CONCEPTUAL COST ESTIMATION MODELS FOR BRIDGE PROJECTS

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Adikie Esinam Essegbey

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Construction Management and Engineering

November 2021

Fargo, North Dakota

North Dakota State University
Graduate School

Title

CONCEPTUAL COST ESTIMATION MODELS FOR BRIDGE PROJECTS

By

Adikie Esinam Essegbey

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Eric Asa
Chair

Dr. Majura Selekwa

Dr. Abdul-Aziz Banawi

Approved:

November 28, 2021
Date

Dr. Xuefeng (Michael) Chu
Department Chair

ABSTRACT

Conceptual cost estimating is typically completed early in the project lifecycle when 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. Therefore, the objective of this research is to develop a conceptual cost estimate model for bridge infrastructure projects. In this study, a systematic literature review and meta-analysis of Mean Average Percentage Errors (MAPEs) of cost models was undertaken to identify the various cost estimation methods, the input variables that have adopted in the development of models and determine the impact of cost estimation method on the accuracy of cost prediction. The research study utilized regression analysis, decision tree and random forest methods for cost prediction of Wisconsin bridges. A comparison of the three models that were developed revealed that random forest cost estimation method yielded better cost prediction.

ACKNOWLEDGMENTS

My utmost gratitude goes to my advisor, Dr. Eric Asa, for his dedicated support and timely guidance during this research. I am sincerely grateful to my committee members, Dr Abdul-Aziz Banawi and Dr. Majura Selekwa who provided advice without reservations. I would like to express my profound thanks to my parents for their endless support, advise and encouragement throughout my masters' degree journey. I would like to thank my other family and friends for encouraging me every step of the way in graduate school. Also, I would like to express my appreciation to department of construction management and engineering of North Dakota State University for giving me the financial support throughout my masters' degree study.

DEDICATION

I dedicate this thesis to God, my family, and friends for their unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

WsDO .....Wisconsin Department of Transportation

MAPE .....Mean Average Percentage Error

ANN.....Artificial Neural Network

MA .....Meta-Analysis

MLR.....Multiple Linear Regression

DT .....Decision Tree

RF.....Random Forest

CHAPTER 1. INTRODUCTION

1.1. Background

Bridges represent a critical component of the United States' transportation infrastructure system. The need for accurate and more reliable cost estimates for transportation infrastructure projects has been more important than ever given the historical overruns of major capital projects (Markiz and Jrade 2014; Winalytra et al. 2018). With more accurate estimates, funding allocations can be proactive and more closely matched with the specific needs of each project (Fragkakis et al. 2011). Cost estimation method enhancements promote fiscal responsibility through improved budgeting.

Conceptual cost estimates are prepared at the early phase of the project; based on a description of the project or on very limited drawings of the project, as such they tend to be least accurate estimates. Typically, the conceptual estimates are used to study the feasibility of a project or to compare potential design alternatives (for example, a concrete structure versus a steel structure or three stories versus four stories (Perterson 2018). In general, conceptual cost estimates are essential to ascertain the viability of a project before it can be initiated. At such early stages of the project, information is least readily available, thus obtaining a reliable cost data might be challenging (Chou et al. 2005). In most cases, conceptual estimates are beset by limited scope definition with high potential for scope changes and also, they tend to be prepared within limited time.

In order to alleviate the uncertainty due to lack of detailed information at the early phase, probabilistic models, cost-based reasoning, machine learning methods such as neural networks, among other have been proposed to predict the cost estimate at the early phases (Kim and Kim 2010). This research study proposes a model to support the estimation of construction costs for

bridge projects in the early stages when the information available is limited using proven simulation models.

1.2. Problem Statement

Establishing reliable construction cost estimates of bridges is quite difficult at the early phases of most projects. Information available for estimating bridge construction cost is normally very limited at this stage e.g., length and width of bridge, number of lanes and others. The accuracy of the conceptual estimate at this phase is critical for decision making process in the construction of such capital projects. Therefore, the importance of early estimates to owners and their project teams at the early stages of projects is critical (Dimitriou et al. 2018). Too low of an estimate may results in project overruns, loss of public trust, and the potential misguide approval of projects that may not be of high priority on the basis of a cost-benefit analysis (Idowu and Lam,2020). On the other hand, too high of an estimate can result in underfunding as well as having insufficient projects and programmed funds to address critical transportation needs in the state.

Generally, cost estimation is integrated with different variables: project complexity, undefined scope of the project, and the uncertainties of underground construction conditions. It is worth noting that prediction models that are proposed should be developed while ascertaining the risks associated with the projects in order to make the model useful. Therefore, it is essential to develop a model that considers risks. Thus, it is from this observation that this research study intends to develop a statistical/ mathematical model for estimating construction costs using categorized variables in a way that ascertain the best estimates of the project costs.

1.3. Research Goal and Objectives

The aim of this research is to develop a conceptual cost estimate model that would reliably predict the cost of bridge infrastructure projects. This will assist the Wisconsin Department of

Transportation (WsDOT) in predicting bridge cost by data driven methods. The detailed objectives to assist in achieving this goal are to:

1. Identify input variables used for developing models that predict bridge project costs.
2. Evaluate the accuracy of previous models through meta-analysis of mean average percentages.
3. Develop a multiple linear regression model, decision tree model and random forest model using significant predictor variables: and compare the performance of the developed models.

1.4. Research Methodology

The research approach consisted of exploratory data analysis, formulation of hypothesis, development of multivariate regression model, decision tree and random forest model. The various stages of the research approach are shown in Figure 1.

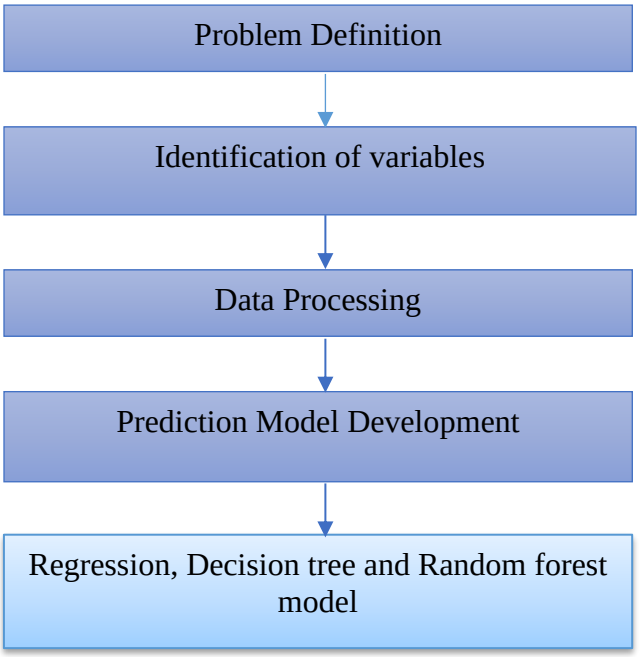


Figure 1: Research Framework

1.5. Research Contribution

At the end, this study provides a probabilistic predictive model as an alternative approach for conceptual construction cost estimation. The predictive model will partly aid in addressing infrastructure funding issues. All of the parties involved in infrastructure construction, from legislators to contractors, will have some level of confidence in the cost estimates developed to ensure efficient and effective funding of infrastructure projects.

Equally, the research can benefit stakeholders at the early stages of a project as the model will assist in improving results of the projects' feasibility studies and leading to better decision making since the accuracy of estimation is a critical factor in the success of construction projects, for which cost overruns are a major problem, especially with the current emphasis on tight budgets.

1.6. Outline of the Thesis

This thesis is made up of five interrelated chapters, each chapter consist of subsections structured to build up the main chapter. This first chapter has presented the background to the research, statement of the problem, research objectives, the research methodology, a summary of the research contributions and the organization of the whole thesis.

The second chapter reviews the existing literature on cost estimation methods for bridges, variables that contribute bridge project costs, estimation methods, theoretical models. Basic descriptive statistics was employed to analyze the estimating variables and meta-analysis was used to compare the accuracy of the models developed by previous authors.

The third chapter assesses the model that was developed by employing multiple linear regression. It addresses multiple linear regression models of the research and gives details about the hypotheses testing, processing of the database, and outlines the steps followed in the development of the model to achieve the research objectives.

The fourth chapter focuses on the development of decision tree and random forest models, and discussion of the results. The fifth chapter which is the final chapter of this thesis presents a summary of the findings from the research, review of the objectives of the study, limitations, and recommendation for further research.

CHAPTER 2. A SYSTEMATIC LITERATURE REVIEW AND META-ANALYSES ON CONCEPTUAL ESTIMATION OF HIGHWAY BRIDGE CONSTRUCTION COSTS

2.1. Abstract

Large sums of capital are currently invested in infrastructure projects in the transport industry. Bridge infrastructure is an essential component and one of the most expensive classes of structures in highway systems. A significant issue for state highway transportation is the challenges related to the accuracy of bridge cost estimates at the conceptual phase of highway bridge construction projects. Even though the construction research sphere is saturated with publications on cost estimation, there is inadequate systematic literature review and meta-analysis of what has been achieved so far in cost estimation of highway bridge construction projects. This chapter synthesizes extant data from previous studies to illustrate the global picture, identify research gaps, and potential benefits to transportation agency stakeholders. A total of twenty-nine papers on cost estimation of bridge projects are systematically reviewed by utilizing content analysis and quantitative data analysis to investigate the frequency of each method over time. Subsequently, a meta-data analysis by Mean Absolute Percentage Error (MAPE) values was performed from 21 studies. The most frequently used cost estimation methods were unit price/ quantity of standard work, regression analysis, and artificial neural networks. A total of 31 input variables were identified. The top three input variables were the weight of steel, the quantity of concrete, and the number of spans. which are materials and design characteristics of the bridge. Results from the meta-analysis indicated that the effect summary of MAPEs was 0.08 with the confidence interval for the cost estimation/MAPE ranging from 0.053 to 0.119, which suggests that the MAPE (cost estimation) in the defined universe could fall anywhere in this range irrespective of the cost estimation method that is adopted. The hypothesis testing result (p-

value<0.05) indicated that the cost estimation method adopted affects the MAPEs. The I^2 statistic of 40.36% showed that the true effect size varies moderately from study to study.

Author Keywords: Conceptual; Cost estimation; Preliminary; Bridges

2.2. Introduction

Cost estimation plays a key role in construction projects. The conceptual cost estimation of bridge projects is regarded as a major activity in the early phase of bridge construction projects. Conceptual cost estimates form the basis for project feasibility studies, serve as an initial budget and financial evaluation tool, enable the comparison of alternative projects, and facilitate excellent and effective decision making at the early stages of projects. The bridge estimating process aims to project, as accurately as possible, the estimated costs for a bridge construction project. Even when grossly inaccurate, early estimates often become the basis upon which all the future projections are judged (Gardner et al. 2016). The accuracy of estimation is a critical factor in the success of any construction project, for which cost overrun is a major problem (Antoniou et al. 2016).

According to Fragkakis et al. (2010), conceptual estimating is done at the initial phase of the project planning process in which limited project information and data are available, and high levels of uncertainty and risk exist. Therefore, accurate conceptual cost estimation is a challenging task, but an essential process used for feasibility and budgeting purposes, the comparison and financial evaluation of alternative projects, and the application of appropriate financing procedures. Elmousalami (2019) opined that a construction project's cost must be estimated within a specified accuracy range. Still, the most significant obstacle of a cost estimate, particularly in the early stage, is the lack of preliminary data, information, conceptual design, and the presence of considerable uncertainties. Consequently, to overcome the challenges with the lack of detailed

information, cost estimation models are used to approximate the cost within an acceptable accuracy range (Chou et al. 2005).

Modern highway bridges play a significant role in the transportation infrastructure. The soaring urban and interurban traffic needs generate an ever-increasing pressure for allocating of funding toward the construction and maintenance of highway infrastructure, including bridges. Flyvbjerg et al. (2003) highlighted that the magnitude of cost overruns in highway bridge projects is substantially high due to several factors including error in estimation, change in designs and project scope, ground conditions, failure to identify and quantify risks, and others. Flyvbjerg (2007) noted that nine out of ten transportation infrastructure projects have cost overrun, for which tunnels and bridges have an average cost overrun of 34%. Flyvbjerg (2007) added that the cost overrun for the San Francisco Oakland Bay Bridge retrofit's cost overrun was more than 100%.

However, there have been several attempts to improve the accuracy of conceptual cost estimation of bridges. Creese and Li (1995) proposed a neural network model for the cost estimation of timber bridges. The estimation accuracy of the neural network method was influenced by historical data. On the other hand, the complex relationship obtained by neural networks makes an explanation of the estimate more difficult. Morcous et al. (2001) used a sample of 22 prestressed concrete bridges in training and testing an artificial neural network. The results showed that ANN is an efficient tool for developing a cost estimation model. Hollar et al. (2010) developed a multilinear regression model with data from 505 North Carolina Department of Transportation (NCDOT) bridge projects awarded for construction from 1999 through 2008. They concluded that, a more accurate prediction of future preliminary costs of engineering projects could be developed by considering numerous parameters.

Furthermore, Fragkakis et al. (2010) developed a conceptual cost estimate model for bridge foundations. They adopted a backward stepwise regression method to derive material estimation models. The coefficient of determination exceeded 77% in all the prediction models, indicating that the proposed models provided a satisfactory and sufficient fit to the data. The estimated p-values and F-values showed that the independent variables and the selected regression models were statistically significant.

Although much effort has been expended on cost estimation methods for bridges from the literature, existing literature reviews are not exhaustive. There has not been a systematic literature review and meta-analysis of conceptual cost estimation of bridges to date. None of these studies provided an illustrative review of what has been achieved in the research domain as well as an extraction and quantitative analysis of data for synthesizing results across the various studies. Adopting a synthesis approach such as a systematic review coupled with meta-analysis of the literature for the research study presents a more realistic and scientific understanding of conceptual estimation of bridge construction projects. In other words, it is required that a narrative summary be followed by a numerical aggregation of the results of various studies to compute the effect summary (Cleophas and Zwinderman 2017). Furthermore, meta-analysis is a process used to integrate the results of various studies to synthesize evidence on a global problem of interest. The advantage of meta-analysis is its transparency in extracting and analyzing knowledge for more accurate decision-making and policy formulation (Cooper et al. 2009). This research aims to employ a systematic literature review and present an overview of existing studies, analyze factors that contribute to cost estimation of bridge projects, and from meta-analysis further examine cost estimation methods and models. This study presents a pool of numerical data from the selected study for a meta-data analysis to synthesize results quantitatively. The study was guided by the

following research questions: (1) what methods have been adopted for the cost estimation of highway bridge projects? (2) what is the trend in the application of cost estimation methods? (3) what are the input variables adopted for cost estimating models/equations of bridge projects? (4) and what is the impact of the estimation methods on the accuracy of cost estimates?

In subsequent sections of this chapter, an overview of cost estimation methods for predicting bridge construction cost and its cost drivers are presented, followed by the research methodology used in this thesis, a discussion of results and finally drawing of conclusions derived from the findings of the study.

2.3. Previous Studies

The estimation process is often carried out during various construction projects and with varying levels of detail and accuracy, depending on the estimation objective. As the project evolves, different types of estimates are usually required. Once the project has advanced to the procurement stage, detailed estimates are typically prepared after completing the detailed project design. Bridge projects have inherent uncertainties and risks; thus, determining final scope and cost bridge project is challenging. Yu et al. (2006) stated that at the preliminary phase of bridge projects, there are occasions where only the basic information such as project size and project scope are known, and estimators have to use the price of similar projects based on historical data and cost predicting models. Kim and Cho (2013) emphasized that an estimate can only be as good as the information it is based on. The level of accuracy of the estimates produced also increases as more information becomes available.

The estimation of the total construction cost of bridges has gained the attention of researchers. Morcous et al. (2001) used a small sample of 22 prestressed concrete bridges constructed in Egypt over the Nile to investigate the cost of prestressed concrete bridges. They

employed ANN with a back-propagation learning algorithm to estimate the concrete volume and prestressing weight in the bridge superstructure. Their testing results indicated that ANNs were sufficient tools for the preliminary quantity estimate of highway bridges as the percentage error was 7.5%. Furthermore, Fragkakis et al. (2010) developed prediction models for the concrete quantities with reinforcing and prestressing steel for three primary bridge deck construction methods using regression analysis. A bootstrap resampling method was used to produce estimate ranges. Similarly, Fragkakis et al. (2011) developed a bridge database with complete data from 157 pier foundations and developed a parametric model for the conceptual cost estimation of concrete bridge foundations. A database with design and structural data for 322 bridge piers was used by Fragkakis et al. (2014) to develop a cost estimate model for piers using regression analysis.

2.4. Methodology

This study was conducted as a systematic literature review and meta-analysis the resulting content- to explore databases to extract pertinent research studies regarding cost estimation for bridge projects. According to Mulow (1994), a systematic literature review is an efficient scientific technique as it is at the top of the evidence level pyramid. Linares-Espinos et al. (2018) added that a systematic literature review minimizes the possibility and extent of biases as the straightforward approach is characterized by a critical investigation, evaluation, and integration of findings of relevant research studies. Kitchenham et al. (2009) asserted that the range of numerous reviewed studies gives an informative context, not obtainable in a single study for implications for practice and policy. This is due to studies addressing similar questions with different eligibility criteria for the various aspects of the research study. A systematic literature review of this kind identifies relations, contradictions, gaps, and inconsistencies in the literature and investigates the underlying reasons (Membah and Asa 2015; Kitchenham et al 2009). The systematic literature review

methodology suggested by Budgen and Breton (2006) was adopted and it is presented in next several pages.

2.4.1. Research questions

To address the main objectives of this research, the questions addressed in this chapter are as follows:

RQ1: What methods were used in the past for the cost estimation of bridge projects?

RQ2: What is the publication trend of bridge cost estimation methods?

RQ3: What are the input variables adopted for cost estimating models/equations of bridge projects?

RQ4: What is the impact of the estimation method used on the accuracy of the cost estimates?

To address RQ1, a number of published papers on the various cost estimation methods for bridges were identified and reviewed. With respect to RQ2, bridge cost estimation publications per year were identified and plotted graphically to depict the trend of publications on cost estimation of bridges. For RQ3, the cost estimation drivers identified in the models used in the various publications were compiled and ranked to determine the frequency of the common cost input variables. Finally, in order to address RQ4, the meta-analysis approach was utilized. The mean absolute percentage error (MAPE) of the model/equation of the author's proposed models was put together and analyzed to determine the accuracy across the various studies' different cost estimations.

2.4.2. Search process

The search process considered the title, abstract, keywords, content, and conclusion of each research paper. The search strategies were broadened to ensure that relevant articles were not

overlooked. The primary terms used were "conceptual cost estimation" or "preliminary cost estimation" and "bridges" in the title, abstract, and keywords fields of the search engines. The search was conducted in electronic databases such as the Web of Science (WOS), Google Scholar, Scopus, and the American Society of Civil Engineers (ASCE). The focus of the search was restricted to the various databases instead of books or reports assuming that significant findings from these books or reports will be cited in the journals. Also, the journal papers were peer-reviewed.

A uniform search pattern was utilized for some databases while the string of words was moderately adapted to fit the database format to identify the majority of research studies. Each journal and/or conference paper was explored to obtain information. Articles and conference papers were identified based on the search process. Additionally, to create a complete list of articles, the abstracts of the identified articles were read. For some, the entire paper was read when the abstract was not clear-cut. Once the articles were identified, detailed content analysis was carried out to (a) profile the identified articles based on the type of journal, and year of publication; (b) explore the cost drivers (input variables) and cost estimation methods specified in the articles; and (c) systematically identify, categorize, and rank common cost drivers (input variables) of bridge projects. Content analysis is a research technique for determining major facets and valid inferences either quantitatively or quantitatively, depending on the research issue addressed (Chan et al. 2009). The Mean Absolute Percentage Error (MAPE) of the papers was compiled and subjected to meta-analysis.

2.4.3. Inclusion and exclusion criteria

The articles were filtered based on the inclusion and exclusion criteria adopted for this systematic literature review. The inclusion criteria used for selection were: (a.) the article should

be a peer-reviewed journal and conference papers published from 1990 to 2020 which are centered on bridges; (b.) the article should identify input variables for cost estimation models of bridges; and (c.) the article should identify cost estimation models or methods for bridges.

Research studies that were not centered on transportation infrastructure or were not published within the date range or were not relevant to research questions were excluded. Informal literature surveys, and the duplicates of research studies existing in different journals, were also excluded.

2.4.4. Study selection

The search across the databases using the string “conceptual cost estimation" or "preliminary cost estimation" and "bridges" in the title, abstract and keywords field of the search engines yielded 194 articles for consideration (See Figure 2). Fifty-nine (59) articles that were duplicated in other journals and conferences were excluded as well as 78 articles that were not centered on bridge projects. This phase resulted in 57 publications for further consideration. Twenty-eight (28) articles were removed as they did not discuss cost estimation methods and cost estimation cost drivers of highway bridges resulting in 29 articles.

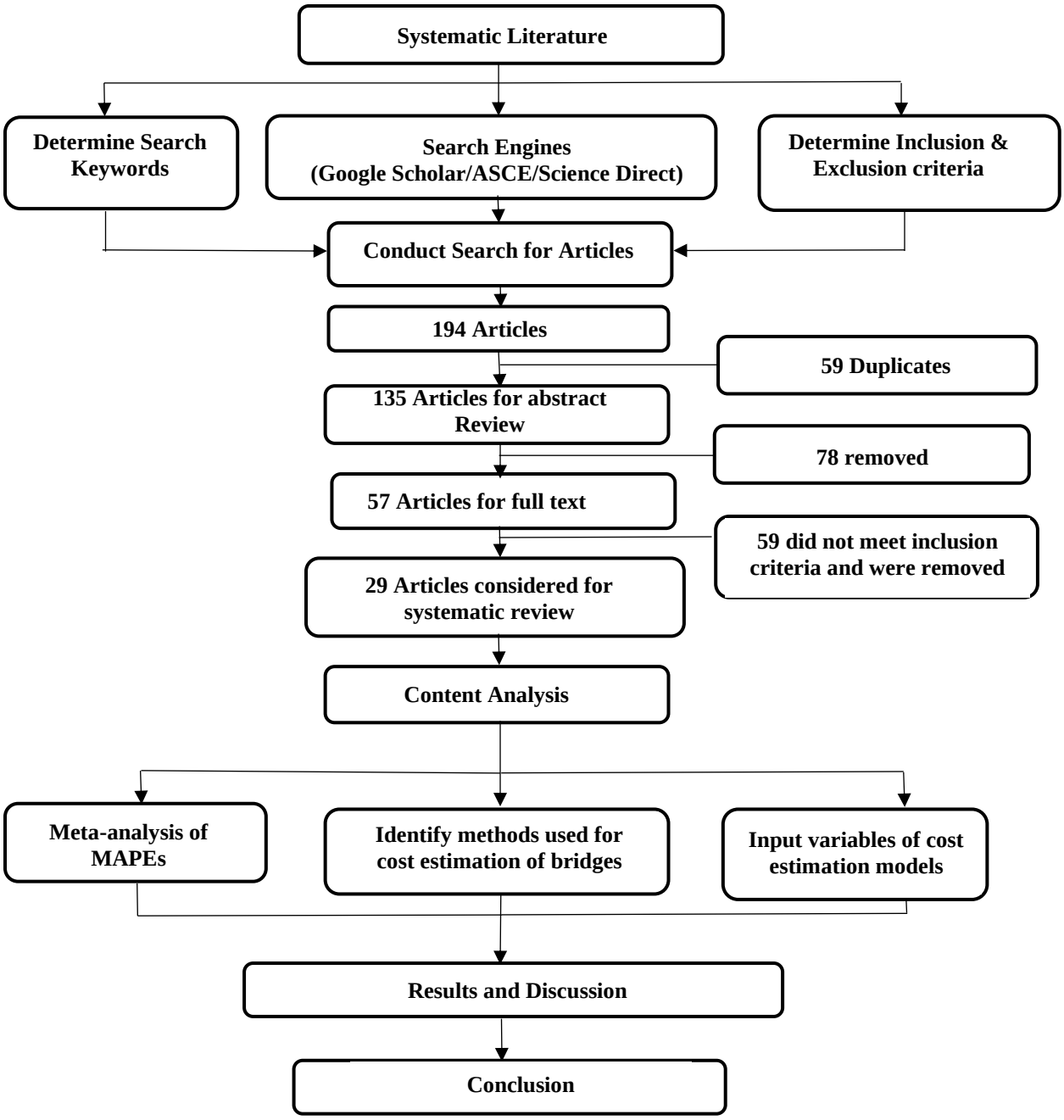


Figure 2: Flow of Methodology

2.4.5. Content analysis

Content analysis is a useful method for gathering and organizing information as well as examining trends and patterns in documents (Gupta et al. 2018). The qualitative content analysis focuses on grouping data into categories. In contrast, quantitative content analysis determines the numerical values of categorized data such as frequencies, ratings, and rankings by counting the

number of times a topic is mentioned. Both qualitative and quantitative content analyses were utilized in the research study. The cost-driving factors used for estimating the cost of bridge projects of cost estimation of bridges of the selected articles and conference papers were reviewed. The articles were then grouped under the author's name, and the year they were published.

2.4.6. Meta-analysis procedure

A meta-analysis (MA) is an empirical tool for explaining the differences in value estimates across studies. Meta-analysis is a technique that is used for analyzing the results of multiple studies statistically. It could be applied to a group of empirical studies rather than theoretical studies and produces quantitative results (Shelby and Vaske 2008). It calculates the combined effect size for a particular relationship by considering the effect size from multiple studies, representing the overall strength of that relationship for all the combined analyses. The effect size is a measurement of the strength of a particular relationship. It is commonly measured in terms of either correlation coefficient or standardized mean difference or odds-ratios from each study (Wirtz, Sparks, and Zimbres 2017). Generally, meta-analysis consists of three steps: (1) review of literature for the relevant studies; (2) coding of the studies and calculation of effect size; and (3) analysis of data (Osbaldiston and Schott 2012). Figure 3 is a summary of the meta-analysis process. This technique is widely used in medical research, social research, psychology, quality control and assurance, engineering, marketing, and public policy analyses (Shelby and Vaske 2008; Janakiraman et al. 2016; Pelaez et al. 2017).

Before conducting an MA, the studies were critically analyzed to determine if they could be included in the review, as well as determine whether or not a statistical combination of their results is feasible. Moreover, the quality of the included studies will directly affect the validity of the conclusions. Although a strict inclusion criterion could be set, studies included in an MA will

always be different from each other, defined as heterogeneity (Neyeloff et al. 2012). Heterogeneity may make the results of different studies different, not by chance, but by the evaluation of the results, or the analyses used. There are statistical ways to quantify heterogeneity (Linares-Espinos et al. 2018). For instance, If I^2 is greater than 50%, the heterogeneity among the studies is deemed high. It may affect the validity of the MA results, so other alternatives should be considered (Ahn and Kang 2018). Thus, the results of the studies should be interpreted carefully. It is worth noting that, in certain situations, MA is not always suitable to adopt for every research study when the heterogeneity between the papers is very high, or when a systematic review (SR) consists of non-randomized studies. The most suitable means to present the results is to exhibit them graphically in a forest plot without merging them statistically or tabulating the results of each research study (Linares-Espinos et al. 2018). However, for this study, results were presented graphically with a forest plot and tabulated as well to give a meaningful interpretation.

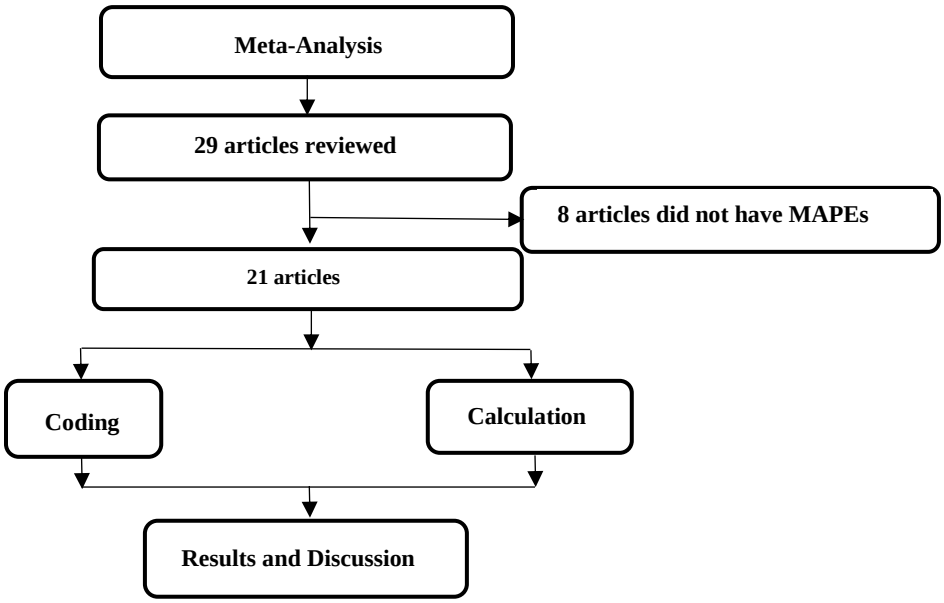


Figure 3: Meta-analysis Methodology

Out of the 29 studies on cost estimation of bridges, 21 had sufficient data to compute or estimate 21 independent effect sizes. The information required for the meta-data analysis in this

study is the Mean Absolute Percentage Error (MAPE) values which is the common method for presenting the results of delay studies. The primary outcome of a meta-analysis is the effect summary. The calculation of the effect summary depends on the model selected. A fixed or random model could be selected based on the heterogeneity of the studies (Higgins 2008). A fixed model assumes that differences in the studies are due to sampling error, while a random model considers differences in the sampling population. Many researchers use a random model (Borenstein et al. 2011). A random model requires the pooling of the sample size (n) and the effect size (es) – considered to be the Mean Absolute Percentage Error (MAPE) values from the selected articles. These two values are used to compute the effect summary according to the following equations (Neyeloff et al. 2012):

Table 1: List of Equations Utilized in the Meta-analysis

Name	Equation	Reference
Standard Error (SE)	$\frac{MAPE}{\sqrt{MAPE * n}}$	Neyeloff et al. 2012
Variance (Var)	SE^2	Neyeloff et al. 2012
Individual study weights (w)	*1/variance*	Neyeloff et al. 2012
Cochran's Q statistic (Q)	$\sum(w * MAPE^2) - \frac{[\sum(w*MAPE)]2}{\sum w}$	Neyeloff et al. 2012
Modification constant (v)	$= \frac{Q-(k-1)}{\sum w-(\frac{\sum w^2}{\sum w})}$	Neyeloff et al. 2012
Modified study weight (w_v)	$\frac{1}{(variance\ + v)}$	Neyeloff et al. 2012
Effect summary (es)	$\frac{\sum(w_v * MAPE)}{\sum w}$	Neyeloff et al. 2012
Standard Error for effect summary (SEes)	$\sqrt{\frac{1}{\sum wv}}$	Neyeloff et al. 2012

Definition:
K = number of studies.

In this study, for the meta-data analysis, a spreadsheet was prepared which incorporated the statistical formulae. It is worth mentioning that meta-analyses are not merely an aggregation of the average values of various studies. Instead, it considers the standard errors and variances to compute weights used to modify the effect size (in this case, Mean Absolute Percentage Error (MAPE values) from various studies as depicted in Table 2.

Table 2: MAPE Values Obtained from Selected Studies for Meta-data Analysis

Author(s)	MAPE (es)	Standard Error (SE):	Variance (Var):	Individual study weights (w):	w*es	$w*(es^2)$	w^2
Marinelli et al .2015	0.12	0.04251	0.001807	553.2953621	68	8.3572	306135.7577
Marinelli et al. 2015	0.17	0.04973	0.002473	404.2806183	68	11.4376	163442.8183
Marinelli et al. 2015	0.16	0.04904	0.002405	415.6479218	68	11.1248	172763.1949
Dimitriou et al. 2018	0.11	0.04108	0.001688	592.3344948	68	7.8064	350860.1537
Dimitriou et al. 2018	0.14	0.04527	0.00205	487.804878	68	9.4792	237953.599
Dimitriou et al. 2018	0.16	0.04868	0.002370	421.8362283	68	10.9616	177945.8035
Bouabaz and Hamami 2008	0.00	0.00883	0.00007	12800	32	0.08	163840000
Kim 2011	0.12	0.03110	0.00096748	1033.613445	123	14.637	1068356.754
Kim and Hong 2012	0.01	0.00738	5.44776E-05	18356.16438	134	0.9782	336948770.9
Fragkakis et al. 2010	0.11	0.03745	0.001402564	712.9798903	78	8.5332	508340.324
Hollar et al. 2013	0.43	0.03304	0.001092	915.6908665	391	166.957	838489.763
Winalytra et al.2018	0.07	0.07311	0.005346	187.0503597	13	0.9035	34987.83707
Winalytra et al.2018	0.07	0.07109	0.005053	197.869102	13	0.8541	39152.18152
Antoniou et al. 2018	0.01	0.01706	0.000291	3434.343434	34	0.3366	11794714.83
Antoniou et al. 2018	0.01	0.01740	0.000302	3300.970874	34	0.3502	10896408.71
Morcous et al. 2001	0.08	0.05838	0.003409	293.3333333	22	1.65	86044.44444
Winalytra et al.2018	0.06	0.06615	0.004376	228.4710018	13	0.7397	52198.99864
Winalytra et al.2018	0.03	0.04859	0.002361	423.4527687	13	0.3991	179312.2473
Yu 2006	0.03	0.01279	0.00016	6106.870229	160	4.192	37293863.99
Kim et al. 2009	0.01	0.01790	0.000320	3120	39	0.4875	9734400

For the total set of 21 studies being investigated, a weighted mean effect size was computed by weighting the effect size of each study by the inverse of its variance. The precision of each mean effect estimate was determined using the estimated standard error of the mean to calculate the 95 percent confidence interval. Using a random-effects model, the heterogeneity of the effect size distribution (the *Q*-statistic) was computed to indicate the extent to which variation in effect sizes was not explained by sampling error alone. Furthermore, the I^2 was calculated to quantify the heterogeneity. The I^2 is expressed in percentage of the total variability in a set of effect sizes

due to true heterogeneity, that is, to between-studies variability. Next, a series of post-hoc subgroup and moderator variable analyses were conducted using Excel (See Table 3).

Table 3: Results from the Meta-analysis

Author(s)	Year	MAPE	Standard Error (SE):	Variance (Var):	Individual study weights (w):	CI Lower	CI Upper
Marinelli et al.	2015	12.29%	0.042513	0.001807	553.295362	8.332543	28.955086
Marinelli et al.	2015	16.82%	0.0497346	0.002474	404.280618	9.747979	36.315959
Marinelli et al.	2015	16.36%	0.0490498	0.002406	415.647922	9.61376	35.58752
Dimitriou et al.	2018	11.48%	0.0410881	0.001688	592.334495	8.053276	27.586511
Dimitriou et al.	2018	13.94%	0.0452769	0.00205	487.804878	8.874277	31.688555
Dimitriou et al.	2018	16.12%	0.0486887	0.002371	421.836228	9.542983	35.205965
Bouabaz and Hamami	2008	0.25%	0.0088388	7.81E-05	12800	1.732412	3.7148232
Kim	2011	11.90%	0.0311043	0.000967	1033.61345	6.09645	24.092899
Kim and Hong	2012	0.73%	0.0073809	5.45E-05	18356.1644	1.446655	3.6233109
Fragkakis et al.	2010	10.94%	0.0374508	0.001403	712.97989	7.340361	25.620722
Hollar et al.	2013	42.70%	0.0330465	1.09E-03	915.690867	6.477115	55.654231
Winalytra et al.	2018	6.95%	0.0731174	5.35E-03	187.05036	14.33101	35.61202
Winalytra et al.	2018	6.57%	0.0710904	0.005054	197.86102	13.93372	34.43744
Antoniou et al.	2018	0.99%	0.0170639	2.91E-04	3434.34343	3.344523	7.6790464

2.5. Interpretation of Results

In this section, the summary of the results of the analysis of literature is presented. Trends in cost estimation methods are provided as well as the cost drivers that are considered in the estimation of bridge projects. This is followed by a discussion highlighting some key findings.

2.5.1. RQ1: Cost estimation methods for bridge projects

By reviewing the methods used by the various authors, it is obvious some methods have been overlooked throughout literature and others have been used more frequently by other authors over the years. As seen in Figure 4, it is evident that unit cost is used often as it has the highest frequency, followed by regression analysis and artificial neural networks.

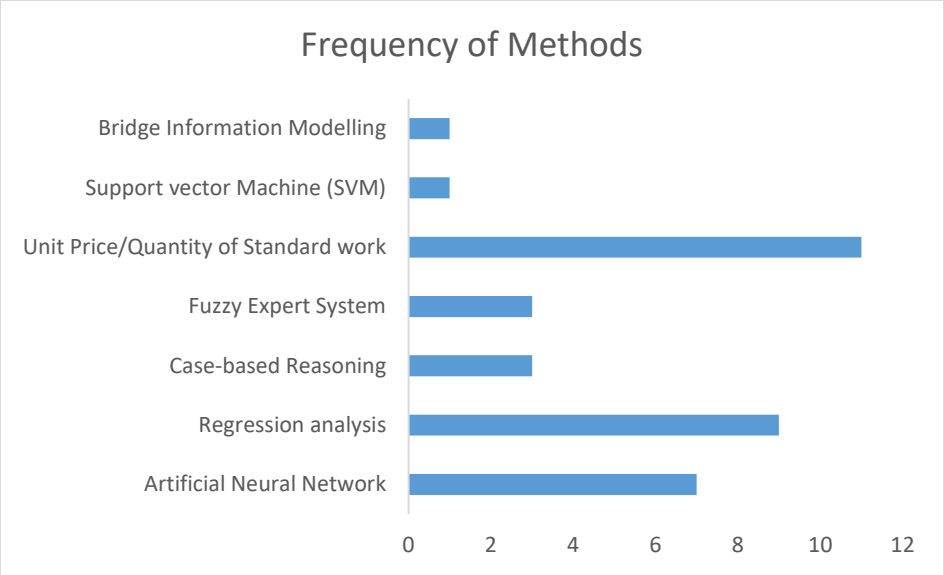


Figure 4: Frequency of Cost Estimation Methods

Unit price estimates are obtained by establishing the number of works and their unit cost for each work section of the bridge project. With less detailed design, specification, and other relevant information, inaccurate estimates could be obtained while using the unit price estimate approach (Kim and Cho 2013). Regression analysis is often used for parametric cost estimation during the conceptual phase of bridge projects. It is a technique that utilizes the mathematical relationship of independent variables which are the estimating cost factors and the dependent variable, that is the project cost to be estimated (Antoniou 2016; Dimitriou 2018). The cost estimating models are developed by applying regression analysis to historical data. The advantages include the level of accuracy provided by the regression analysis method and the simplicity in its usage (Kim et al. 2010).

However, the regression method is limited by the development of an appropriate mathematical model that best fits the historical project cost data (Jai et al. 2016). Artificial Neural Network (ANN) involves three layers which include input, hidden, and output. The neural network model operates like the human brain by learning from past cases to predict the outputs. By creating

layers of arbitrary data, the input variables are transformed into output variables. The data from past projects is used to train the model, which develops relationships within the database to predict the output variables. Finally, the trained model is used to predict the output variable by recognizing patterns in the trained data. Even though the number of inputs and outputs is not limited, which is an advantage of ANNs, the number of hidden layers and neurons is defined. On the other hand, one of the drawbacks is establishing the number of neurons which is time-consuming. Case-based reasoning (CBR) is an alternative to an expert system, which is based on rule-based reasoning. Reasoning in CBR is based on experience or memory. A case-based reasoner solves new problems by adopting solutions that were used to solve old problems. The other methods identified include support vector machines (SVM), Bridge Information Modeling (BrIM), and Fuzzy Expert Expert System (FES).

2.5.2. RQ2: Publications trends

Identifying the trend of the cost estimation methods enables researchers to know the cost estimation methods that have been utilized in the past and forecast the future direction of cost estimation research of bridge projects. Figure 5 shows the distribution of the methods for the various years identified in the literature. It is an overview of the trend which shows the publication rate increased after 2009.

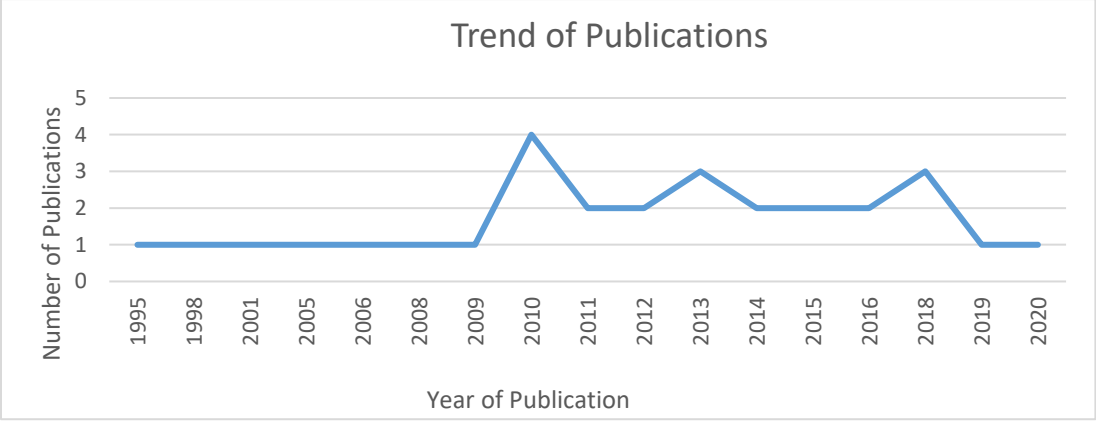


Figure 5: Trend of Bridge Cost Estimation from 1990 to 2020

It was observed that single methods were adopted in the cost estimation of bridge projects from 1995 to 2009. Several studies were conducted to improve cost estimation in bridge projects between 1990 to 2020, with the highest number of publications (4) recorded in 2010. Despite the steep increase in publication over the years, the number of studies recorded from 2011 through 2020 fluctuated from the previous years. However, as shown in Figure 5, hybrid methods have been adopted with the emerging development of modern artificial intelligence (AI) techniques from 2010 to 2020, which is in accordance with the assertions of Winalytra et al. (2018). An active area of research on analog-based cost estimation is applying AI (artificial intelligence) techniques to capture the domain experts' estimation knowledge and mimic the decision processes of estimators. There have been promising results in the application of various AI techniques for construction cost estimation, including applications of expert systems (ESs), case-based reasoning (CBR), artificial neural networks (ANNs), neuro-fuzzy systems (NFS), and statistics regression approaches to cost estimation.

2.5.3. RQ3: Input variables adopted for cost estimating models/equation of bridge projects

The input variables were identified and defined through an intensive literature review for the required numerical analysis. The study of various research publications relevant to the cost estimation of bridge construction projects explored the most influential input variables (Adel et al. 2016). Meseret et al (2019) stated that it is not necessarily true that increasing the number of input variables in an early estimate may seem to improve the accuracy of the estimate. However, selection of the appropriate input variables can improve the accuracy of the cost estimate, particularly at the early stages of the project development. In this study, 32 input variables or attributes were compiled from the literature and ranked.

In order to select the input variables as shown in Table 4, the frequency of the variables in the literature, and a variable whose information is available at an earlier stage were considered as the criteria. The common input variables included length of span, number of spans, type of foundation, the weight of steel, and concrete volume. These input variables are preferred because of their high correlation with the construction cost of bridges. The validity in preliminary or conceptual cost-estimation practice is that "project size" is the most significant and important input variable in highway construction projects. Regarding project size or length, it can be generalized that, the larger the project, the more expensive it will be. This finding was supported by other researchers (Mahmid 2011; Gradner et al. 2016). Elfaki et al. (2014) also proved that the project size and labor hours strongly correlate. The number of bridges in the project scope greatly affects the cost of construction. It has a direct relationship with the construction cost as it enlarges the overall scope of the project. The results are in line with Morcous et al. (2001), which indicates that the cost of superstructure concrete and prestressing steel represents a significant percentage of the total cost of structural bridge construction. Thus, materials such as concrete and steel are usually not overlooked when developing equations or models for cost estimation of bridges.

Similarly, since a conceptual design is available at the early phase of bridge projects, variables such as length of span, number of spans, type of foundation are commonly used. Morcous et al. (2001) opined that one of the essential attributes in the superstructure design includes the span of the length of bridges. Morcous et al. (2001) conducted further studies to establish a correlation between concrete volume and span length and the correlation between the weight of steel and the span length. The correlation analysis resulted in a coefficient of determination of 67% and 51%, respectively. The superstructure has a considerable influence on the building costs of concrete bridges **(**Fragkakis et al 2010).

According to Konstantinidis and Maravas (2003), depending on the construction method adopted, the construction cost varies from 35% to 53% of the overall bridge construction cost. Similarly, for the type of foundation, Fragkakis et al. (2011) mentioned that foundations have a substantial effect on the construction cost of modern concrete bridges; their cost, depending on the construction process used and the bridge design scheme, varies from 19 to 27% of the overall bridge construction cost. According to Antoniou et al (2018), 84% of the cost is consumed by reinforced concrete construction (reinforcement and concrete), while 12% corresponds to the cost of earthworks and only 4% to waterproofing, joint construction, and drainage. Overall, the common inputs comprise of basic material and pertinent design parameters which are commonly known at the early phase of the bridge construction projects.

Table 4: Input Variables of the Cost Estimation of Bridge Projects

Authors	Creese and Li 1995	Bakhoum et al. 1998	Morcous et al. 2001	Chou et al. 2005	Yu 2006	Chou and Connor 2007	Bouabaz and Hamami 2008	Kim et al. 2009	Kim and Kim 2010	Van de lit and Stone 2010	Hollar et al. 2010	Fragkakis et al. 2010	Fragkakis et al. 2011	Kim 2011	Kim and Hong 2011	Marzouk and Hisham 2012	Hollar et al. 2013	Oh et al. 2013	Kim and Cho	Markiz and Jrade 2014	Fragkakis et al. 2014	Jai et al. 2016	Behmardi et al. 2015	Marinelli et al. 2015	Antoniou et al. 2016	Antoniou et al. 2018	Dimitriou et al. 2018	Winalytra et al. 2018	Markiz and Jrade 2019	Juszczyk 2020	Total	Rank
Project Type						x																								x	2	**16**
Project Duration						x					x																				2	**16**
Location						x			x						x							x								4	**11**
Bridge type			x						x						x			x											x	6	**9**
Construction method		x	x																						x			x		4	**11**
Contract type		x	x																											2	**16**
Number of bridges						x																								1	**21**
Bridge Material						x												x											x	3	**14**
Design for AADT																					x									1	**21**
Bridge deck area	x				X							x	x							x			x	x		x				8	**6**
Length of bridge									x					x	x							x								4	**11**
Number of lanes									x																					1	**21**
Correction factor			x																											1	**21**
Concrete volume			x	x			x					x	x					x		x			x		x	x				10	**2**

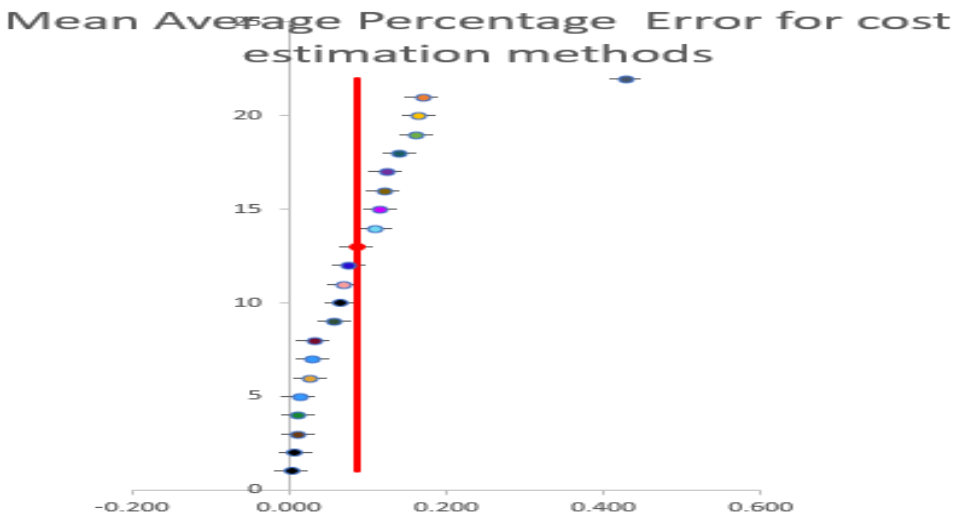
Table 4: Input Variables of the Cost Estimation of Bridge Projects (Continued)

Author	Creese and Li 1995	Bakhoum et al. 1998	Morcous et al. 2001	Chou et al. 2005	Yu 2006	Chou and Connor 2007	Bouabaz and	Kim et al. 2009	Kim and Kim 2010	Van de lit and	Hollar et al. 2010	Fragkakis et al. 2010	Fragkakis et al. 2011	Kim 2011	Kim and Hong 2011	Marzouk and Hisham	Hollar et al. 2013	Oh et al. 2013	Kim and Cho	Markiz and Jrade 2014	Fragkakis et al. 2014	Jai et al. 2016	Behmardi et al. 2015	Marinelli et al. 2015	Antoniou et al. 2016	Antoniou et al. 2018	Dimitriou et al. 2018	Winalytra et al. 2018	Markiz and Jrade 2019	Juszczyk 2020	Total	Rank
Pile length					x																										1	21
Services and ancillaries								x																							1	21
Area of Slab								x	x																						2	19
Number of spans									x		x			x	x							x	x		x	x		x			9	3
Length of pier								x	x				x					x							x	x	x				7	7
Type of foundation				x	x		x	x	x										x	x									x	x	9	3
Volume of webs	x																														1	21
Volume of bridge decks	x																														1	21
Weights of steel	x		x									x	x					x	x	x	x			x		x	x				11	1
Length of span		x	x		x					x				x	x									x		x	x				9	3
Number of girders										x										x											2	16
Height of bridge														x	x			x													3	14
Types and number of bridge shoe																		x													1	21
Size of sidewalk																												x			1	21
Railing Type																												x			1	21
Width of bridge									x	x													x			x		x			5	10
Type of superstructure		x	x		x			x							x														x	x	7	7

2.5.4. RQ4: The impact of cost estimation method used on the accuracy of cost estimates

The analysis is based on twenty-one studies that evaluated the cost estimation methods or models of bridge projects. The MAPE of the cost estimation of each bridge project was identified as the effect size(es). The studies in this analysis were sampled from a universe of possible studies defined by specific inclusion/exclusion rules as outlined in the previous section. For this reason, the random-effects model was employed for the analysis. From Figure 6, it was observed that a study had a high MAPE of 40%. This outlier could have resulted from errors in the data used for the model. The confidence interval for the cost estimation/MAPE is from 0.053 to 0.119, which indicates that the MAPE (cost estimation) in the defined universe could fall anywhere in this range. This range does not include a MAPE of 1.0. Thus, the MAPE ratio is probably not 1.0. Similarly, the Z-value for testing the null hypothesis (that the MAPE is 1.0) is 5.139, with a corresponding p-value is < 0.000. The null hypothesis which is, the cost estimation methods adopted do not affect the accuracy of cost estimation of bridge projects was rejected. The conclusion is that the cost estimation method adopted affects the MAPE**.**

The Q-statistic provides a test of the null hypothesis that all studies in the analysis share a common effect size identified as the studies MAPEs. If all studies shared the same effect size, the expected value of Q would be equal to the degrees of freedom (the number of studies minus 1). The Q-value is 33.533 with 20 degrees of freedom. Thus, the observed dispersion is less than expected by chance with an I^2 equal to 40.36%. It follows that the true effect size varies moderately from study to study. The I^2 statistic indicates what proportion of the observed variance reflects differences in true effect sizes rather than sampling error. Here, I^2 is 40.36%. T^2 is the variance of true effect sizes is 0.004005. T is the standard deviation of true effects is equal to 0.063285.



LEGEND

	MAPE	Cost Estimation Method
	0.427	Multiple Linear Rgression/ Hollar et al. (2013)
	0.168	FFANN / Marinelli et al. (2015)
	0.164	FFANN / Marinelli et al. (2015)
	0.161	FFANN / Dimitriou et al. (2018)
	0.139	FFANN / Dimitriou et al. (2018)
	0.123	FFANN/ Marinelli et al. (2015)
	0.119	CBR/ Kim (2011)
	0.115	FFANN / Marinelli et al. (2015)
	0.109	MLR/ Fragkakis et al. (2010)
	0.086	Summary Effect
	0.075	ANN-Back propagation/ Morcous et al. (2001)
	0.070	MLR/ Winalytra et al. (2018)
	0.066	MLR/ Winalytra et al. (2018)
	0.057	ANN-Back propagation/ Winalytra et al. (2018)
	0.031	ANN-Back propagation/ Winalytra et al. (2018)
	0.030	Standard quantities/ Oh et al. (2013)
	0.026	ANFIS/ Yu (2006)
	0.013	Quantity of standard work/ Kim et al. (2009)
	0.010	MLR/ Antoniou et al. (2018)
	0.010	MLR/ Antoniou et al. (2018)
	0.007	CBR/ Kim and Hong (2012)
	0.003	FFANN/ Bouabaz and Hamami (2008)

Figure 6: Forest Plot of Mean Average Percentage Error for Cost Estimation Methods

2.6. Conclusion

The systematic literature search found most of the search results by creating search clusters, and step-by-step filtering. This research study explored a systematic review of previous articles to determine research trends, identify the cost variables that are adopted in cost estimation models

and influence the construction cost of bridges, and ascertain the impact of cost estimation methods on the accuracy of highway bridge costs at the conceptual stages of projects. Selected journals from 1990 to 2019 were identified, giving rise to 29 articles on bridge cost estimation at the initial stages of a project. The trend in conceptual cost estimation research with regard to the distribution of publications per year was analyzed. Furthermore, the methods used in estimating construction costs at the conceptual stages of bridge construction cost were categorized. The cost drivers used in bridge cost estimation models were identified and ranked. The results indicate that despite the steep increase in the number of publications between 2009 and 2010, the number of studies published from 2010 to 2019 fluctuated. This shows a need to increase research efforts to enhance conceptual cost estimation methods because of the rapid advancement of transportation infrastructure globally.

The findings from the study showed that the top three cost estimation methods adopted for bridge construction cost were unit price/quantity of standard work, regression analysis, and artificial neural network. Some research studies combined case-based reasoning and genetic algorithms; multiple regression analysis and artificial neural networks; artificial neural network and genetic algorithms; artificial neural networks and support vector machines; factor analysis and multivariate regression; fuzzy expert system; and bridge information management system to augment the proficiencies of single methods. In relation to the cost drivers, the most used variables included the weight of steel, the volume of concrete, number of spans, type of foundation, which are materials and characteristics of the bridge. These influence the models that were used for cost prediction. The meta-analysis results indicated that the type of cost estimation method adopted for predicting the estimates of bridge projects impacts the accuracy of the prediction. The hypothesis

testing result (p-value<0.05) indicated that the cost estimation method adopted affects the MAPEs. The I^2 statistic of 40.36% showed that the true effect size varies moderately from study to study.

This research paper's distinctive contribution to the body of knowledge is its comprehensive statistical analysis of the data to evaluate and present primary insight into the accuracy of the cost estimation models from the selected literature. Future research could focus on the geographic differences between cost estimation practices, the main elements affecting the quality of cost estimation and cost management practices, and its relationship to cost estimation in bridge projects.

CHAPTER 3. COST ESTIMATION MODEL FOR WISCONSIN BRIDGE PROJECTS USING REGRESSION ANALYSIS

3.1. Abstract

In construction field, an adherence to budget is one of the principal contributors to the success of projects. The adherence to the budget can be achieved when the estimated budget is closed to the actual cost. For the owners, cost estimation is necessary as a guidance in determining the amount of investment. Therefore, it is very important to know the estimation of the project cost by using the limited data before the detailed plans and specifications of the project are identified. Hence, the primary objective of this study was to develop linear regression models to predict the construction cost of bridges, based on 200 sets of data collected within Wisconsin State. Nine potential independent variables were identified which include deck area, road width, number of spans, span length, piling size, bridge length and deck width. Among the variables, location, deck width and deck area variables were significant in the model, suggesting that they are the key linear cost drivers in the data. The regression models for forward and backward regression gives an R^2 42.31% and 41.06% respectively.

3.2. Introduction

Estimating construction cost is an essential part of planning and preparation for construction projects. The estimated construction cost is used for verifying the feasibility study on the facilities or in evaluating design alternatives. Thus, it becomes a key element in the decision-making process involved in the project. In particular, when the projects are bigger and many projects are being undertaken at the same time, accuracy of the estimated construction cost in the planning phase is essential in the efficient use of the budget. The lack of information required for estimating construction cost in the planning phase, however, results in inaccuracy. Thus, it is

necessary to devise a method that would improve the accuracy of construction cost estimation during the planning phase. Many researchers have studied and developed models that can be used to improve the estimation process, based on various methods.

It is worth noting that early estimates are critical to the initial decision-making process for the construction of capital projects. As such, the importance of early estimates to owners and their project teams cannot be overemphasized. Early estimates are typically plagued by limited scope definition and thus high potential for scope change which are often prepared under time constraints. Furthermore, reliable cost data are often difficult to obtain during the conceptual stages of a project, particularly if basic design and geographic issues remain unresolved. Early estimates, even when grossly inaccurate, often become the basis upon which all future estimates are judged with future estimates sometimes being adjusted to be consistent with early estimates.

Several studies have found that clients are generally dissatisfied with the initial cost advice provided by their construction professionals (Bouras 2018). Bouras (2018) concluded that there is a need to provide more accurate and robust forecasts of construction costs. While it is widely held that a perfect estimate is not possible and even the best possible estimate will always contain a number of key risks, the goal of the forecaster is a practicable level of accuracy Smith 1995. Hollar et al (2013) that a vital consideration with any method of estimating is the accuracy by which anticipated costs can be predicted. Regression analysis is one of the modeling techniques adopted, which have been used to develop models to estimate the cost of construction. However, predominantly, the model relies on the use of historic but recent cost data.

Regression analysis (RA) represents one of the most widely used methods for conceptual cost estimation during the early project stages. Despite its various drawbacks, in particular the requirement of a defined mathematical form that best fits the available historical data, the difficulty

in accounting for the large number of variables present in a construction project, as well as the numerous interactions among them (Woldesenbet and Jeong, 2012). RA provides adequate accuracy, benefits from the parsimonious use of the parameters and presents simplicity in its use.

Aiming to minimize the prediction error in conceptual estimates, Fragkakis and Lambropoulos (2004) collected actual cost information from a large sample of 119 concrete bridges and overpasses constructed between 1999 and 2003 as part of the Egnatia Motorway in Greece. They divided the actual bridge construction cost into earthworks, foundation, substructure, superstructure, and accessories and proposed the average cost values for four major deck construction methods. For example, concerning bridges consisting of precast prestressed beams and reinforced concrete slabs, the average cost percentages for the five elements above are 4.70%, 26.70%, 15.90%, 34.40%, and 18.30% respectively.

Likewise, Fragkakis et al. (2010) applied regression analysis on a database consisting of 68 structures and derived material estimating models for three widely used deck construction methods (cantilever construction, precast prestressed beams, cast-in-situ box girders). Bridge cost estimation guidelines, mostly based on historical bid data, have also been developed by Departments of Transportation (DoT) in United States.

Similarly, Fragkakis et al. (2011) employed backward stepwise regression to develop models for predicting material quantities for different types of foundations and estimated the total foundation costs. In the study, it was noted that the results of the models reconciled with the opinions of the expert when they were interviewed and also, the goodness of fit of the final model was satisfactory irrespective of the factors that impact foundation design.

Also, Hollar et al. (2013) assessed the costs of preliminary engineering of bridges by analyzing 461 bridge projects in North Carolina, USA, between 2001 and 2009. Authors further

developed a model using multiple linear regression. he MLR model for PE cost ratio prediction includes four numerical and four categorical variables, representing project specific parameters and interactions. Although the prediction error percentage was greater that desired (42.7%), the regression analyses did yield useful insights into PE cost-estimating.

However, Winaltra et al. (2018) adopted both multiple linear regression and artificial neural network in order to attain the suitable estimation model. The results of the analysis showed that bridge span and width were the significant factors influencing cost. The correlation value of bridge span is 89.0%, bridge width is 74.2%, the size of the sidewalk is 66.1%, and railing's type is 46.1% as identified factors that affect the cost of the superstructure. The aim of the paper is to explore cost estimation models by utilizing historic data to predict construction cost at the conceptual stage and compare the predictions of multiple regression.

3.3. Methodology

3.3.1. Data acquisition

The data used in this study were obtained from Wisconsin Department of Transportation Data website, Highway Structures Information System (HSIS). Highway Structures Information System (HSIS) is a repository of data consisting of construction cost of bridges, the design parameters of the bridges, locations, and the conditions of bridge projects within the Wisconsin Department of Transportation (DOT). For this study, a sub database was established to cover bridge projects with no missing data only from the original dataset. The resulting dataset consisted of year of bridge construction, materials used, configuration of the bridge, design parameters, and project cost. Bridge project costs were based on year of construction which differs for the different projects.

3.3.2. Calculation of the present cost of construction costs

The bridge projects selected for analysis are spread from 2013 to 2019. Thus, construction cost of bridges needed to be adjusted for inflation. Furthermore, in order to compare costs consistently between systems in different years for the analysis and account for the time value of money, costs data were normalized and converted to 2020 dollars using the Chained Fisher Construction Cost Index is used. Time adjustments using the historical cost indexes was done using the following formula:

$$Index\ for\ Year\ 2020\ Index\ for\ \text{the}\ Year\ that\ the\ data\ is\ based\ on \times Cost\ for\ Year\ that\ the\ data\ is\ based\ on = Cost\ for\ Year\ 2020$$

3.3.3. Input and output variables

The data collected were grouped into independent input variables and dependent output variables. The regression model detailed in this research study is based on construction cost only. An extensive literature review discovered various predictor input variables. These were finally reduced to 9 variables, which it was believed would be known at the early estimating stage the stage at which the models are intended to be used. The influence of time was accommodated through the use of the Wisconsin Department of Transportation Chained Fisher Construction Cost Index 2010-Q1 by adjusting all the data sets projects to a common base date.

3.3.4. Potential model

Linear regression analysis has, in the past, been performed by using raw cost as the dependent variable. However, there are a number of assumptions implicit in this choice of variable which must be addressed.

- The standard deviation in the error associated with the dependent variable cost remains constant throughout the domain,

- The error should be normally distributed, and
- The effect of any variable is always expressed in terms of a fixed cost increase or decrease, irrespective of project size or type.

With respect to the standard deviation of error remaining constant, the cost of a small project can vary by the same monetary amount as a large project. This is highly unlikely to be the case and is much more likely to be proportional to the size or cost of the infrastructure. Since regression modeling reduces the squares of the errors, it will be fundamentally biased towards minimizing the errors for very large projects, where the errors are highest. Consequently, it is not likely to be suitable predictor of the cost of relatively smaller projects. Since the costs of projects of the data for this research study vary between $132,834 and $18 M, the impacts of errors in the cost of the largest projects are a number of orders of magnitude more than those of the smallest projects. Thus, the effect will be evident. These criticisms raise questions as to the meaningfulness of models produced by using raw cost as the predictor for a linear regression model.

The estimation of the cost functions is based on multiple linear regression models, whose general form is

$$Y = b_0 + b_1X_1 + b_2X_2$$

where Y = dependent variable (i.e., estimated cost).

Xi = independent variables; and

bi = estimated coefficients.

The simple linear and polynomial regression can be considered special cases of the multiple linear regression. Linear regressions can also be used with relationships that are not inherently linear (e.g., power function), but can be linearized after a mathematical transformation.

3.3.5. Correlation

A correlation number gives the degree of association between two variables (Shinde 2018). It is important to explore possible correlations between the dependent and the independent variables in modelling to better understand the data set. The correlation test was performed to determine the strength level of the relationship between each influence factor (independent variable) to the cost change. Linear regression models are sensitive to outliers, non-linearity, and collinearity (Bouras 2018); hence there is a need to identify likely problems. Table 5 depicts the correlation between every two variables and each variable with project cost (dependent variable). In this figure, there are no highly inter correlated variables. Hence, we keep all of these variables when selecting and preparing the features to use in the modelling. On the other hand, roadway width and deck area variables have a slightly higher correlation with the project cost when compared to other variables displayed in Table 5.

Table 5: Correlation of Input Variables with Project Cost

	Deck Area	Roadway Width	Bridge Length	Number of spans	Span Length	Deck width	Piling Size
Deck Area	1.000	-0.092	-0.095	-0.036	-0.13	0.015	-0.11
Roadway Width	-0.092	1.000	0.18	-0.11	-0.043	0.42	0.282
Bridge Length	-0.095	0.18	1.000	0.26	-0.096	-0.459	0.202
Number of spans	-0.036	-0.11	0.26	1.000	-0.089	0.15	0.035
Span Length	-0.13	-0.043	-0.096	-0.089	1.000	0.29	-0.29
Deck width	0.015	0.42	-0.049	0.15	0.29	1.000	0.099
Piling Size	-0.11	0.2	0.25	0.36	-0.29	0.099	1.000
Cost	0.64	0.70	0.18	-0.339	0.37	0.42	-0.03

3.3.6. Predictive models

Two methods were attempted to create a predictive regression model. One of the methods was a forward stepwise regression modeling technique. One of the issues with this technique is that a variable that correlates well with a number of significant variables could be added ahead of other variables, because it appears to encapsulate those variables. If this encapsulating variable has

a higher significance than the individual variables themselves then the variable will be included first. When other variables are considered for addition to the model, some of the information contained in them will already be present in the model, which will make them appear less significant than they really are. One possible way of avoiding this problem is to perform a backward modeling technique. Thus, both forward and backward modeling were performed.

Model testing and validation are vital processes to give the reliability of the models before they can be adopted. Whenever the regression models are developed, it is imperative to verify the goodness of fit of the model and the statistical significance of the model and of the estimated parameters (Oredein et al. 2011). The assumptions need to be tested for validating the goodness of fit, namely:

(1) linearity and additivity of the relationship between dependent and independent variables.

(2) statistical independence of the errors.

(3) homoscedasticity of the errors; and

(4) normality of the error distribution.

The residuals analysis was carried out. For this purpose, the normal QQ (quantile-quantile) plot of residuals, for normality of the error distribution and the residuals against fitted value, for variance homoscedasticity plots were used. Furthermore, in order to verify the goodness of fit of the model, the p-value as well as the coefficient of determination, R^2, were calculated. The p-value of the model indicates its statistical significance of the results. The p-value of calculated for the parameters indicates the parameters' significance. The model or parameter is significant if the p-value <0.05). The coefficient of determination relates to the statistical measure of how well the regression equation estimates the actual data points.

3.4. Summary of Models Found

A total of two regression models were developed. The number of variables in the various models slightly changed. The smallest number of variables employed was four, in the backward stepwise model. The largest was seven variables in the forward models. The results of the model as seen in table 6 and table 7 indicates that the forward regression model performed better as the R^2 was of the forward and backward regression were 38.74% and 38.33%.

Table 6: Results of Forward Regression Model

Residuals:

Min	1Q	Median	3Q	Max
-9.266e-04	-2.583e-04	-1.109e-05	2.495e-04	1.630e-03

Residual standard error	0.000369
Multiple R-squared	0.4231
Adjusted R-squared	0.3891
F-statistic	9.690
p-value	5.47e-16

Coefficients:

	Estimate Std.	Error	t value	Pr(>|t|)
Intercept	3.375e-03	5.470e-04	6.170	3.97e-09 ***
Location/Urban	-1.632e-04	8.454e-05	-1.930	0.055049
Log (Roadway width)	1.283e-05	1.894e-04	0.068	0.946061
Log (Deck area)	-2.644e-04	7.283e-05	-3.630	0.000364***
Log (Number of span)	9.363e-05	8.694e-05	1.077	0.282842
Log (Span length)	3.600e-05	7.346e-05	0.490	0.624655
Log (Deck width)	-1.992e-04	1.167e-04	-1.708	0.089335
Log (Piling size)	3.109e-04	2.088e-04	1.489	0.138119

Table 7: Results of Backward Regression Model

Residuals:

Min	1Q	Median	3Q	Max
-9.246e-04	-2.641e-04	-6.310e-06	2.425e-04	1.630e-03

Residual standard error	0.000367
Multiple R-squared	0.4106
Adjusted R-squared	0.379
F-statistic	19.11
p-value	2.2e-16

Coefficients:

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	3.266e-03	5.154e-04	6.337	1.59e-09 ***
Location/Urban	-1.701e-04	8.228e-05	-2.068	0.0400 *
Log (Deck area)	-2.049e-04	3.359e-05	-6.100	5.60e-09 ***
Log (Deck width)	-2.341e-04	9.743e-05	-2.402	0.0172 *
Log (Piling size)	3.036e-04	2.040e-04	1.488	0.1384

3.5. Significance of Variables

Overall, the backward selection process generated models with more variables than the forward techniques. This seems to imply that through the adoption of backward selection it is likely to obtain more significant variables than utilizing forward selection. A probable reason for this is that there are a number of variables that, while not necessarily exerting a significant influence on cost in themselves, do correlate well with a number of cost significant variables that do. Thus, if this variable is included, it is possible that the influence of a number of other cost significant variables is also implicitly taken into account by the inclusion of this variable.

3.6. Model Testing and Validation

The normal quantile-quantile plot of residuals (QQ plot) in Figure. 7 and Figure 8 indicates that the regression residuals are almost normally distributed (most of the points are very close to the line), however, there are a few outliers. Furthermore, the residuals versus fitted value plot Figure 9 and 10 validates that the assumption of homoscedasticity. The errors appear to have constant variance, with the residuals scattered randomly around zero.

In addition, to verify the goodness of fit of the models developed, the statistical significance of the results that is, the (p-value of the models) and the coefficient of determination, R^2, are shown in Table 6 and Table 7.

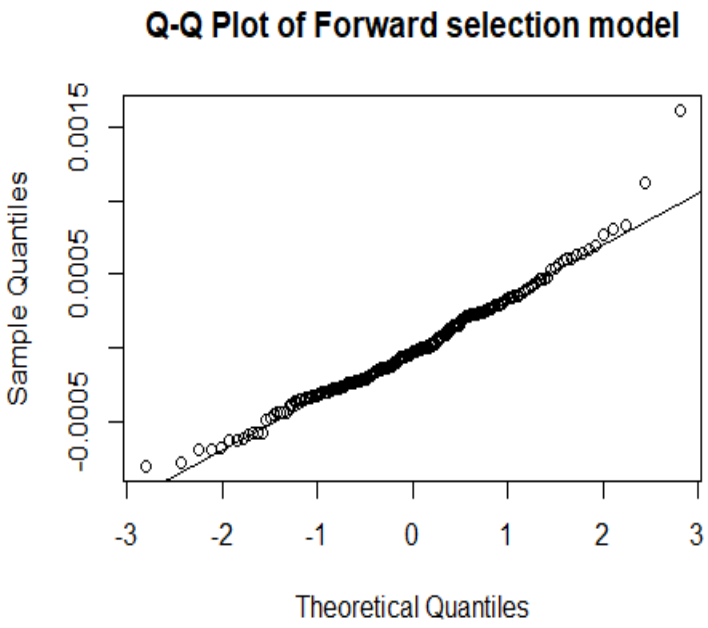


Figure 7: Q-Q plot of Forward Selection Model

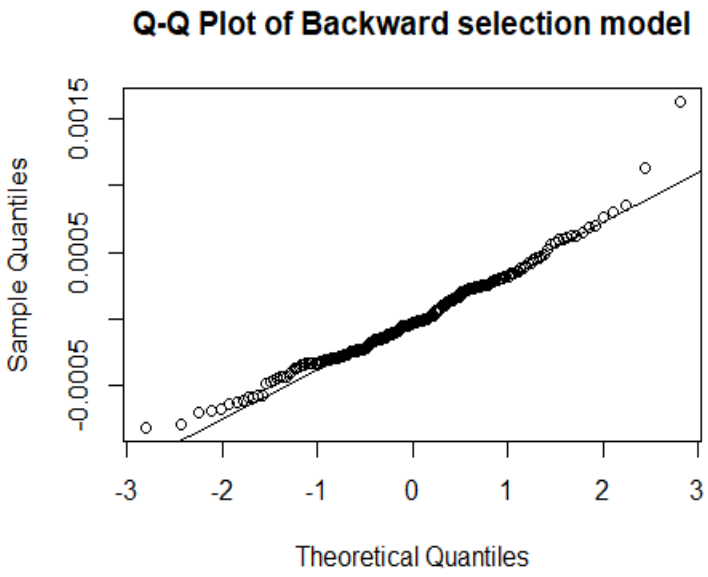


Figure 8: Q-Q plot of Backward Selection Model

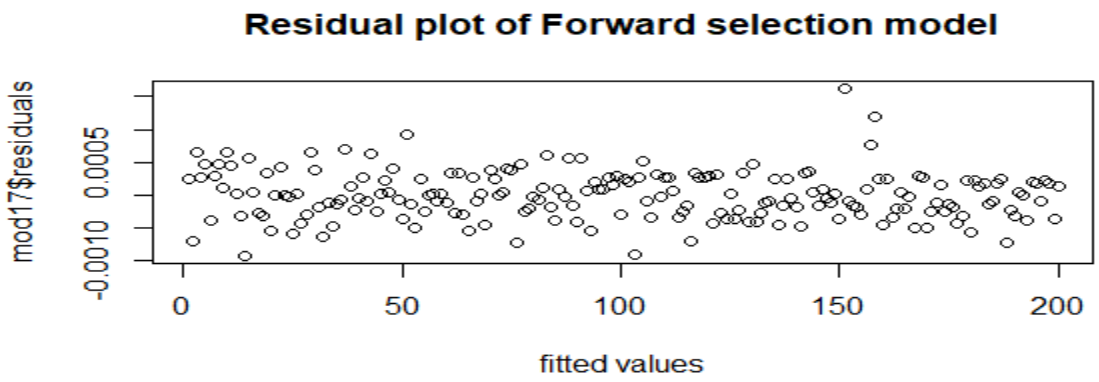


Figure 9: Residual Plot of Forward Selection Model

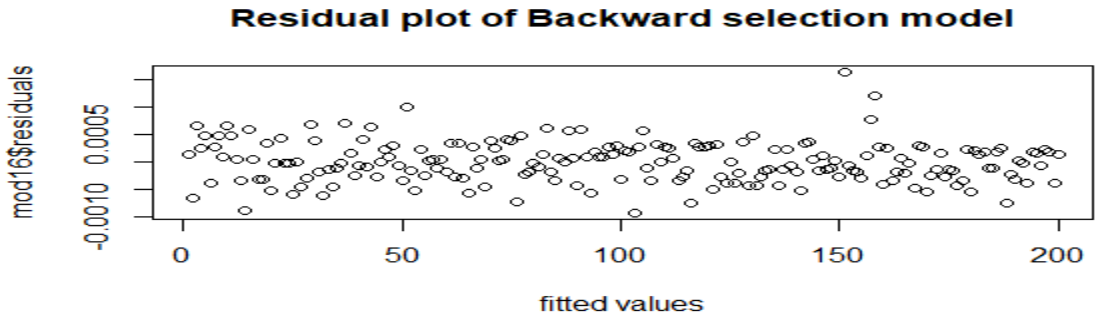


Figure 10: Residual Plot of Backward Selection Model

3.7. Performance

The two models all performed somehow in the same way as shown in Table 8. The backward model outperformed the other models by when their adjusted r-square is compared. Nevertheless, the differences between all the models are small. In addition, it is worth noting that the R^2 value on the models were relatively similar, despite having fewer variables in the backward model than the other model. Considering that the prediction potential of the models is very close, it could be more appropriate to look at the spread of error.

The spread of error cost can be evaluated by taking into account the scatter plots of the value of the residual. The R^2 value of 42.31% and 41.06% indicates that almost 42% of the error could be explained for by this phenomenon. The errors in the prediction by the model seems to

suggest that some vital cost drivers of the construction cost are not being properly represented. The vital cost drivers may have been omitted from the data set. This could be due to missing cost driver variables were not captured in the database. The scatter plots seem to follow the assumptions made in the selection of the dependent variable. The R^2 value of the trend line is also similar to that found in the scatter plots. This indicates that the proportion of error associated with this phenomenon appears to be the same for both forward and backward models.

Table 8: Performance of Regression Models

	Forward Regression Model	Backward Regression Model
R^2	0.4231	0.4106
F-Statistic	9.69	19.11
Significance (p-value)	5.471E-16	< 2.2E-16
Residual standard error	0.0003694	0.0003665

In assessing the cost models, the measures employed was the R^2 on cost. The R^2 on cost value is the value of R^2 gained when the models' prediction is expressed as the raw cost of the project. The adjusted R-squared value is used to measure the performance of a regression model developed. The adjusted R-squared value ranges from 0% to 100%. If the value gets closer to 1, that indicates that the model explains the data better. The adjusted R-squared value increases only if the addition of a new input/independent variable improves the performance of the existing model. An R-squared of 42.31% and 41.06% were recorded for the three models, which were consistent with previous studies and showed proportion of variance of the response variable explained by the predictor variables (Jeong and Woldesenbet 2012).

The adjusted R-squared value of 42.31% shows that the independent variables in this model can explain variability between the independent variables and the dependent variable up to 42.45%. This result was consistent with the argument of Hollar et al. (2013) that the performance of regression model is a little lower than the other cost estimation methods such as Artificial Neural

Network (ANN) Hollar et al. (2013). This can be attributed to inaccuracy of data and relatively the low number of data points used in this study. In this study, only 200 projects are considered in model development. Furthermore, Ritz (2004) proposes an acceptable accuracy range of ±25-30% for construction cost estimates prepared prior to the project's conceptual design. However, the expected prediction error of a simplified model would be smaller than that of an extended model under the following conditions:

1. When the data are very noisy
2. When the true absolute values of the left-out parameters are small
3. When there is high multicollinearity among the predictors; and
4. When the sample size is small, or the range of input values is small. The accuracy of the models could be attributed to these conditions that Wu et al (2008) highlighted.

On the other hand, these results were consistent with the argument of Magdum et al. (2018) that the performance of regression model is a little lower. This study found some problems in the analysis due to relatively small number of data, poor data uniformity, and poor data distribution pattern. Problems related to poor of uniformity of data and data distribution pattern are caused by several factors such as compilation of data from different types of bridges, causing differences in design, work quantity, and work item price b. Possibility of error analysis. The problems in the data will affect the analysis process as well as influence the magnitude of error value in modeling results. This was consistent with Ji et al (2010), who states that generally, the data of construction project cost contains noise, such as internal error and abnormal values which can negatively affect the accuracy of cost estimation.

3.8. Conclusion

The conceptual cost estimate method presented herein addresses the construction cost of bridges in Wisconsin State. The MLR model for the construction cost of bridges includes four numerical and four categorical variables, representing project specific parameters and interactions. Although the prediction error percentage was greater than desired (42%), the regression analyses did yield useful insights into construction cost-estimating. The equation is able to predict the construction cost of bridges from its design parameters. The proposed regression equations were statistically checked regarding their significance and the results confirmed the proposed equations' ability to capture 42% of the variables' variability. Furthermore, relevant statistical checks confirmed that the data sample used for the development of the equations was partially free from the multicollinearity problem, while the assumptions of the correct application of the regression methodology were verified. This research provides a cost prediction model for bridges which is particularly valuable for the projects' stakeholders, as it only requires as input, data available at the early design stage.

CHAPTER 4. CONCEPTUAL COST ESTIMATION MODELS OF WISCONSIN BRIDGES: A COMPARISON OF THE ACCURACY OF DECISION TREE AND RANDOM FOREST MODELS

4.1. Abstract

Inaccurate cost estimates have substantial effects on the final cost of construction projects and erode profits. Cost estimation at conceptual phase is a challenge as inadequate information is available. For this purpose, approaches for cost estimation have been explored thoroughly, however they are not employed extensively in practice. The main goal of this paper is to compare the performance of various models in predicting the cost of construction bridge projects at early conceptual phase in the project development. In this study, on the basis of the actual project data, two modeling algorithms which include multiple linear regression, decision tree and random forest are used to forecast the construction cost of Wisconsin bridge projects. The two models were then compared based on the R-squared and Mean Absolute Percentage Error. The findings revealed that random forest outperforms multiple linear regression and decision tree in realizing better prediction accuracy. The R-squared of decision tree and random forest random forest cost model were 47.70% and 52.70% respectively. It is anticipated that a more reliable cost estimation model could be designed in the early project phases by using a random forest regression technique in the development of a bridge construction cost estimation model. In conclusion, the practitioners in the Department of Transportation can make sound financial decisions at the early phases of the project development in Wisconsin.

4.2. Introduction

A conceptual cost estimate is dependent on the conceptual design of projects at the early design phase with least scope definition (Fragkakis et al. 2011). The conceptual cost estimates are

employed for variety of reasons, such as, determining the feasibility of projects, developing the initial budget and financial assessment, and evaluating alternative projects (Sonmez 2004). Conceptual cost estimate provides the lowest predicted accuracy due to the minimal details available, but the highest degree of complexity and necessity. Regardless of the lack of information about the project at the conceptual phase, public institutions need these cost estimates for statewide fiscal funding provisions (Anderson et al. 2007). The success of a construction project is determined by meeting the client's standards in terms of schedule, cost, and quality of work. For decision-makers to monitor the overall project, accurate cost estimate in the preliminary stage is critical (Magdum et al. 2018).

In addition, the importance of early estimation from the viewpoint of owners and related project teams cannot be over-emphasized (Kim et al. 2014). Adequate estimation of construction cost is key factor in any type of construction projects. However, based on the analysis of 258 transport infrastructure projects worth $90 billion (U.S.), it was found that in the vast majority of projects actual costs were significantly higher than initially estimated, e.g., 34 % higher on an average for bridges and tunnels (Bouras 2018). From past research studies, it is evident there exists a challenge in accurately estimating the cost of projects at the conceptual phase (Bouras 2018; Chau et al 2018; Mayer et al. 2019). In order to curb the challenge and estimate the construction project cost more accurately and rapidly, this study puts forward a method of cost estimation of construction project based on machine learning algorithms. So, the study is going to discourse our work on predicting the cost of bridge construction projects with few project features or attributes. This is to provide all contracting parties accurate information about the expected cost of bridge projects at its early phase with minimal errors. Upon the completion of the different modelling algorithms, an evaluation of each model is conducted by comparing its accuracy. The goal of the

paper is the assess cost estimation models by using historic data to predict construction cost at the conceptual stage and evaluate the predictions of decision tree and random forest model.

4.3. Literature Review

Predicting construction costs is one of the most important preliminary steps in any construction project since cost prediction is crucial to avoid construction delays and ensure successful project completion (Rafei 2018). Various estimation techniques and methods are available. Studies of project performance prediction in the construction industry have employed a variety of different approaches. The use of regression techniques is a well-established approach to the prediction of project performance in the construction industry (Elmousalami 2019; Nouralli and Osaloo 2020). Regression methods and Artificial Neural Network (ANNs) have been successfully applied to the prediction of project performance in the construction industry. However, the efficiency of such prediction methods is limited. Regression techniques require a large amount of statistical data; moreover, the accuracy of such approaches is affected by the generally held assumption that the independent variables are independent of one another and normally distributed.

With the improvements in computing capability, the latest cost estimating techniques tend to use more complex approaches and larger data sets. Machine learning algorithms as part of artificial intelligence, which allow exploring multi- and non-linear relationships between variables and final costs, have been employed in recent years (Mostafa 2003; Hollar et al. 2013; Marinelli et al. 2015).

In research performed for Texas DOT in the early 2000s, Chou et al. (2005) developed a probabilistic cost estimation tool that focused on 22 major work items that accounted for roughly 80% of total cost. Unlike other traditional models that are affected by untreated historical data, the

probabilistic model developed by Chou et al. (2005) provided confidence bounds for an estimate, which helps control error, accounts for probability, and considers the independent or correlated relationships between the major work items. As with any other estimating method, the effectiveness of probabilistic models hinges on the quality of the data available to estimators. (Behmardi, et al. 2013). However, these statistical approaches focused solely on prediction of aggregate costs using historical data and have neglected prediction methods directly incorporating construction trends, economies of scale, and many site-specific factors, such as expected production rates and labor and material costs.

Jeong and Woldesnbet (2012) used historical data provided by Oklahoma DOT to estimate the preliminary engineering (PE) costs of roadway projects. Three models were developed, which were regression models, decision tree models and neural network models to obtain the optimum roadway PE cost model. The results of this project are expected to significantly influence how efficiently and economically highway projects are planned, executed, and managed in the early stages of a project. From the study, findings indicated that, based on the comparisons, the neural network model performs better than the regression and decision tree models for 6 major plan development task outputs: summarizing sheets, storm control, construction sequence, detail sheets, pay item quantities, and preliminary engineering cost. With regard to plan and profile, cross-section sheets, and mass diagram sheets, the regression model outperforms decision tree and neural network models. The decision tree models outperform the neural network and regression models in drainage, traffic sheets, and typical section.

Decision trees are models that show sequences of decisions about attributes in the branches leading to predictions in the leaves (Witten, 2017). Models resulting from decision tree analysis predict the value of a root or target variable using input variables. The source dataset is split into

nodes from the root node based upon classification features using recursive partitioning, where the subgroups are split in a manner that classifies them into groups (Denison et al. 2002). In binary recursive partitioning, the tree is split into two nodes: a group that has the same features as the target value, and a group that does not, based upon a decision criterion (which can be viewed as a yes/no question) at each node. The recursive partitioning is halted when splitting a subset no longer improves the quality of the model or some pre-determined stopping criteria are met. Ngo et al. (2020) opined that top application of decision tree as a tool were found to be cost estimation, delay prediction, and energy management, and consumption prediction.

In the use of alternative cost estimation methods, random forest has been adopted by a few studies for cost prediction as well. The random forest (RF) is a bagging ensemble learning model that can produce accurate performance without overfitting issues (Breiman 2001). RF algorithms draw bootstrap samples to develop a forest of trees based on random subsets of features. Therefore, some features may be selected more than once, whereas others might never be selected (Breiman 2001). RF is more robust against noisy data or big data than the decision tree algorithm (Dietterich 2000). The key limitation of the RF algorithm is that it cannot interpret the importance of features or the mechanism of producing the results. RF does not search for the best split variables that diminish the correlation among the developed trees and the strength of every single tree. As a result, RF decreases the generalization error (Breiman 2001). However, RF cannot interpret the produced predictions. An extremely randomized tree (ERT) algorithm merges the randomization of random subspace to a random selection of the cut-point during the tree node splitting process. Extremely randomized tree mainly controls the attribute randomization and smoothing parameters (Geurts et al. 2006). Wang and Ashuri (2016) have developed a highly accurate model based on random tree ensembles to predict a construction cost index in which the model's accuracy has

reached 0.8%. Miljan et al (2020) investigated Seven state-of-the-art machine learning techniques for estimation of construction costs of reinforced-concrete and prestressed concrete bridges. The techniques were artificial neural networks (ANN) and ensembles of ANNs, regression tree ensembles (random forests, boosted and bagged regression trees), support vector regression (SVR) method, and Gaussian process regression (GPR).

A database of construction costs and design characteristics for 181 reinforced-concrete and prestressed-concrete bridges is created for model training and evaluation. Although ensemble methods, such as ensembles of ANNs, regression tree ensembles using boosting and SVR with RBF kernel, perform well, they require a considerable amount of time to train the models, especially if the number of base models in the ensemble is high. The findings confirmed that methods based on machine learning eliminate the biases introduced by human factor and offer a fast and reliable tool for the construction industry to estimate construction costs of concrete bridges, even in early implementation stages, when only the basic technical and economic characteristics are available.

The main problem in estimation of transport infrastructure project costs is significant deviation between the estimated costs and the real, actual construction costs, due to underestimation in the initial project phases, when the costs are evaluated in order to decide whether the transport infrastructure should be built. Therefore, there is a definite need for prediction methods that are more robust and more reliable.

4.4. Methodology

Random Forest (RF) is an ensemble machine learning techniques by combining and averaging the results from multiple decision trees. Random forests models are used for classification purpose. Random forests are a combination of tree forecasters such that every tree

depends upon the values of an independently sampled random vector where same distribution is used for all trees in the forest. This technique consists of generating a set of trees that vote for the most prevalent class. There are two important characteristics of using Random forests. The first characteristic is that the generalization error converges with the increase in number of trees increases and second characteristic is that this type of learning does not suffer from over fitting.

A decision tree model is selected as one approach for directed knowledge discovery to develop cost prediction models. A decision tree model is a tree like structure that predicts target variables through a set of prediction rules (Berry and Linoff, 2010). Decision trees are drawn with a root node at the top by taking all the data and splitting it into branches or decision nodes. This process continues until it reaches the bottom node, or leaf node, based on the values of independent variables. During the splitting process, for each split or decision node or leaf, the number of observations is recorded, and the observation which has higher nodes is distributed to the lower nodes. A decision tree model is selected as one approach for directed knowledge discovery to develop cost prediction models. A decision tree model is a tree like structure that predicts target variables through a set of prediction rules (Berry and Linoff, 2010). Decision trees are drawn with a root node at the top by taking all the data and splitting it into branches or decision nodes. This process continues until it reaches the bottom node, or leaf node, based on the values of independent variables. During the splitting process, for each split or decision node or leaf, the number of observations is recorded, and the observation which has higher nodes is distributed to the lower nodes. In addition, decision trees are widely used for solving classification problems. The decision tree is constructed continuously based on the feature that best satisfies the branching rule. This process is then performed iteratively for each branch (Djukova and Peskov 2007). Classification and regression decision trees deal with predicting a dependent variable based upon a predictor

variable. The response variable in the former includes a finite set of values, while in the latter contains continuous or discrete set of variables (Loh 2008). Decision trees are good substitution for basic regression methods. Decision trees are mainly constructed based on those attributes in the dataset that are pertinent to the classification case, thus it can be mostly regarded as a feature selection problem (Perner 2015). According to Karca et al (2020) a decision tree is a classifier that works with recursive partition of the instance space. It is used to represent a supervised learning approach. It is a simple graphical model where non-terminal nodes represent tests on one or more attributes and terminal nodes give decision outcomes. This tree consists of one root, branches, internal nodes, and leaves. Each node corresponds with a certain feature or characteristic or feature and the branches correspond with a range of values or decision outcomes. A decision tree works with local regions that are identified in a series of recursive splits in a smaller number of steps that implements divide and conquer paradigm. A decision tree works with input data and uses decision rules for future predictions.

To develop a prediction model using decision trees and random forests, this study consisted of data preparation and running the data through R-Programming. Raw data were processed before being input into a training model. Randomly selected dataset from the training set is used and a random subset of that is used at each step to grow a decision tree. To predict new observations from the test set, each case is funneled through each decision tree. The results are aggregated and averaged to produce the test prediction. Results from an individual decision tree tend to overfit the training data leading to poor accuracy for prediction of new data. By averaging multiple decision trees, RF has the added ability to improve flexibility and accuracy while reducing overfitting issues. Due to the random nature of the decision process, RF is considered a "black box" method because it is difficult to gain a full understanding behind the specific rationale and logic creating

decision trees (Zhang & Wang, 2009). This also means relatively lower training and prediction speeds compared to other machine learning-based algorithms. However, the advantages by using combinations of a variety of decision trees is that it handles irrelevant features without overfitting and have relatively high prediction accuracy (Zhang & Wang, 2009).

4.5. Measuring Prediction Accuracy between Models

The data-driven construction estimating methods under consideration for this study rely on historical project cost databases in relating various project attributes to actual construction project costs. It is worth noting that, once an estimator has confidence that the identified cost relationships are stable and reasonably robust, the resulting cost function can then be used to forecast future costs of construction projects. Thus, two different metrics were used to compare the best model configuration between decision tree and random forest to ultimately determine the technique that produced the best prediction results based on the provided dataset. The Mean Absolute Percent Error (MAPE) and R-squared was chosen as the error measures. MAPE denotes the overall average of deviations between predicted and actual estimates in absolute values expressed as a percentage of the actual estimate. Even though, the performance of prediction models will depend on the soundness of the underlying assumptions (such as the linearity and continued validity of the relationships).

The optimized configuration for each method that generated the highest accuracy was subsequently compared against each other on the basis of several metrics based on the independent test set. The metrics used were the following:

- Coefficient of Determination or R Squared (R^2) measures the degree for which the response explains the predictor variables. It is generally defined by:

$$R^2 = 1 - \text{RSS} / \text{TSS}$$

where R^2= coefficient of determination

RSS= sum of squares of residuals

TSS= total sum of squares

- Mean Absolute Percent Error (MAPE)

$$\text{MAPE} = (100/\text{n}) \sum |Pi - Ai| / Ai$$

Where, n = the number of testing data-points

P_i = the predicted construction cost

A_I = the actual construction cost.

4.6. Results and Discussion

This section presents the results of comparing the prediction accuracy of random forest model relative to decision tree model. The r-squared value is used to measure the performance of a regression model developed. The r-squared value could be between 0 and 100%. If the R-square value is close to a hundred% or 1 then, it shows that the model explains the data better. An R-squared of 37.70% and 32.60% for the DT model and RF model respectively were consistent with previous studies (Shin 2018). The results of adjusted R-squared value of 37.70% and 32.60% indicated that the independent variables in this model can explain variability between the independent variables and the dependent variable up to 37.70% and 32.60% respectively.

A decision tree model for estimating construction cost is shown in Figure 11. Based on the model, the roadway width is the root node for splitting, which further splits into deck area, piling size and deck width on another branch. Random forest model developed twenty (20) trees. Seven

(7) of the trees had roadway width as the root node, five (5) trees had location as the root node, five trees had deck area as root node, one tree had span length and two trees had bridge length as the root nodes. From fig 2, it is evident that the variable significance varied across both decision tree and random forest models. Deck area had the highest weight in the development of the decision tree model while bridge length had the highest weight in the development of random forest model.

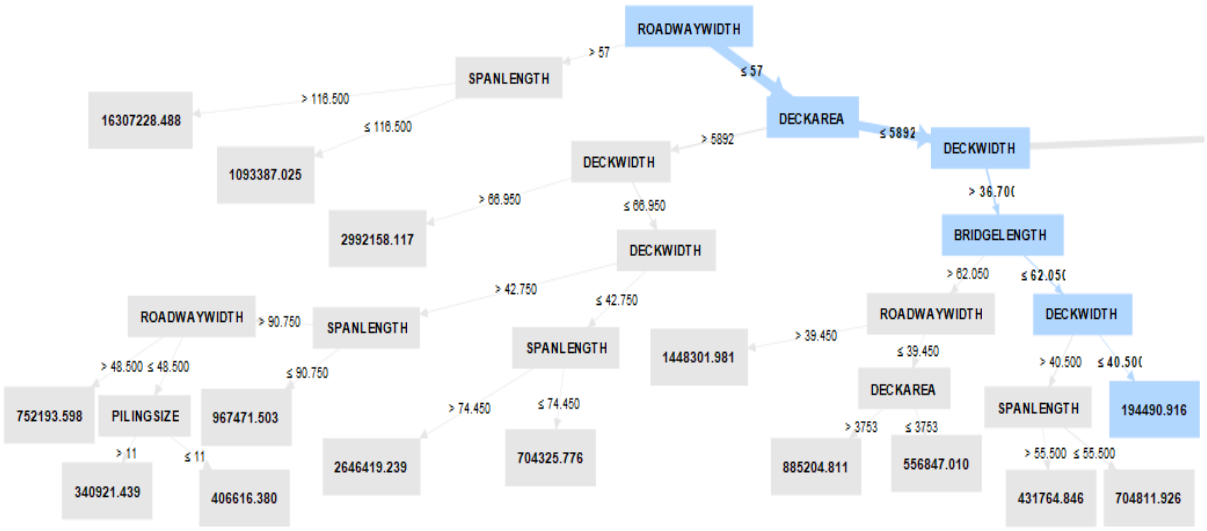


Figure 11: A Section of the Decision Tree Model

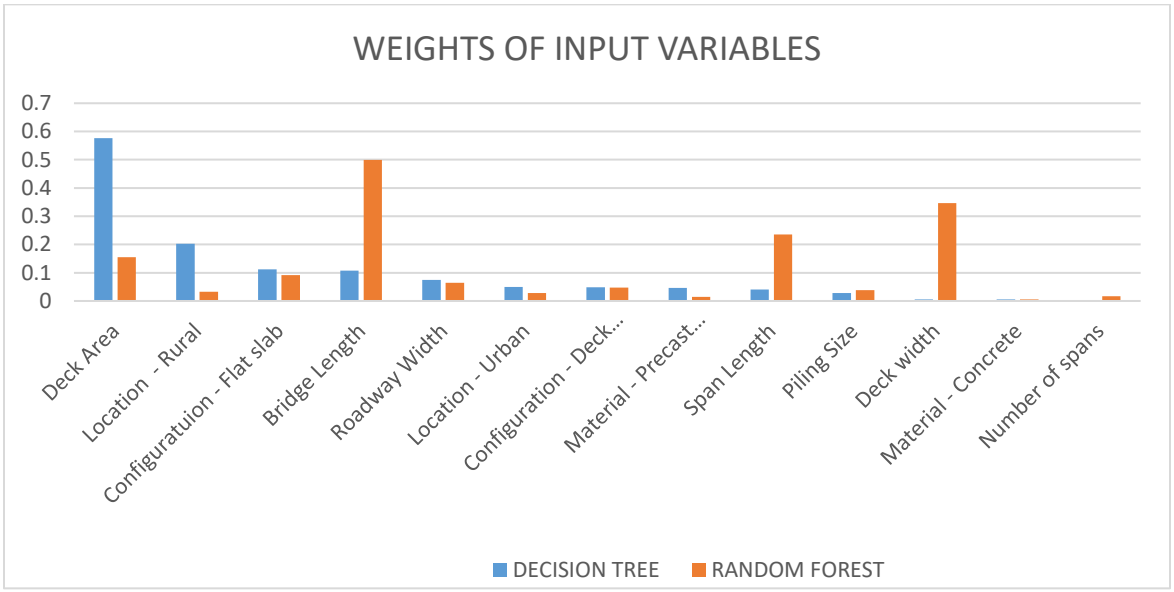


Figure 12: Weights of Input Variables

4.6.1. Evaluation of performance

This section presents the results of comparing the prediction accuracy of the two model. On balance, the performance of the two models were evaluated using the model accuracy measures tabulated in table 9

Table 9: R-square and MAPE Results

Model	R-Square	MAPE
Decision Tree	37.70%	19.71%
Random Forest	32.60%	15.03%

MAPE values for all estimating models, shown in Table 9, indicate that the proposed models are able to predict the actual cost with an average error of less than 30%. This error is considered acceptable According to National Cooperative Highway Research Program (NCHRP) research report, conceptual estimating techniques with few project definitions can produce a project estimate with an accuracy range of +40/-20% to +120/-60%. Similar accuracy ranges are reported in the Association for the Advancement of Cost Engineering International (AACE) Total Cost Management Framework (Stephenson 2015) which propose an accuracy range of ±40% for planning/ feasibility estimates prepared prior to conceptual design**.** The results may be evaluated based on the potential impact of left-out parameters, multicollinearity of input parameters, and sample size of training data sets.

The importance of each dependent variable to cost estimation was given, as shown in Figure 12. These values indicate the importance of each variable for the construction cost estimation in the model. Finally, the tree structures in the model were provided as shown in Figure 11. This shows the estimation rules, such as the applied variables and their influence on the proposed model. In terms of the estimation accuracy, the RF model showed slightly better results than the DT model. These results mean that the RF model has remarkable performance and

moreover, the RF model provided additional information, that is, an importance plot and structure model, which helps the estimator comprehend the decision-making process intuitively. Consequently, these results reveal that a RF has potential applicability in preliminary cost estimations. It can assist estimators in avoiding serious errors in predicting the construction costs when only limited information is available during the early stages of a building construction project. Based on the analysis and model comparison, the random forest model outperforms and decision tree models.

4.7. Conclusion

Several factors, including the dynamic nature of bridge projects for many construction inputs and cost factors unique to individual construction projects, pose challenges for an accurate estimate of construction project costs. The need for early estimates further compounds the problem because an exact specification of the final cost function is not attainable at the early stages of planning and design. As a result, public agencies, which are often faced with considerable challenges to improve the accuracy of their early cost estimates, may prefer simplified models over relatively more extensive approaches. This research project developed two different methods including a decision tree model and random forest model to assist in estimating bridge construction costs. The adjusted R^2 value ranges were 37.7% and 32.60% for DT model and RF model respectively. The performance of the models measured by using the MAPE were 19.71% and 15.03% for DT model and RF model respectively where the best performance of 15.03% was obtained from the random forest model. The results clearly revealed that RF has more accuracy in prediction with less error value when compared with DT. It can be concluded that the prediction done with RF portrays a strong degree of coherency with actually collected cost data of bridge project against DT. So, this study will be helpful the contracting parties in the bridge construction

industry and the future works. Finally, as the relatively superior prediction power of early estimating models presented in the study indicates, such improvements are expected to be of considerable value to public agencies tasked with setting budget allocations based on early estimates. Furthermore, as the paper demonstrates, such agencies stand to reap sizable benefits from a review of their early estimating practices, as long as such models can be kept current through continuous monitoring and calibration of their prediction performance to reflect shifting agency needs and priorities.

CHAPTER 5. CONCLUSIONS AND RECOMMENDATION

5.1. Introduction

Department of Transportation are required to provide transportation infrastructures according to projected budget and time allocations. DOTs need various cost estimates for different functions during the life cycle of bridge projects. Consequently, there is a considerable interest in ensuring accuracy and reliability in cost estimation of bridge construction particularly at the conceptual stage due to its inherent challenges. Despite the numerous research studies that are carried out, the problem of inaccurate estimation of construction costs of bridge projects still persists.

In order to contribute to the alleviation of the problem, this research study explored past articles on cost estimation of Bridge projects to determine research trends, identify the cost variables that are adopted in cost estimation models and influence the construction cost of bridges, and ascertain the impact of cost estimation methods in the accuracy of the prediction of construction cost. This study employed three cost estimation methods, namely regression analysis, decision tree and random forest. This chapter provides a summary of the study findings to ascertain whether the objectives and research questions outlined in the introductory section of the thesis have been met. Furthermore, the chapter presents the implications of the study findings and gives suggestions for future research.

5.2. Findings and Conclusions of Systematic Literature Review

This research study explored a systematic review of previous articles to determine research trends, identify the cost variables that are adopted in cost estimation models and influence the construction cost of bridges, and ascertain the impact of cost estimation methods in the accuracy of the prediction of construction cost. Selected journals from 1990 to 2019 were identified, giving

rise to 29 articles on estimation of bridge costs. The trend of conceptual cost estimation research with regard to the distribution of publications per year was analyzed. Furthermore, the methods used in estimating construction costs at the conceptual stages of bridge construction cost were categorized and the cost drivers adopted for cost estimation models in estimating construction cost of bridges were identified and ranked. The results indicate that despite the steep increase in the number of publications between 2009 and 2010, the number of studies published from 2010 to 2019 fluctuated during that period. This indicates there is a need to reinforce research efforts to enhance conceptual cost estimation methods in view of the rapid advancement of transportation infrastructure globally. The findings from the study disclosed that the top three cost estimation methods adopted for bridge construction cost were unit price/quantity of standard work, regression analysis, artificial neural network. A few research studies combined case-based reasoning and genetic algorithms; multiple regression analysis and artificial neural networks; artificial neural network and genetic algorithms; artificial neural networks and support vector machines; factor analysis and multivariate regression; fuzzy expert system, and bridge information management system to augment the proficiencies of single methods. In relation to the cost drivers, the most commonly used variables included the weight of steel, the volume of concrete, number of spans, type of foundation, which are materials and characteristics of the bridge. These influence the models that are used for the cost prediction. The results of the meta-analysis indicated that the type of cost estimation method adopted for predicting the estimates of bridge projects impacts the accuracy of the prediction.

5.3. Findings and Conclusions of Multiple Linear Regression Model

The conceptual cost estimate method presented herein addresses the construction cost of bridges in Wisconsin State. The MLR model for the construction cost of bridges consisted of

variables, representing design parameters and interactions. The summary of findings is listed as follows:

- Although the prediction error percentage was greater than desired (42%), the regression analyses did yield useful insights into construction cost-estimating. The equation is able to predict the construction cost of bridges from its design parameters.
- Even though, two out of nine input variables were statistically significant (p-value< 0.05) for the prediction of construction cost. The remaining variables showed no significant relationship. The proposed regression equations were statistically checked regarding their significance and the results confirmed the proposed equations' ability to capture 43% of the variables' variability. The model was found to be statistically significant with a p-value less than 0.05 with estimates of 1.82e-15.
- Furthermore, relevant statistical checks confirmed that the data sample used for the development of the equations was partially free from the multicollinearity problem, while the assumptions of the correct application of the regression methodology were verified.

5.4. Findings and Conclusions of Decision Tree and Random Forest Model

This research study developed decision tree model, and random forest model to assist in estimating bridge construction costs. The summary of findings of the two proposed models are as follows:

- Similar R-square values of 0.3770 and 0.3260 were recorded for decision tree and random forest respectively.
- MAPE values of 19.71% and 15.03% were recorded for decision tree and random forest respectively and showed very similar measures of errors for all models.

- The variable significance varied across both decision tree and random forest models. Deck area had the highest weight in the development of the decision tree model while bridge length had the highest weight in the development of random forest model.
- The results clearly revealed that random forest has more accuracy in prediction with less error value when compared with decision tree. It can be generalized that the prediction done with random forest portrays a strong degree of coherency with the actual cost data of bridge project against decision tree.

5.5. Contributions to the Body of Knowledge

- This research proposed three models using multiple linear regression, decision tree and random forest. The proposed model is not only efficient to the problem in the thesis, but also applicable to other prediction problems.
- The proposed models contribute to a more accurate cost estimation. the random forest model outperforms the other two models in three prediction scenarios. The random forest provides a more accurate cost forecast compared to the decision tree and multiple linear regression. In this field of research problem, a little improvement of the prediction suggests a huge advancement of the accuracy in budget estimation.
- The proposed random forest and decision tree models can efficiently handle both numerical and categorical variables. Not like linear regression, the categorical variables need to be treated as several dummy variables, or in some other methods the preliminary manipulation is required to manually transform categorical variables into numerical ones, the implemented algorithms in this research could automatically detect the categories with built-in converting function.

- Furthermore, the models are quite interpretable and easy to understand. They can also be used to identify the most significant variables in your data set.
- In summary, the proposed models provide great benefits to state DOTs in preparing more accurate budgets and cost estimates for bridge projects.

5.6. Recommendations

The recommendations outlined in this section, if implemented, could streamline the modeling process for any future model updates.

- From the literature review and meta-analysis in chapter 2, areas of possible improvement in the way the studies report bridge construction estimation performance metrics were identified. The studies reviewed in the literature review section employed various statistical measures to assess the performance of the cost estimation model which limited the section of articles to be used for the meta-analysis for the accuracy comparison. Thus, the different performance metrics provided do not facilitate an extensive comparison of results among empirical bridge cost estimation studies.
- To retain prediction accuracy, the models recommended in this study should be updated periodically using data from recent bridge projects. When creating or updating a regression model, a dataset of the highest quality reasonably obtainable is essential for ensuring that the model can produce reliable and accurate estimates. Thus, data compiled in the HIS data website for WsDOT as well as any data repository of other DOTs should be fully complete with parameters as well as all the other essential information about the bridge projects.
- Cost estimation models with higher level of accuracy can be achievable in different stages of project development by considering more data and input variables from

various projects. Additional pertinent input variables could make the model outcome more reliable.

5.7. Limitations and Future Research

- The proposed cost estimation models have limitations that should be addressed. The reason is idealizations and assumptions have been made during their development. In particular, the multiple linear regression model, decision tree and random forest conceptual cost estimation models, explained in this research, were developed by a limited number of executed bridge projects; and estimate of project costs assigned for future projects based on historical data gathered from these projects. Consequently, the model needs to be applied for projects have similarity to analyzed projects so that the outcome can be trustworthy. The availability of more project data can help to develop a more accurate and reliable estimation.
- Also, the model considered only a particular type of bridge, which limits the model in terms of the variables that were considered. The variables employed were specific to the type of bridge. Thus, it may not be useful for some bridges if the input variables of other types of bridges were not in included.
- To examine the disparities between a current cost estimate to previous estimates for bridge construction projects, there is a demand for uniform estimation performance metrics. Future research could focus on establishing guidelines for developing and applying cost-estimating performance metrics and generate additional performance metrics to assess the accuracy of bridge cost estimation models. Thus, new models that will be developed could be evaluated according to a standard performance metric.

- The dataset creation process used in this project involved several additional steps because of the lack of other design parameters information in the database that was used in developing the models, the error percentage would have been minimized and potentially improving the models.
- The selection of the suitable model should have also been dependent on the time efficiency in using the model to get a prediction. Based on the time frame required for the cost estimate to be ready, an efficient model will be needed to give prediction within the shortest period of time. Further studies could research on the selection of a suitable model based on time efficiency.

REFERENCES

Ahn, E., & Kang, H. (2018). Introduction to systematic review and meta-analysis. *Korean Journal of anesthesiology*, *71*(2), 103.

Adel, K., Elyamany, A., Belal, A. M., & Kotb, A. S. (2016). Developing parametric model for conceptual cost estimate of highway projects. *International Journal of Engineering Science*, *6*(7), 1728-1734.

Anderson, S., Molenaar, K. R., and Schexnayder, C. J. (2007). Guidance for cost estimation and management for highway projects during planning, programming, and preconstruction. National Cooperative Highway Research Program (NCHRP) Rep. 574, Transportation Research Board, Washington, DC.

Antoniou, F., Konstantinidis, D., & Aretoulis, G. N. (2016). Analytical formulation for early cost estimation and material consumption of road overpass bridges. Research Journal of Applied Sciences. Engineering and Technology, 127, 716-725.

Bouras CB-T. (2018) Regression models to house price prediction.

Bakhoum, M., Morcous, G., Taha, M., and El-Said, M., (1998). Estimation of quantities and cost of prestressed concrete bridges over the Nile in Egypt. Journal of Egyptian Society of Engineers/Civil, 37(4), pp.17-32.

Behmardi, B., Doolen, T., & Winston, H. (2015). Comparison of predictive cost models for bridge replacement projects. Journal of Management in Engineering, 31(4), 04014058.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Bouabaz, M., & Hamami, M. (2008). A cost estimation model for repair bridges based on artificial neural network. American Journal of Applied Sciences, 5(4), 334-339.

Chau AD, Moynihan GP, Vereen S. Design of a Conceptual Cost Estimation Decision Support System for Public University Construction. Constr Res Congr 2018, Reston, VA: American Society of Civil Engineers; 2018, p. 629–39. doi:10.1061/9780784481295.063.

Chou, J. S., Wang, L., Chong, W. K., & O'Connor, J. T. (2005). Preliminary cost estimates using probabilistic simulation for highway bridge replacement projects. In Construction Research Congress 2005: Broadening Perspectives (pp. 1-10).

Creese, R. C., & Li, L. (1995). Cost estimation of timber bridges using neural networks. Cost Engineering, 37(5), 17.

Cleophas, T. J., & Zwinderman, A. H. (2017). Modern meta-analysis. *Modern Meta-Analysis. https://doi. org/10.1007/978-3-319-55895-0*.

Dennison, D.G.T., Holmes, C.C., Mallick, B.K., and Smith, A.F.M. (2002). Bayesian Methods for Nonlinear Classification and Regression. John Wiley & Sons.

Dimitriou, L., Marinelli, M., & Fragkakis, N. (2018). Early bill-of-quantities estimation of concrete road bridges: an Artificial Intelligence-based application. Public Works Management & Policy, 23(2), 127-149.

D.J. Lowe, M.W. Emsley, A. Harding, Predicting construction cost using multiple regression techniques, J. Constr. Eng. Manage. 132 (2006) 750-758.

Djukova EV, Peskov N (2007) A classifcation algorithm based on the complete decision tree. Pattern Recognit Image Anal 17(3):363–367

Elfaki, A. O., Alatawi, S., & Abushandi, E. (2014). Using intelligent techniques in construction project cost estimation: 10-year survey. *Advances in Civil Engineering*, *2014*.

Elmousalami, H. H. (2019). Intelligent methodology for project conceptual cost prediction. *Heliyon*, *5*(5), e01625.

Fragkakis, N., Lambropoulos, S., and Pantouvakis, J.P., (2010). A cost estimate method for bridge superstructures using regression analysis and bootstrap. Organization, technology & management in construction: an international journal, 2(2), pp.182-190.

Fragkakis, N., Lambropoulos, S., & Tsiambaos, G. (2011). Parametric model for conceptual cost estimation of concrete bridge foundations. Journal of infrastructure systems, 17(2), 66-74.

Flyvbjerg, B. (2007). Cost overruns and demand shortfalls in urban rail and other infrastructure. *Transportation Planning and Technology*, *30*(1), 9-30.

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport reviews*, *23*(1), 71-88.

Gardner, B. J., Gransberg, D. D., & Jeong, H. D. (2016). Reducing data-collection efforts for conceptual cost estimating at a highway agency. *Journal of Construction Engineering and Management*, *142*(11), 04016057.

Gupta, S. K., Gunasekaran, A., Antony, J., Gupta, S., Bag, S., & Roubaud, D. (2019). Systematic literature review of project failures: Current trends and scope for future research. *Computers & Industrial Engineering*, *127*, 274-285.

Gunduz M, Sahin HB (2015) An early cost estimation model for hydroelectric power plant projects using neural networks and multiple regression analysis. J Civ Eng Manag 21(4):470–477

Higgins, J. P. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International journal of epidemiology*, *37*(5), 1158-1160.

Hollar, D., Arocho, I., Hummer, J., Liu, M. and Rasdorf, W., 2010, March. Development of a regression model to predict preliminary engineering costs. In ITE 2010 Technical Conference and Exhibit: Meeting Transportation's 21st Century Challenges.

Hollar, D. A., Rasdorf, W., Liu, M., Hummer, J. E., Arocho, I., & Hsiang, S. M. (2013). Preliminary engineering cost estimation model for bridge projects. Journal of construction engineering and management, 139(9), 1259-1267.

H.S. Ji, M. Park, H.S. Lee, Data preprocessing based parametric cost model for building projects: Case studies of Korean construction projects, J. Constr. Eng. Manage. 136 (2010) 844-853.

Janakiraman, N., Syrdal, H. A., & Freling, R. (2016). The effect of return policy leniency on consumer purchase and return decisions: A meta-analytic review. *Journal of Retailing*, *92*(2), 226-235.

Jia, J., Ibrahim, M., Hadi, M., Orabi, W., Ali, M., & Xiao, Y. (2016). Estimation of the total cost of bridge construction for use in accelerated bridge construction selection decisions. In Transportation Research Board 95th Annual Meeting (No. 16-6305, p. 17).

Juszczyk, M. (2020). On the Search of Models for Early Cost Estimates of Bridges: An SVM-Based Approach. Buildings, 10(1), 2.

Kermanshachi, S., & Safapour, E. (2020). Gap Analysis in Cost Estimation, Risk Analysis, and Contingency Computation of Transportation Infrastructure Projects: A Guide to Resource and Policy–Based Strategy Establishment. Practice Periodical on Structural Design and Construction, 25(1), 06019004.

Kim, K.J., Kim, K., and Kang, C.S., (2009). Approximate cost estimating model for PSC Beam bridge based on quantity of standard work. KSCE Journal of Civil Engineering, 13(6), pp.377-388.

Kim, K. J., & Kim, K. (2010). Preliminary cost estimation model using case-based reasoning and genetic algorithms. Journal of Computing in Civil Engineering, 24(6), 499-505.

Kim, B. S., & Hong, T. (2012). Revised case-based reasoning model development based on multiple regression analysis for railroad bridge construction. Journal of construction engineering and management, 138(1), 154-162.

Kim, S. B., & Cho, J. H. (2013). Development of the Approximate Cost Estimating Model for PSC Box Girder Bridge based on the Breakdown of Standard Work. Journal of the Korean Society of Civil Engineers, 33(2), 791-800.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, *51*(1), 7-15.

Konstantinidis, D., & Maravas, A. (2003). Egnatia Motorway concrete bridges statistics. In *ASECAP Study and Information Days proceedings of the international conference in Portoroz Slovenia* (pp. 92-109).

Linares-Espinós, E., Hernández, V., Domínguez-Escrig, J. L., Fernández-Pello, S., Hevia, V., Mayor, J., ... & Ribal, M. J. (2018). Methodology of a systematic review. *Actas Urológicas Españolas (English Edition)*, *42*(8), 499-506.

Loh WY (2008) Classifcation and regression tree methods. In: Encyclopedia of statistics in quality and reliability. Wiley, pp 1–13

Mahamid, I. (2011). Early cost estimating for road construction projects using multiple regression techniques. *Construction Economics and Building*, *11*(4), 87-101.

Marinelli, M., Dimitriou, L., Fragkakis, N., and Lambropoulos, S., (2015). Non-parametric bill-of-quantities estimation of concrete road bridges' superstructure: An artificial neural networks approach. In 31st Annual Association of Researchers in Construction Management Conference, ARCOM 2015.

Markiz, N., & Jrade, A. (2014). Integrating a fuzzy-logic decision support system with bridge information modelling and cost estimation at conceptual design stage of concrete box-girder bridges. International Journal of Sustainable Built Environment, 3(1), 135-152.

Markiz, N., & Jrade, A. (2019). Integrating an expert system with BrIMS, cost estimation, and linear scheduling at conceptual design stage of bridge projects. International Journal of Construction Management, 1-16.

Marzouk, M., & Hisham, M. (2012). Applications of building information modeling in cost estimation of infrastructure bridges. International Journal of 3-D Information Modeling (IJ3DIM), 1(2), 17-29.

Meharie, M. G., Gariy, Z. C. A., Ndisya Mutuku, R. N., & Mengesha, W. J. (2019). An effective approach to input variable selection for preliminary cost estimation of construction projects. *Advances in Civil Engineering*, *2019*.

Molenaar, K. R. (2005). Programmatic cost risk analysis for highway megaprojects. Journal Construction Engineering and Management. 131 (3): 343–353. https://doi.org/10 .1061/(ASCE)0733-9364(2005)131:3(343).

Morcous, G., Bakhoum, M.M., Taha, M.A. and El-Said, M. (2001) . Preliminary quantity estimates of highway bridges using neural networks. In Proceedings of the Sixth International Conference on the Application of Artificial Intelligence to Civil & Structural Engineering Computing (pp. 51-52).

Mulrow, C. D. (1994). Systematic reviews: rationale for systematic reviews. *Bmj*, *309*(6954), 597-599.

Neyeloff, J. L., Fuchs, S. C., & Moreira, L. B. (2012). Meta-analyses and Forest plots using a microsoft excel spreadsheet: step-by-step guide focusing on descriptive data analysis. *BMC research notes*, *5*(1), 1-6.

Ngo, M. (2012) UK construction industrys responses to government construction strategy BIM deadline and applications to civil engineering education. in Proceedings of 1st Civil and Environmental Engineering Student Conference.

Oh, C. D., Park, C., & Kim, K. J. (2013). An approximate cost estimation model based on standard quantities of steel box girder bridge substructure. KSCE Journal of Civil Engineering, 17(5), 877-885.

Pelaez, A., Chen, C. W., & Chen, Y. X. (2019). Effects of perceived risk on intention to purchase: A meta-analysis. *Journal of Computer Information Systems*, *59*(1), 73-84.

Perner P (2015) Decision tree induction methods and their application to big data. In: Xhafa F, Barolli L, Barolli A, Papajorgji P (eds) Modeling and processing for next-generation big-data technologies. Springer, Cham, pp 57–88

Qian L, Ben-Arieh D (2008) Parametric cost estimation based on activity-based costing: a case study for design and development of rotational parts. Int J Prod Econ 113(2):805–818

Sonmez, R. (2004). Conceptual cost estimation of building projects with regression analysis and neural networks. Can. J. Civ. Eng. 31 (4): 677–683. https://doi.org/10.1139/l04-029.

Shelby, L. B., & Vaske, J. J. (2008). Understanding meta-analysis: A review of the methodological literature. *Leisure Sciences*, *30*(2), 96-110.

Winalytra, I., Nugroho, A. S. B., & Triwiyono, A. (2018). Cost Estimation Model for I-Girder Bridge Superstructure Using Multiple Linear Regression and Artificial Neural Network. In Applied Mechanics and Materials (Vol. 881, pp. 142-149). Trans Tech Publications Ltd.

Wirtz, J. G., Sparks, J. V., & Zimbres, T. M. (2018). The effect of exposure to sexual appeals in advertisements on memory, attitude, and purchase intention: A meta-analytic review. *International Journal of Advertising*, *37*(2), 168-198.

Woldesenbet, A., and Jeong, H.-S. (2012). Historical data driven and component-based prediction models for predicting preliminary engineering costs of roadway projects. Proc., Construction Research Congress, ASCE, Reston, VA, 417–426.

Yu, W. D. (2006). PIREM: a new model for conceptual cost estimation. Construction management and economics, 24(3), 259-270.