sample is shown in Figure 5a.

For the purposes of validating this method two additional samples of 85 projects were selected from the 169 possible training projects. In each of the samples one of the attributes was improperly represented relative to the control sample. The highway classification was misrepresented in Figure 5b (Sample I) and the lengths of the projects were misrepresented in Figure 5c (Sample II) relative to the control sample.

(a) Sampling Control and Test Data from the Population

(b) Control Sample vs Sample I

(c) Control Sample vs Sample II
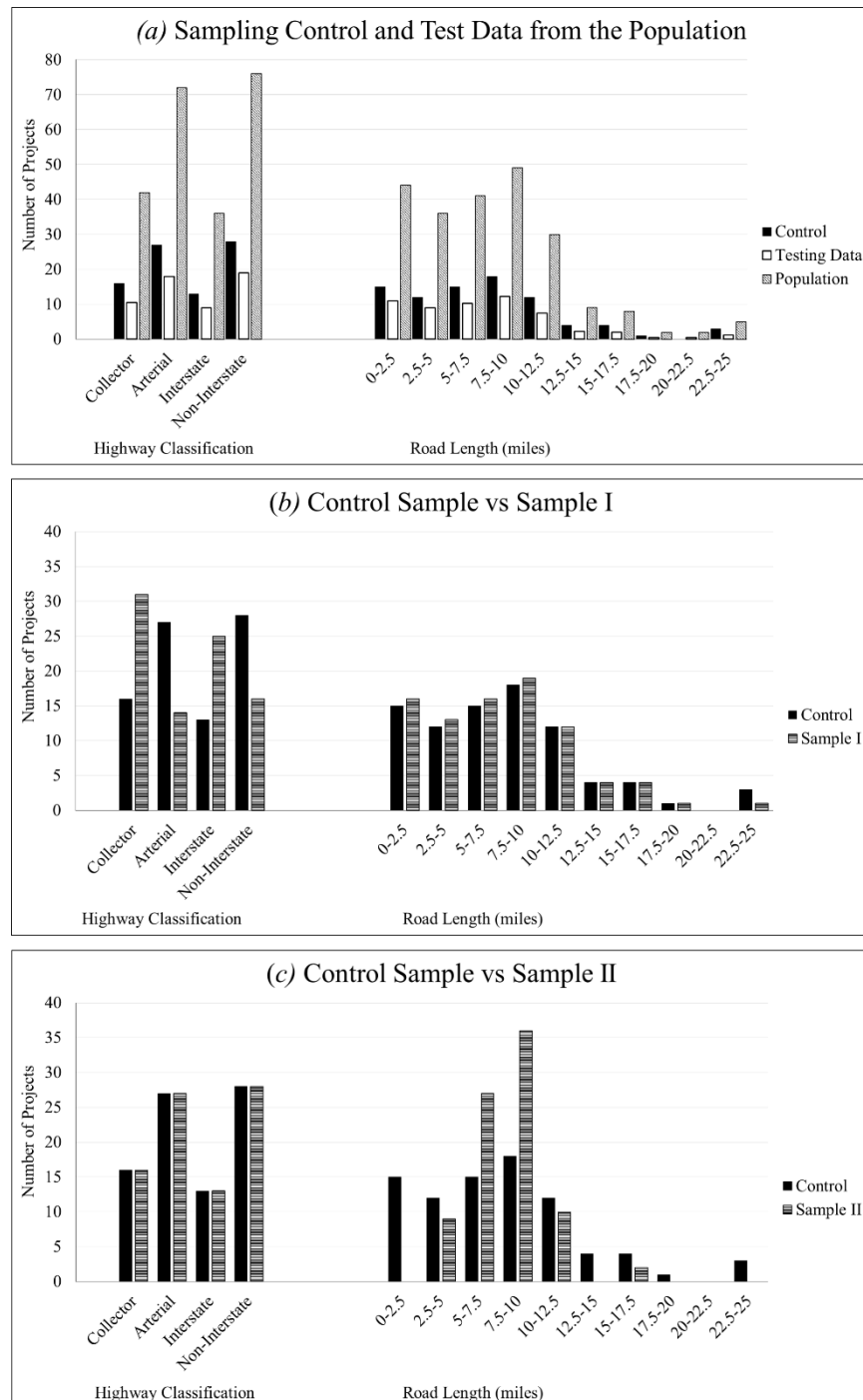
FIGURE 5 Distribution of Sample I, II and III based on three key input variables.

Results: The nine attributes from the 169 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 trained data is shown in Figure 6a. The 57 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 57 test data-points is shown in Figure 6b.
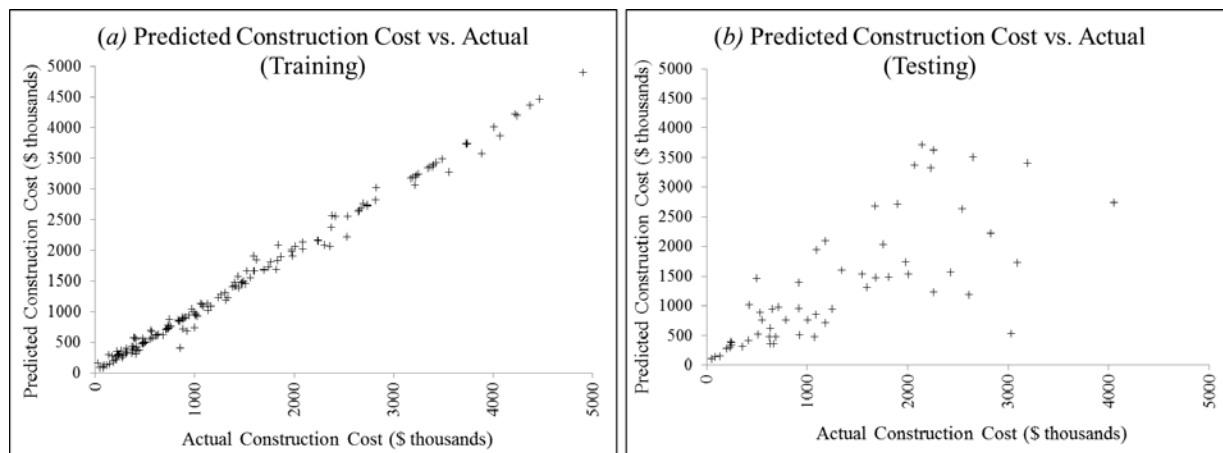
FIGURE6 (*a*) Training the artificial neural network (*b*) Validating the artificial neural network with the test data.

The MAPE for the test 57 projects was 41.6%. Further improving the accuracy of this model was not the goal here, so research into sampling this population of 169 training projects continued. A separate model was trained and tested for the Control Sample, Sample I and Sample II. The same 57 projects were used to the test the error of these trained models.

Results of all four ANN models created are shown in Table 2. It was not surprising that no single sample out-performed using the entire population to predict the construction cost. 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 Ghanzanfer 1987; Setwayati 2002; Gunaydin 2004; Tatari 2010; Gunduz 2011; Rajkumar and Bardina 2003; Bode 2000).

TABLE 2 Error in the testing data

Sample	MAPE with the test data
Entire Population (169 projects)	41.6%
Control Sample (85 projects)	60.6%
Sample I (85 projects)	61.5%

IOWA STATE UNIVERSITY
Institute for Transportation

Sample II (85 projects)	76.3%

It was observed that Sample I performed almost as well as the control sample. This is unusual because the distribution of the highway classifications in the sample did not match that of the population. On the other hand Sample II performed much worse at predicting the construction cost in comparison to Sample I and the Control Sample. On visual inspection of Sample II (Figure 5c) the distribution of length attributes was much more significantly misrepresented than the highway classifications in Sample I (Figure 5b).

This finding suggests that key attributes of the population only need sufficient representation in the sample data-base 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 supports the concept that more data increases the accuracy and reliability of data-driven CCE models (Bell and Ghanzanfer 1987; Setwayati 2002; Gunaydin 2004; Tatari 2010; Gunduz 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 data-bases in the CCE models, 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 data-bases used and no sampling method used for selection of projects.

REFERENCES

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.

Alshanbari, H. (2010). "Impact of Pre-Construction Project Planning on Cost Savings," Master of Science Thesis, University of Florida, 52pp.

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.

Bajaj, A., D.D. Gransberg, and M.D. Grenz, "Parametric Estimating for Design Costs" *2002 Transactions*, AACE, Int'l, Portland, Oregon, June 2002, pp. EST.08,01 – EST.08.06.

Bazeley, P., and L. Richards. (2000). *The NVivo: Qualitative Project Book.* SAGE Publications Inc., California.

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.

Brunsman, A., K.F. Robson, and D.D. Gransberg, ""Parametric Estimating for Environmental Remediation Projects," *2008 Transactions*, AACE, International, Toronto, Ontario, Canada, July 2008, pp EST.06.1-EST.06.8.

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 and 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.

IOWA STATE UNIVERSITY

Institute for Transportation

Deis, D., Schneider, H., Wilmot, C., and Coates C. (2004). "A Simulation Approach To In-House Versus Contracted Out Cost Comparisons," J. of Public Procurement, 4(1), 43–66.

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

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

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.

Eisenhardt, K. M. "Better Stories and Better Constructs: The Case for Rigor and Comparative Logic." Academy of Management Review, 16(3), 1991, pp.620-627.

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.

Government Accounting Office (GAO) Using Structured Interviewing Techniques, GAO/PEMD-10.1.5, Government Accounting Office, Washington, D.C., June 1991, 191pp.

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,* 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.

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.

IOWA STATE UNIVERSITY
Institute for Transportation

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.

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.

Jennings, W. "Why costs overrun: risk, optimism and uncertainty in budgeting for the London 2012 Olympic Games." *Construction Management and Economics,* Vol. 30(6), (2012), pp. 455-462.

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

Janacek, J. (2006). "Construction Costs Going Through the Roof?" Presentation, 2006 *Public Works Officer Institute*, Los Angeles, California.

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.

Mak, S. and D. Picken, Using Risk Analysis to Determine Construction Project Contingencies. American Society of Civil Engineers, *Journal of Construction Engineering and Management*., Vol. 126 Issue 2, 2000, pp.130–136.

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.

Miles, M. B., and Huberman, A. M. "Qualitative Data Analysis: An Expanded Sourcebook." Sage, Thousand Oaks, CA 1994.

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

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

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.

IOWA STATE UNIVERSITY
Institute for Transportation

Montana Department of Transportation. (2007). *Cost Estimation Procedure for Highway Design Projects.*
http://www.mdt.mt.gov/other/roaddesign/external/report_templates_guidance/costest_procedure _jan07rev.pdf. Accessed July 14, 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 J. Bardina. (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.

IOWA STATE UNIVERSITY
Institute for Transportation

Tatari, O., and M. Kucukvar. (2011). "Cost premium prediction of certified green buildings: A neural network approach." *Building and Environment*, 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 J.D. Stevens. Improving Conceptual Estimating Methods Using Historical Cost Data. (1997). In *Transportation Research Record: Journal of the Transportation Research Board, No. 1575,* Transportation Research Board of the National Academies, Washington, D.C., 1997, 127-131.

Yin, R. *Case Study Research: Design and Methods*. Sage, New York. 2008, 176pp.

IOWA STATE UNIVERSITY
Institute for Transportation