Project 2365 | November 2024

SJSU SAN JOSÉ STATE UNIVERSITY

MTI MINETA TRANSPORTATION INSTITUTE

Estimating Models for Engineering Costs on the State Highway Operation and Protection Program (SHOPP) Portfolio of Projects

Elhami Nasr, PhD
Nigel Blampied, PhD
Tariq Shehab, PhD
Vinit Kanani

CSUTC California State University Transportation Consortium

CALIFORNIA STATE UNIVERSITY LONG BEACH

MINETA TRANSPORTATION INSTITUTE

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation's transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the Mineta Consortium for Equitable, Efficient, and Sustainable Transportation (MCEEST) funded by the U.S. Department of Transportation, the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1 and the Climate Change and Extreme Events Training and Research (CCEETR) Program funded by the Federal Railroad Administration. MTI focuses on three primary responsibilities:

Research

MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce and labor. MTI research publications undergo expert peer review to ensure the quality of the research.

Education and Workforce Development

To ensure the efficient movement of people and products, we must prepare a new cohort of transportation professionals who are ready to lead a more diverse, inclusive, and equitable transportation industry. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business: a Master of Science in Transportation Management, plus graduate certificates that include High-Speed and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

Information and Technology Transfer

MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

Report 24-35

Estimating Models for Engineering Costs on the State Highway Operation and Protection Program (SHOPP) Portfolio of Projects

Elhami Nasr, PhD

Tariq Shehab, PhD

Nigel Blampied, PhD

Vinit Kanani

November 2024

A publication of the
Mineta Transportation Institute
Created by Congress in 1991

College of Business
San José State University
San José, CA 95192-0219

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
24-35		

4. Title and Subtitle	5. Report Date
Estimating Models for Engineering Costs on the State Highway Operation and Protection Program (SHOPP) Portfolio of Projects	November 2024
	6. Performing Organization Code

7. Authors	8. Performing Organization Report
Elhami Nasr, PhD ORCID: 0009-0001-7683-9596 Tariq Shehab, PhD Nigel Blampied, PhD ORCID: 0000-0001-8597-6691 Vinit Kanani	CA-MTI-2365

9. Performing Organization Name and Address	10. Work Unit No.
Mineta Transportation Institute College of Business San José State University San José, CA 95192-0219	11. Contract or Grant No. ZSB12017-SJAUX

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
State of California SB1 2017/2018 Trustees of the California State University Sponsored Programs Administration 401 Golden Shore, 5th Floor Long Beach, CA 90802	14. Sponsoring Agency Code

15. Supplemental Notes

10.31979/mti.2024.2365

16. Abstract

The State Highway Operation and Protection Program (SHOPP) is crucial for maintaining California's 15,000-mile state highway system, which includes projects like pavement rehabilitation, bridge repair, safety enhancements, and traffic management systems. Administered by Caltrans, SHOPP aims to preserve highway efficiency and safety, supporting economic growth and public safety. This research aimed to develop robust cost-estimating models to improve budgeting and financial planning, aiding Caltrans, the California Transportation Commission (CTC), and the Legislature. The research team collected and refined comprehensive data from Caltrans project expenditures from 1983 to 2021, ensuring a high-quality dataset. Subject matter experts validated the data, enhancing its reliability. Two models were developed: a statistical model using exponential regression to account for non-linear cost growth, and an AI model employing neural networks to handle complex relationships in the data. Model performance was evaluated based on accuracy and reliability through repeated testing and validation. Key findings indicated that the new models significantly improved the precision of cost forecasts, reducing the variance between predicted and actual project costs. This advancement minimizes budget overruns and enhances resource allocation efficiency. Additionally, leveraging historical data with current market trends refined the models' predictive power, boosting stakeholder confidence in project budgeting and financial planning. The study's innovative approach, integrating machine learning and big data analytics, transforms traditional estimation practices and serves as a reference for other state highway programs. Continuous improvement and broader application of these models are recommended to further enhance cost estimation accuracy and support informed decision-making in transportation infrastructure management.

17. Key Words	18. Distribution Statement
Cost estimating, Neural networks, Regression analysis, Project portfolio management, State Highway Operation and Protection Program (SHOPP)	No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	72	

Form DOT F 1700.7 (8-72)

DOI: 10.31979/mti.2024.2365

Mineta Transportation Institute
College of Business
San José State University
San José, CA 95192-0219

Tel: (408) 924-7560
Fax: (408) 924-7565
Email: mineta-institute@sjsu.edu

transweb.sjsu.edu/research/2365

ACKNOWLEDGMENTS

The authors acknowledge the valuable advice and assistance provided by personnel from the California Department of Transportation (Caltrans). Caltrans personnel who suggested this research, endorsed the proposal, and met the authors to discuss the progress included James Monroe, Jr., Omar Elkhayat, and Anupkumar Khant.

Dr. Hilary Nixon provided the authors with valuable advice on the procedures to be followed on this Grant and Dr. Lisa Blampied very kindly reviewed and commented on the draft report, as did James Davis, retired Chief Deputy Director of Caltrans, and Robert So, retired Deputy Director for Program and Project Management in Caltrans District 7 (Los Angeles). Messrs. Davis and So also provided valuable advice throughout the project.

This research was funded, partially or entirely, by a grant from the Trustees of the California State University and the prime sponsoring agency State of California AB115 (Chapter 20) of 2017 and contract number 08720115-SJAUX. The authors gratefully acknowledge this support.

Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the California State University system in whole or part, the Trustees of the system, or Caltrans.

Contents

List of Figures

List of Tables

Executive Summary

The State Highway Operation and Protection Program (SHOPP) is essential for maintaining California's extensive 15,000-mile state highway system, encompassing projects like pavement rehabilitation, bridge repair, safety enhancements, and traffic management systems. Administered by Caltrans, SHOPP ensures highway efficiency and safety, and supports economic growth.

This research aimed to develop advanced cost-estimating models to improve budgeting and financial planning for SHOPP projects, benefiting Caltrans, the California Transportation Commission (CTC), and the Legislature. The study analyzed comprehensive data from Caltrans project expenditures from 1983 to 2021, incorporating feedback from subject matter experts to ensure data quality.

Two cost-estimating models were developed: a statistical model using exponential regression and an AI model employing neural networks. Both models were rigorously tested for accuracy and reliability. The findings demonstrated a significant improvement in cost estimation precision, reducing variances between predicted and actual costs, thereby minimizing budget overruns and optimizing resource allocation.

The enhanced models leverage historical data and current market trends, refining predictive accuracy and increasing stakeholder confidence in project budgeting and financial planning. This innovative approach integrates machine learning and big data analytics, transforming traditional estimation practices. The research team recommends continuous model improvement and broader application to further support informed decision-making in transportation infrastructure management.

1. Introduction

1.1 Overview of the SHOPP

The State Highway Operation and Protection Program (SHOPP) is a critical component of California's transportation infrastructure strategy, focusing on maintaining and preserving the state highway system. Administered by the California Department of Transportation (Caltrans), SHOPP is a four-year program updated biennially to ensure the continued efficiency and safety of highways across the state. The program encompasses projects such as pavement rehabilitation, bridge repair, safety enhancements, and traffic management systems, reflecting significant investment to support and sustain the state's transportation infrastructure. By continually repairing and rehabilitating the State Highway System (SHS), SHOPP protects the substantial investment made over decades to manage approximately 15,000 miles of SHS. This extensive network includes interstate routes, numbered highways, and other state-owned assets such as bicycle and pedestrian facilities, culverts, Transportation Management Systems, safety roadside rest areas, and maintenance stations. The program funds safety and condition improvements, damage repairs, and highway operational and modal enhancements on the SHS. SHOPP projects focus on capital improvements that enhance existing infrastructure without adding new highway lanes. The program also addresses compliance with the Americans with Disabilities Act (ADA) and stormwater control requirements, ensuring accessibility and environmental compliance. Thus, SHOPP is vital for sustaining and improving California's transportation infrastructure, ensuring it remains safe, efficient, and resilient for all users (Caltrans, 2024a).

1.2 Importance of Cost Estimation in Highway Operations

Accurate cost estimation is fundamental to the successful execution of highway operations and maintenance projects. It provides a basis for budgeting, resource allocation, and financial planning, ensuring that projects are completed within their allocated budgets and timelines (Nevett and Goodrum, 2022). For large-scale programs like the SHOPP, precise cost estimation is essential to maximizing available funds, avoiding cost overruns, and ensuring timely project delivery. Inaccurate cost estimates can lead to significant project delays, budget shortfalls, and inefficient use of resources, ultimately impacting the reliability and safety of transportation infrastructure. Therefore, the ability to predict costs accurately is vital for maintaining the integrity of the highway system and delivering value to taxpayers.

1.3 Defining Portfolio and Portfolio Management

Before discussing the scope of this study, it is essential to define "portfolio" and "portfolio management." According to Wei et al., most of the Caltrans budget is dedicated to projects, each falling into a "program component" established by the Legislature. The term "portfolio" describes collections of projects similar to what Caltrans refers to as a "program component" (Wei et al., 2023). According to the Project Management Institute (PMI), a portfolio is "a collection of projects, programs, subsidiary portfolios, and operations managed as a group to achieve strategic objectives" (PMI, 2017). In addition to an overall project portfolio, organizations often manage sub-portfolios. For example, the totality of all Caltrans projects constitutes a portfolio. Each component is a sub-portfolio, and each Caltrans district has its own portfolio and sub-portfolios of projects.

Key features of a portfolio include multiple projects within each portfolio, projects selected with the intent to help the organization achieve its strategic objectives, and the addition of new projects generally occurring at regular intervals. Overall, the portfolio consists of all actions that the organization undertakes or intends to undertake to bring about desired changes (Wei et al., 2023).

1.4 Objectives of the Study

The primary objective of this research is to develop robust cost-estimating models for the SHOPP. This study aims to support Caltrans, the California Transportation Commission (CTC), and the Legislature in improving the accuracy and management of cost estimates for SHOPP projects. The specific objectives are as follows:

1. Identify Cost Norms: Establish norms related to the overall costs of SHOPP projects, providing a baseline for future cost estimates.

2. Model Comparison: Evaluate the effectiveness of parametric regression models and neural network models in estimating project costs, and determine the most accurate and reliable approach.

3. Tool Enhancement: Provide Caltrans with improved tools for estimating and managing costs at the portfolio level.

4. Support CTC and Legislature: Assist the CTC and the Legislature in reviewing and assessing the overall costs of the SHOPP, ensuring transparency and accountability in the use of public funds.

By achieving these objectives, the study aims to enhance the accuracy and reliability of cost estimates for SHOPP, supporting better financial management and decision-making processes for California's highway operations and ensuring the efficient use of resources in maintaining and improving the state's transportation infrastructure.

1.5 Scope of the Project

This research focuses on developing and testing cost-estimating models using historical data from Caltrans's project expenditures as primary outputs. The scope includes:

1. Analyzing annual data sets of all Caltrans State Highway project expenditures from 1983 to 2021.
2. Developing cost-estimating models using parametric regression and neural network techniques.
3. Evaluating the performance of these models based on criteria such as accuracy and reliability.

Additionally, the research will address the benefits of using 2-standard deviations versus the 90th and 10th percentiles, considering whether the 2-standard deviations apply given the increasing variation with size, and evaluating the value of considering smaller sub-portfolios.

1.6 Structure of the Report

The report is organized into several sections to provide a comprehensive overview of the research process and findings:

1. Introduction & Background: An overview of the SHOPP, the importance of cost estimation, objectives of the study, scope of the project, and structure of the report.
2. Literature Review: A detailed review of existing research related to project cost estimating methods, including regression and artificial intelligence approaches.
3. Data Analysis and Development: A description of the data sets used, data processing techniques, and the development of a workable data set for SHOPP projects.
4. Model Development: An in-depth explanation of the model development process, including the use of parametric regression and neural networks, and the criteria for evaluating model performance.
5. Results and Discussion: Presentation of the findings, comparison of model performance, and discussion of the implications for cost estimation in highway operations.
6. Conclusion and Recommendations: Summary of the research outcomes, conclusions drawn from the study, and recommendations for future research and practical applications.

This structure is intended to provide a clear and systematic presentation of the research, facilitating a better understanding of the methodologies and results, and supporting the goal of improving cost estimation practices for the SHOPP.

2. Literature Review

This chapter introduces the concept of project portfolio management and provides an overview of the portfolio management process and the role of portfolio management in an organization. It also discusses the benefits of portfolio management and the challenges associated with managing a project portfolio.

2.1 Overview of Project Portfolios and Portfolio Management

This section builds on the concept of a project portfolio introduced in section 1.3 of this report. PMI offers this definition: "Portfolio management is the centralized management of one or more portfolios to achieve strategic objectives. It is the application of portfolio management principles to align the portfolio and its components with the organizational strategy" (PMI, 2017). A portfolio, as noted in section 1.3, is defined as "a collection of projects, programs, subsidiary portfolios, and operations managed as a group to achieve strategic objectives" (PMI, 2017). Virtually every organization has strategic objectives and projects that it plans to undertake to achieve those objectives, even if it does not use the terms "project portfolio," "strategic objectives," and "projects."

Elbok and Berrado assert that the primary goal of project portfolio management (PPM) is to maximize business value while maintaining the strategic alignment of the organization (Elbok and Berrado, 2017). Project portfolios are dynamic in nature and require continuous assessment and adjustment to respond to internal and external changes (PMI 2017).

Some of the research literature describes processes in PPM. These include: screening, project selection, resource allocation, and performance monitoring, as described below.

1. Project Selection and Screening: This process involves analyzing doable projects based on strategic fit, financial value, risk, and resource availability. Techniques such as productivity index and scoring models are commonly used to help decision-making (Bhuiyan, 2011).

2. Resource Allocation: Effective allocation of resources is important to make sure that the project has the needed human, financial, and technical resources. Klingebiel and Rammer (2014) investigate whether allocation of resources to a wider range of projects improves the performance.

3. Performance Monitoring and Balancing: The projects must be monitored and informed adjustments need to be made. Tadeu de Oliveira Lacerda et al., (2011) suggest the use of a multicriteria decision-aiding methodology.

Hadjinicolaou and Dumrak indicate that PPM practices offer many benefits, including increased cost savings, maximized resource usage, investment in the right areas, and repeatable success

(Hadjinicolaou and Dumrak, 2017). Moreover, Prencipe and Tell note that PPM supports the sharing and maintaining of knowledge learned from completed projects and utilization in the future ones (Prencipe and Tell, 2001).

However, Martinsuo suggests the selection of the portfolio is often not well-planned and rational. Instead, it is more political and path-dependent (Martinsuo, 2013). Although changes are required to optimize the portfolio and for customer satisfaction, Christiansen and Varnes note that the selection is also influenced by political and emotional factors instead of rational choices (Christiansen and Varnes, 2008). Additionally, the portfolio involves risk, which is dynamic and complex in nature, and the organization needs to be prepared for known threats (PMI, 2017).

PPM involves the coordination and control of several projects for an organization. It aligns projects with the organization's proposed strategy, optimizes resource allocation, and prioritizes projects based on their fit and returns. While a rational decision-making process is standard, recent literature suggests that negotiation and bargaining among stakeholders and the context in which PPM is used play an important role in its effectiveness and success.

Farshchian et al. claim that the selection of projects for portfolios is a challenging process and has been an active area of research for the past 40 years (Farshchian et al., 2017). Others note that practical solutions for the selection of project portfolios have been proposed for many types of projects, including highways and bridges (Son et al., 2019; Son and Khwaja, 2022). While some researchers have used simple methodologies to select projects located within the same geographical zones (Avineri and Cohen, 2018), others have used more sophisticated optimization and artificial intelligence techniques to select projects within limited budget environments (Gabriel et al., 2006; Liu and Wang, 2011; Zhang et al., 2023; Shakhsi-Niaei et al., 2015). Techniques such as cost-to-benefit ratio (CBR) and net present value (NPV) have also been used to select project portfolios in multiple conflicting criteria environments (Frej et al., 2021; Sun et al., 2021). In addition to the formation of project portfolios based on physical locations, budget limitations, NPV, and CBR, some researchers have proposed various project portfolio models based on labor and equipment requirements (Beşikci et al., 2015, Taghaddos et al., 2012; Gajpal and Elazouni, 2015).

2.2 Overview of the Caltrans State Highway Portfolio

Caltrans is responsible for coordinating the transportation infrastructure in the state; for planning, developing, and maintaining the State Highway System; and for interregional transportation. Its mission is to "provide a safe and reliable transportation network that serves all people and respects the environment." Wei et al. indicate that the Caltrans budget for Fiscal Year 2021–2022 consisted of 160 line items, which could be categorized into State Highway projects, Local Assistance, State Highway maintenance and operations, Interregional Passenger Rail, and other miscellaneous projects. State Highway projects accounted for 45 of these line items and made up most of the Caltrans budget (Wei et al., 2023; State of California, 2021).

Wei et al. describe how the Legislature is responsible for setting strategic objectives through legislation. As noted in its definition above, the purpose of portfolio management is to achieve strategic objectives. Wei et al. describe how Caltrans uses the term “program components” to identify each sub-portfolio of the Caltrans portfolio. They further indicate that these components are listed in Chapter 4 of the Project Development Procedures Manual (Caltrans, 2024b), and that each sub-portfolio/component has specific objectives and rules defined by the Legislature and the California Transportation Commission.

The SHOPP is a performance-driven project portfolio built on principles of asset management. The projects in the 2022 SHOPP were developed based on this framework established through the California Transportation Asset Management Plan and the State Highway System Management Plan. The SHOPP focuses on the improvement of four primary asset classes: pavements, bridges, drainage, and transportation management systems. Since the year 2000, the SHOPP has been the largest sub-portfolio within the overall California State Highway portfolio.

Wei et al. (2023) observed that:

1. There are many program components (sub-portfolios) in the Caltrans project portfolio.

2. Program components are added by the Legislature and come to an end when the Legislature removes the enabling statutes.

3. Each program component results from actions taken by the Legislature.

4. Each program component has its own goals and rules, established by the Legislature, often with additional details added by the CTC.

2.3 Conceptual Cost Estimation Models

According to Blampied (2018), a conceptual cost estimate is “[a] cost estimate that is made when a problem or opportunity has been named, before any project work has been done, before any project charging codes have been established, before any possible solutions or alternatives have been identified, and taking at most a few minutes, in order to decide whether to start a feasibility study.”

Conceptual cost estimation is an important step in the early phase of planning and developing the project. Several methods have been developed and improved over time for better accuracy and reliability. This section discusses and contrasts four documented approaches to conceptual cost estimation, namely analogous, parametric, bottom-up, and neural networks.

2.3.1 Analogous

According to PMI, analogous estimating is "a technique for estimating the duration or cost of an activity or a project using historical data from a similar activity or project" (PMI, 2019). PMI further mentions that "[a]nalogous techniques, also known as top-down estimating, are used when less information is available, the new project is very similar to a previous project, or the estimators are very experienced with what is going to be estimated. These techniques are preferred for early estimates when detailed information is not available" (PMI, 2017). Analogous estimation is very common in portfolios in which a project placeholder may be needed to evaluate the entire portfolio."

Analogous estimation is often used in the early stages of the project development when detailed information is limited. It requires an expert with deep knowledge of similar-sized projects and depends on their judgment. The American Association of State Highway and Transportation Officials (AASHTO) refers to this method as "judgment" in their Practical Guide to Cost Estimating (AASHTO, 2013).

While analogous estimation can provide the initial estimates, its accuracy is only dependable if the referenced data consists of projects of similar size and the expertise of the estimator.

2.3.2 Parametric

PMI defines parametric estimations as "an estimating technique in which an algorithm is used to calculate cost or duration based on historical data and project parameters" (PMI, 2019). PMI further notes that "[p]arametric estimating techniques are designed to provide some mathematical equations to perform estimating. Parametric estimating is based upon historical information of very similar projects but takes into consideration scale differences by identifying unit/cost duration from past projects and scaling the information to the required number of units in the current project" (PMI, 2019). AASHTO refers to this method as "stochastic" estimating.

Parametric models are useful in the early stages of project planning when detailed information is not yet available. They can provide more consistent estimates than analogous methods. This approach is referred to as "multiple regression modeling" in the Transportation Research Board's (TRB) Guidebook on Estimating Highway Preconstruction Services Costs (Gransberg et al., 2016).

Caltrans developed a notable example of parametric estimating for transportation projects known as Person-Year Project Scheduling and Cost Analysis (PYPSCAN). PYPSCAN used several input parameters such as project type, function, capital cost, environmental type, location, right-of-way information, and weather zone to estimate resource needs for different project phases.

Multiple linear regression (MLR) parametric modeling is a widely used method for cost estimation (Hollar and Rasdorf, 2013; Araya et al., 2020). MLR with 8 project attributes and multi-level Dirichlet process linear regression (MDPLR) with 13 attributes were compared by Hollar and Rasdorf (2013). Jeong and Woldesenbet (2012) developed a regression model for preliminary engineering cost estimation consisting of several plan sheets and engineering hours as the model variables. A multiple regression model to estimate the engineering hours for bridge replacement projects consisting of project attributes was developed (Araya et al., 2020). Blampied et al. (2023) developed three different multiple regression models to estimate preconstruction hours using a set of 138 Caltrans pavement projects and 21 input variables. These were the additive exponential, multiplicative exponential, and linear regression models described later in this report.

2.3.3 Bottom Up

According to PMI, bottom-up estimation is defined as "a method of estimating project duration or cost by aggregating the estimates of the lower-level components of the work breakdown structure" (PMI, 2019). PMI adds: "Bottom-up techniques, also called deterministic or detailed estimating, are applied as the estimating tool of choice for estimating costs and resource requirements when detailed project data become available" (PMI, 2019).

This approach is usually used later in the project lifecycle when more detailed information is available. While bottom-up estimating can provide highly detailed and potentially more accurate estimates, it is generally not feasible for conceptual cost estimation due to the lack of detailed project information at early stages.

Although PMI lists bottom-up estimation as a third approach, separate from analogous and parametric, the most common technique for estimating each of the detailed elements is an analogous estimate. According to analogous estimation, after breaking down the work into detailed elements, the estimator then uses expert judgment to estimate the cost or duration of the detailed element. In some cases, the estimator may sub-contract elements of the project, in which case the sub-contractor nevertheless develops an analogous estimate for their elements of the work. Some firms, particularly large firms, maintain databases of the detailed records of prior contracts and use their data to develop parametric estimates of detailed elements. It is more common, however, to use the databases to find analogous work and then to develop analogous estimates. With current advances in artificial intelligence, this can lead to the recent developments described in the following section.

2.3.4 Recent Developments in Conceptual Cost Estimation Methods: Artificial Intelligence

Recent advancements in artificial intelligence have led to the development of novel techniques for estimating costs at the conceptual stage, and multiple studies have been conducted for developing neural network-based cost estimation processes (Xue et al, 2020; Tijanić et al, 2020; Matel et al, 2022).

Benefits of Artificial Neural Networks (ANNs) over other conceptual cost estimation techniques include:

1. They can handle non-linear relationships between input and output variables.

2. Their accuracy improves over time as more data becomes available.

3. They can capture complex interactions between project features that might have been missed by parametric models.

Even though artificial intelligence (AI)-based methods like ANNs are promising, they also present challenges. These include the requirement of large high-quality datasets for training and the difficulty in explaining the black-box nature of ANN.

AI aims to mimic human thought processes, and the AI methods for cost estimating are essentially powerful methods of analogous estimating. In conventional analogous estimating, estimators rely upon their experience to develop estimates by expert judgment. As noted for bottom-up estimating above, estimators may supplement their expert judgment with a search through databases of past projects, enabling them to broaden the experience upon which they can draw to find an analogous project, work element, or situation. AI expands upon that ability further by providing the power of a computer to search for analogies. As with human reasoning, this search is most often performed through processes of elimination, although some versions of AI, such as Case-Based Reasoning, may use parametric tools rather than elimination to find the best analogies.

2.3.5 Summary

Although the discussion above has addressed four types of estimating, there are fundamentally only two, namely analogous and parametric. Bottom-up and AI estimation are variations that allow for a potentially more accurate version of analogous estimating, in some cases adding the use of parametric tools.

2.4 Applications to Highway Projects

Conceptual cost estimation is required in the planning and management of highway projects to provide early estimates of the budget required for project completion. It is also important for budgeting, allocating resources, and evaluating project viability. This section provides an overview of the applications of conceptual cost estimation methods in highway projects, drawing from studies conducted in the United States and elsewhere.

2.4.1 Applications to Transportation Departments in the United States

The National Cooperative Highway Research Program (NCHRP) has published several reports on cost estimation and management. These include NCHRP Report 574, which offers guidance on cost estimation during the planning, programming, and pre-construction phases (NCHRP, 2007). Similarly, NCHRP Report 625 provides procedures for right-of-way cost estimation and management, highlighting best practices and procedures (Anderson et al., 2007; Anderson et al., 2009).

Additionally, state-specific studies provide insights into cost estimation practices. For instance, the Washington State Department of Transportation has developed a Cost Estimating Manual for Projects that outlines the methods and tools that can be used for cost estimation (Washington State DOT, 2023). This manual suggests the use of parametric, historic bid-based, cost-based, and risk-based methods for cost estimation at different stages of project development. The Texas Department of Transportation (TxDOT) has also published a Construction Cost Estimating Guide for roadway and bridge projects, suggesting the use of three-point estimating for base cost estimation, allowances for not yet quantifiable items at a given project phase, and contingencies for unknown risks. Caltrans has published cost-estimating guidelines, suggesting the use of previous-bid items for similar projects, and complete analysis methods for the overall project cost estimation factoring in all costs.

2.4.2 International Practices

Various international studies and reports provide insights into the applications of conceptual cost estimation methods for highway projects. In Europe, the European Cooperation in Science and Technology's (COST) Action TU1003 has helped to standardize cost estimation processes across various countries by providing a unified approach to estimating road construction costs (Action TU1003, 2013). Also, the European Commission's report on the comparison of road infrastructure costs provides an analysis of the cost estimation methodologies and their application across the member countries (Doll and van Essen, 2008).

In Australia, the Australian Transport Assessment and Planning (ATAP) has published research on cost estimation methods for road projects. Their suggestions are similar to the guide published by TxDOT, which divides the cost estimation process into three stages: base cost estimation, contingency allowances, and escalation allowances.

Global perspectives on cost estimation are supported by research from international organizations such as the World Bank. The World Bank's guidelines for highway project cost estimation provide a framework for implementing well-made cost estimation practices for developing countries (Watanatada., et al., 1987).

2.5 Cost Estimating for Project Portfolios

Son and Khwaja consider the ratio of preliminary engineering to construction capital costs on a set of 628 Texas DOT bridge projects. This appears to be the most similar example in the literature to the present study. They conclude that their proposed methodology based on the ratio of preliminary engineering cost to construction cost for bridge projects displayed poor prediction accuracy for individual projects, with Mean Absolute Percentage Error ranging from 42% to 5,276%, but provided a much better accuracy at the portfolio level (Son and Khwaja, 2022). The objective and the accuracy of project cost estimates differ based on the project phase. The cost estimate can be divided into scoping cost estimates and detailed cost estimates. The former is performed during the initial phase while the latter is done when detailed information is available (Hessami et al., 2017).

The effective management of project portfolios is often limited by the uncertain cost estimation methods which are not consistent and objective throughout the portfolio (van Niekerk and Bekker, 2014). Van Niekerk and Bekker developed an ANN that estimates contingency percentages for cost and duration due to systemic risks, reducing subjectivity and enabling more consistent risk management across a project portfolio.

Although cost estimation is a technical process, it sometimes takes into account non-technical factors that form the basis of the costs (Akintoye, 2000). Carr (1989) mentions that only limited importance is given to establishing a fundamental base, which could assist in accurate estimates and be helpful in decision-making.

3. Data Collection, Analysis, and Set Development

3.1 Data Assembly

In our earlier research (Blampied et al., 2023), we obtained four sets of data from Caltrans. These four sets were:

1. Annual data sets of all Caltrans State Highway project expenditures from 1983 to 2021.
2. Detailed Caltrans project bid item data on more than 1,000 projects for which bids were opened between 2016 and 2021.
3. Data on the primary outputs from each Caltrans project in the SHOPP.

That research developed models specifically for pre-construction costs on individual pavement projects within the SHOPP. It used the bid items (data set 2) and primary performance measures (data set 3) as predictors of pre-construction costs.

Because the prior research focused on only a small subset of projects, the full range of data in the three data sets was not used. As a first order of business on this new project, the three data sets were prepared for use by this and future groups of researchers. This involved reformatting the data sets into more usable formats, an effort that occupied the first half of the current project.

Data source 1, the annual data sets of all Caltrans State Highway project expenditures is by far the largest of the three data sets, and required the greatest effort to be developed into a useful format. It consisted of the eighty-nine files listed in Appendix A. The files brought together data from almost forty years, and were in several different formats, determined by the formats of the original sources and the technological limitations that prevailed when those sources were developed.

Together, the eighty-nine source files had 2,396,747 individual records. Each record recorded an expenditure in dollars, and many records also recorded an expenditure of state employee hours that drove the dollar expenditure.

Each file included several fields that categorized the expenditures by a variety of factors. Between them, the source files used the sixteen fields listed in Table 1 to categorize expenditures. Thus, it is possible to obtain reports that summarize expenditures of both dollars and state employee hours by these categories.

Table 1. Fields Used to Categorize the Expenditures

Field name in final data set	Description
TransactionYear	The actual year in which a cost was incurred. Years are measured from July 1 to June 30, and listed as the later year (e.g., 2009 = July 1, 2008 to June 30, 2009).
FundingFiscalYear	Pre-2009 Funding Fiscal Year = the annual budget that paid for the cost (years are measured from July 1 to June 30). Not provided in the post 2009 data.
SourceDistrict	The district that performed the work (and incurred the expenditure).
ChargeDistrict	The district that owns the project.
ProjectCode	New code starting July 1, 2009. The first two characters are the old Charge District. The remaining number is a sequential serial number.
ExpenditureAuthorization	5-or-6-character alphanumeric. EAs beginning 0 to 4 are projects. In the pre-2009 data, the last character of project EAs is the phase. The Charge District + first 5 characters of the EA ("DistEA5") are the project identifier. (Each district has its own series of EAs; therefore, there can be 12 projects with the same EA, one for each of the 12 districts.)
CategoryOfExpenditure	Category of Expenditure, a 1970s code that grouped projects.
Phase	The project phase, introduced as a separate field since July 1, 2009. Prior to 2009, the Phase was the final character of the EA if the EA began with numerals from 0 to 4. See the Description of Expenditure Authorization.
Program	2-Character Program.
Element	2-Character Element, a subdivision of the Program.
Component	3-Character Component, a subdivision of the Element.
Task	3-Character Task, a subdivision of the Component.
PECLabel	Program, Element and Component, with their title.
PECTLabel	Task, with its title.
Type	A grouping of the objects, developed to avoid having excessive data due to the limitations of the old systems.
ObjectLabel	The type of expenditure, post-2009.

The field names listed in Table 1 are the final names used when the eighty-nine source files were combined into a single new file. Not every categorizing field occurred in every source file, and the categorizing fields had a variety of slightly modified names from one source file to another. The first few of the field names and formats are illustrated in Appendix B showing the modifications to these fields that had to be made to create a new combined file.

As an example of a required change, many of the early source files listed years as 2-character numbers. The year 1999, for instance, was given as 99. This creates a problem when one reaches the year 2000, which would be recorded as year 00 and appear to be 99 years before 1999. This is well known, and well documented, as the "Y2K" problem. The 2-character years were therefore converted to 4-character years by adding 19 before those years numbered 50 to 99, making them 1950 to 1999, and adding 20 to other 2-character numbers, which would change the year 00 into 2000 and 50 to 2050. The earliest year of data in the files was for Fiscal Year 1977, and the last year was for Transaction Year 2022, making the year 50 a reasonable cut-off.

Using a Python script and rules such as those displayed in Appendix B, the data from the eighty-nine files listed in Appendix A were combined to form a single comma-delimited text file which was named ExpenditureHistory.csv. This, as previously noted, included 2,396,747 records.

For possible ease of use by those who could not manage such a large file, the data from ExpenditureHistory.csv was uploaded to a Microsoft Access file, named ExpenditureHistory.accdb.

3.2 Quality Control

3.2.1 Purpose

A series of checks were performed to verify that the data from the eighty-nine source files listed in Appendix A had been properly combined into ExpenditureHistory.csv, applying the modifications described in Appendix B. These checks served as a Quality Control measure to ensure the accuracy and completeness of the ExpenditureHistory.csv. They were intended to address both errors in the Python code and errors in human procedures. Quality Control is defined as "the process of monitoring and recording results of executing the quality management activities in order to assess performance and ensure the project outputs are complete, correct, and meet customer expectations" (PMI, 2017).

3.2.2 Method Used

The Quality Control test compared sums from the original expenditure files with the corresponding sums in the combined expenditure history file. The purpose of this test was to ensure that the original expenditure files and combined expenditure file produced the same results. Using Pivot Tables on each original source file, and a new Python Code on the combined expenditure history file, a count was made of the records, and sums were found of the Expenditure Amount and Labor Hours for each Transaction Year, Funding Fiscal Year, Source District, and Charge District.

The Pivot Tables were created for each of the eighty-nine original source files, and Python Reports were run for the entire combined expenditure file.

The ability to create pivot tables is standard in Excel and well-documented on Microsoft.com. Pivot tables are limited, however, by Excel's limitation of 1,048,576 rows of data. This limitation was not a problem in the analysis of the original expenditure files since none of them exceeded this limit. The combined expenditure history file, however, was considerably larger than Excel's limit.

3.2.3 Results

As an example of the results, the new Python Code found that the combined expenditure history file included 182 records with the earliest Funding Fiscal Year (FFY) in the records being 1977. These records together recorded zero State Employee hours and -$413,982 (a negative number) in expenditures. Appendix C lists the number of records for FFY 1977 found in each of the original source files, totaling 182 records. In this instance, then, the sum of the records in the source files matched the corresponding sum in the combined expenditure history file.

As an explanation of Appendix C, the FFY records the year in which expenditures were budgeted. The money budgeted in any FFY can be expended over several years in accordance with rules established by the Legislature and Congress. This is necessary to accommodate the fact that projects and contracts take several years to complete. If a contract is awarded in a given calendar year (the Transaction Year (TYR)), the money for that contract is normally encumbered or obligated and remains available for several years to fund the contract. (The State of California refers to this as "encumbered" or "an encumbrance." The United States uses the term "obligated" or "an obligation.") Although the records begin on July 1, 1983, the start of Transaction Year 1984, then, funds from the earlier budget year were still being expended.

3.3 Quality Assurance

3.3.1 Purpose

In the Quality Control, the single expenditure history file was subjected to a series of tests to ensure that this huge file had been constructed correctly and that the data import from the files listed in Appendix A had been performed correctly.

A subsequent, less extensive check was performed on the single expenditure history file to provide Quality Assurance, defined as "the process of auditing the quality requirements and the results from quality control measurements to ensure that appropriate quality standards and operational definitions are used" (PMI, 2013). This second check therefore consisted of sample audits of some of the measurements in the Quality Control.

3.3.2 Method Used

From the Required Fields column in Appendix B, it can be seen that the following fields are required on every record in the combined ExpenditureHistory.csv:

- TransactionYear

- SourceDistrict
- Program
- Element
- Component
- Task
- ExpeditureAmount
- LaborHours
- DistEA5
- Source

Source, the last field, is a string that identifies the source from which the data was obtained, and is not found in the original expenditure files.

For Quality Assurance, a test was performed on both a selection of the original expenditure files and the corresponding records in the single expenditure history file. The purpose of this test was to ensure that the original expenditure files and single expenditure file produced the same results. Using pivot tables and Power Pivot, a count was made of the records, and sums were then also found of the Expediture Amount and Labor Hours for each combination of the first six required fields, namely Transaction Year, Source District, Program, Element, Component, and Task.

As noted above, the ability to create pivot tables is standard in Excel and well-documented on Microsoft.com. Pivot tables are limited, however, by Excel's limitation of 1,048,576 rows of data. This limitation was not a problem in the analysis of the original expenditure files since none of them exceeded this limit. The single expenditure history file, however, was considerably larger than Excel's limit. For this file, then, Power Pivot was used. Power Pivot is an add-in to Excel and functions in the same manner as pivot tables, but its limitation of 1,999,999,997 rows is more than 1,900 times the Excel limit.

The following nine original expenditure files were selected for this analysis:

- ExpenditureHistory1983-84D1-4.txt
- ExpenditureHistory1987-88D5-8.txt
- ExpenditureHistory1991-92D9-58.txt

- ExpenditureHistory1995-96D59+.txt
- ExpenditureHistory2000-01.txt
- ExpenditureHistory2005-06.xlsx
- Caltrans ROE FY 2011.xls
- Caltrans ROE FY 2016.xls
- Caltrans ROE FY 2021.xls

These files were selected to provide a sample spread through the years in which the data was collected; data from every Caltrans district; and original expenditure files in every format in which they were obtained.

The Quality Assurance, then, produced ten pivot tables. The first nine tables were pivots of the above nine original expenditure files, using standard Excel pivot tables. The tenth was a pivot of the entire 2,396,747-record combined expenditure file using Power Pivot. Sections of the combined-expenditure-file pivot were then copied and pasted alongside the data from the nine original-file pivots, and compared with those original-file pivots.

3.3.3 Results

As an illustration of the results, Appendix D shows the first few rows of the report that was generated for the ExpenditureHistory1983-84D1-4.txt file with the comparable data from the combined ExpenditureHistory.csv. Data from the pivot table of the original data appears on the left, while data from the Power Pivot of the single expenditure history file appears to the right. Similar reports to Appendix D were generated for each of the nine original expenditure files. In every report, the count of the records is the same for each row of the original data as for the single expenditure history file. Similarly, the sum of the expenditures is the same for each row of the original data as for the single expenditure history file, and the sum of the labor hours is the same for each row of the original data as for the single expenditure history file. Although the data in Appendix D is limited to the first few rows of each report, the match does continue throughout the nine reports.

Appendix E summarizes the results. For each of the nine original expenditure files, the original expenditure files had exactly the same number of rows as the combined expenditure file, and the total expenditures on amounts that sometimes totaled more than $5 billion nevertheless matched to the cent. The labor hour total showed discrepancies of up to 0.06 hours (3 minutes and 36 seconds) in summing amounts of over 10 million hours. As timesheets are recorded to an accuracy of no more than 0.1 hours (6 minutes), the discrepancy appears to result from deficiencies in the

Excel software rather than from any error in copying the original data into the combined expenditure file.

3.4 Conclusions

Based on the Quality Control findings and the Quality Assurance tests illustrated in Appendix D, it seems that the combined expenditure history file can be used in future reports with confidence that it correctly reflects the original expenditure files.

4. Model Development

4.1 Overview of Model Development Process

The model development process for estimating costs in the SHOPP portfolio involved several steps to ensure the creation of robust and reliable models. This section outlines the phases of the process, including the utilization of various subject matter experts, as well as ongoing discussions with Caltrans (Districts and Headquarters).

4.1.1 Utilization of Subject Matter Experts

Engaging industry Subject Matter Experts (SMEs) was an aspect of the model development process. These SMEs provided invaluable insights and guidance throughout the project.

The following steps highlight the involvement of these key stakeholders:

1. Subject Matter Expert Consultations: Regular consultations were held with experienced transportation professionals, who both provided practical insights into the complexities of highway project cost estimation and highlighted common challenges faced in the industry.

2. Review and Feedback: Draft models and preliminary findings were shared with SMEs for review and feedback. Their critiques and suggestions were instrumental in refining the models and ensuring their practical applicability.

4.1.2 Consideration of Ongoing Efforts in Caltrans Districts

Discussions with Caltrans districts and headquarters played a crucial role in tailoring the cost estimation models to the specific needs and conditions of the SHOPP projects.

The following points summarize the key aspects of these interactions:

1. Needs Assessment: Initial discussions with representatives from Caltrans helped identify the unique requirements and challenges faced. This needs assessment ensured that the models addressed the variations in project costs.

2. Data Collection and Validation: data on past project expenditures and outcomes were collected and validated through direct communication with Caltrans personnel. These data sets provided a solid foundation for model development and testing.

3. Feedback on Model Prototypes: Preliminary versions of the cost estimation models were shared with Caltrans representatives for feedback. Their practical insights and suggestions were crucial in refining the models to enhance their accuracy and usability.

4. Continuous Engagement: Regular updates and ongoing communication with Caltrans ensured that the models remained aligned with evolving needs and priorities. This continuous engagement facilitated the incorporation of real-time data and emerging trends into the models.

In summary, the model development process for estimating costs in the SHOPP portfolio was a comprehensive and collaborative effort, leveraging the expertise of industry advisors, subject matter experts, and Caltrans district representatives. This approach ensured the creation of robust, reliable, and practical cost estimation models tailored to the specific needs of California's highway operations.

4.1.3 Description of the Data Set Development Process

The initial datasets for the project consisted of the following:

1. Annual datasets of all Caltrans State Highway project expenditures from 1983 to 2021.

2. Detailed Caltrans project bid item data on more than 1,000 projects for which bids were opened between 2016 and 2021.

3. Data on the primary outputs from each Caltrans project in the SHOPP.

The annual dataset consisted of 89 files in different file formats. The relevant fields for each file were extracted to create a single expenditure history dataset consisting of the fields listed in Table 1. This combined expenditure file has 2,396,747 records. A quality assurance check was performed to assess the quality of the extracted data.

Several datasets were developed for the expenditure history dataset by aggregating the costs of different phases across the duration of the project. The projects that were active only between the years 2000 and 2022 and had both the construction and support costs were selected for the model development. The developed datasets consisted of pavement, major damage, safety, and bridge.

4.2 Parametric Models

4.2.1 Three Possible Functions

As discussed in the Literature Review, there are fundamentally only two methods of estimating: (1) analogous, also referred to as expert opinion, and (2) parametric, also referred to as statistical or stochastic. All other named methods are variations of these two.

Throughout the estimating literature, one finds the concepts referred to by more than one name. Thus, we have above that analogous is also called expert opinion, and parametric is also called statistical or stochastic. This pattern continues below.

According to Membah (2016), there are "in general" three possible forms of the parametric, or statistical, function. These are:

1. The additive or "power curve" exponential regression function:

$$\text{Cost} = a + \sum_{i=1}^{i=n} b_i X_i{}^{ci} \qquad \text{(Equation 1)}$$

Wallace (1978) and Akeel (1989) use the word "additive" to describe this function, while Hamaker (1987) refers to it as a "power curve."

As an example, Blampied (2018) developed the following additive exponential regression model for predicting the cost of pedestrian access projects based upon a data set of 39 projects:

$$\text{Cost} = 173{,}250.36 + 135{,}312.73X_1^{0.4019} + 30{,}649.59X_2^{0.4089} + 158.33X_4^{0.6814} + 22{,}595.99X_5^{0.47}$$
(Equation 2)

Where cost is a value in dollars, inflation-adjusted to a January 1, 2012 base date; X1 is the required number of wheelchair ramps; X2 is the length in linear feet of sidewalk to be constructed; X3 is the number of audible traffic signals to be constructed; X4 is the dollar amount paid by the agency to property owners and utility companies for right-of-way (land, easements, and utility relocations); and X5 is the number of hours that employees spent on obtaining right of way.

2. The linear regression function, a special case of the additive exponential with all the "C" exponents set at a value of "1":

$$\text{Cost} = a + \sum_{i=1}^{i=n} b_i X_i \qquad \text{(Equation 3)}$$

Blampied (2018) finds that this is by far the most common form of parametric estimating function, with examples in the literature spanning a time range from 1984 (Kouskoulas) to 2017 (Wang et al.) and 2018 (Elmousalami et al.).

As an example, Blampied (2018) developed the following linear regression model from the same sample set as for equation 2:

$$\text{Cost} = 489{,}337.85 + 4{,}646.14X_1 + 374.19X_2 + 0.X_3 + 7.04X_4 + 226.25X_5 \text{ (Equation 4)}$$

3. The multiplicative, multiplicational, or logarithmic exponential regression function:

$$\text{Cost} = a + b.\prod_{i=1}^{i=n} Xi^{ci} \qquad \text{(Equation 5)}$$

Here, again, the function goes by several names in the literature. Wallace (1978) and Akeel (1989) say "multiplicational," Kouskoulas (1984) prefers "multiplicative", while Hamaker (1987) uses "logarithmic curve." Other writers prefer "modified Cobb-Douglas exponential parametric function" (Forster et al., 1987; Akeel, 1989; Irfan et al., 2012; Lei et al., 2015; Elmousalami et al., 2018).

Using the same sample set as for equations 2 and 4, Blampied (2018) developed the following multiplicative exponential regression model:

Cost = 288,981.67 + 50,956.8X10.3844.X20.1306.X30.5061.X40.00061.X50.2085 (Equation 6)

Whereas factors are added in equation 2, here they are multiplied. X1 through X5 are as in equations 2 and 4, except that they are each increased by one unit to address the problem that a zero for any factor would produce a zero result for the cost in the multiplicative exponential function.

The "logarithmic" form of this function takes the logarithm of each side of the expression to produce:

$$\mathrm{Log(Cost)} = \mathrm{b} + \sum_{i=1}^{i=n} c_i . \log(X_i) \qquad \text{(Equation 7)}$$

4.2.2 Application in the Present Research

In the present research, we began by finding the best-fit versions of each of the above expressions, but then we found that for the very large sets that we were using there is no significant difference between them. Thereafter, we proceeded to calculate only the simplest linear regression form.

We used the engineering support-to-capital ratio, which was the ratio most available to us, although it has some serious flaws. This ratio has been used by Caltrans for many decades and continues to be referenced, but is criticized, especially among experts in project management, because it is a gross single-factor estimate that does not allow for the peculiarities and uniquenesses that characterize each project. This contrasts with Blampied et al. (2023), who used 21 input variables to estimate the preconstruction hours on Caltrans pavement projects.

The engineering support-to-capital ratio is easily available because it is rooted in U.S. Federal regulations that stem from the very first Federal Aid Highway Act, the Rural Post Roads Act of 1916. Following the 1916 Act, recipients of U.S. Federal funds for highway construction were required to report their costs in four categories: preliminary engineering, construction engineering, construction capital, and right-of-way. These categories remain in force and are required throughout the country, with a history of over 100 years.

The State of California further subdivides preliminary engineering into three phases: project initiation (Phase K), environmental studies and permits (Phase 0), and plans specifications and estimates (Phase 1). Construction engineering is recorded as Phase 3 and construction capital is Phase 4. Right of way, which is not considered in this study, is divided between a support element, Phase 2, and capital, Phase 9. The division between support and capital is a requirement of

California state law, and the terms “support” and “capital” have specific meanings in California. The definition of “support,” especially, is unique to California State Government. In essence, “support” consists of the costs required to employ California state employees—salaries, benefits, building rental, utilities, etc. In Caltrans, unlike other California State Agencies, consultants are also counted as support, on the principle that using consultants is an alternative to hiring state employees.

In the past, Caltrans also used a Phase 5 to designate minor construction capital, with Phase 4 being reserved for major project construction capital. The use of Phase 5 was discontinued in 2010 when Caltrans adopted new accounting software, but the data from earlier years does include Phase 5.

Building on the original Federal requirement, phases have come into use in California State Law, using slightly different terms.

- In Government Code 14526.5 (c) (1), which refers to the SHOPP, the phases are listed as (A) project approval and environmental documents, support only [i.e., Phase 0], (B) plans, specifications, and estimates, support only [i.e., Phase 1], (C) rights-of-way [i.e., Phases 2 and 9], and (D) construction [i.e., Phases 4 and 5].

- In Government Code 14529 (b), which refers to the STIP, they are (1) completion of all permits and environmental studies [i.e., Phase 0], (2) preparation of plans, specifications, and estimates [i.e., Phase 1], (3) the acquisition of rights-of-way, including, but not limited to, support activities [i.e., Phases 2 and 9], and (4) construction and construction management and engineering, including surveys and inspection [i.e., Phases 3, 4, and 5].

- In Government Code 14556.13. (b) which refers to the TCRP, they are (1) studies, environmental review, and permits [i.e., Phase 0], (2) preparation of project plans and specifications [i.e., Phase 1], (3) right-of-way acquisition [i.e., Phases 2 and 9], and (4) construction or procurement [i.e., Phases 3, 4 and 5].

- In Article XXII of the California Constitution, which refers to architectural and engineering services, they are permitting and environmental studies [i.e., Phase 0], design phase services [i.e., Phase 1], rights-of-way services [i.e., Phase 2], and construction phase services [i.e., Phase 3].

Although project initiation, Phase K, falls within the federal category of preliminary engineering, it is not included in these state laws because, in the California state system, project initiation is considered to be part of pre-project planning and not a part of the work on the project. Our approach, then, is to consider this ratio:

- Engineering: Phases 0, 1, and 3, i.e., preliminary and construction engineering, as the numerator.

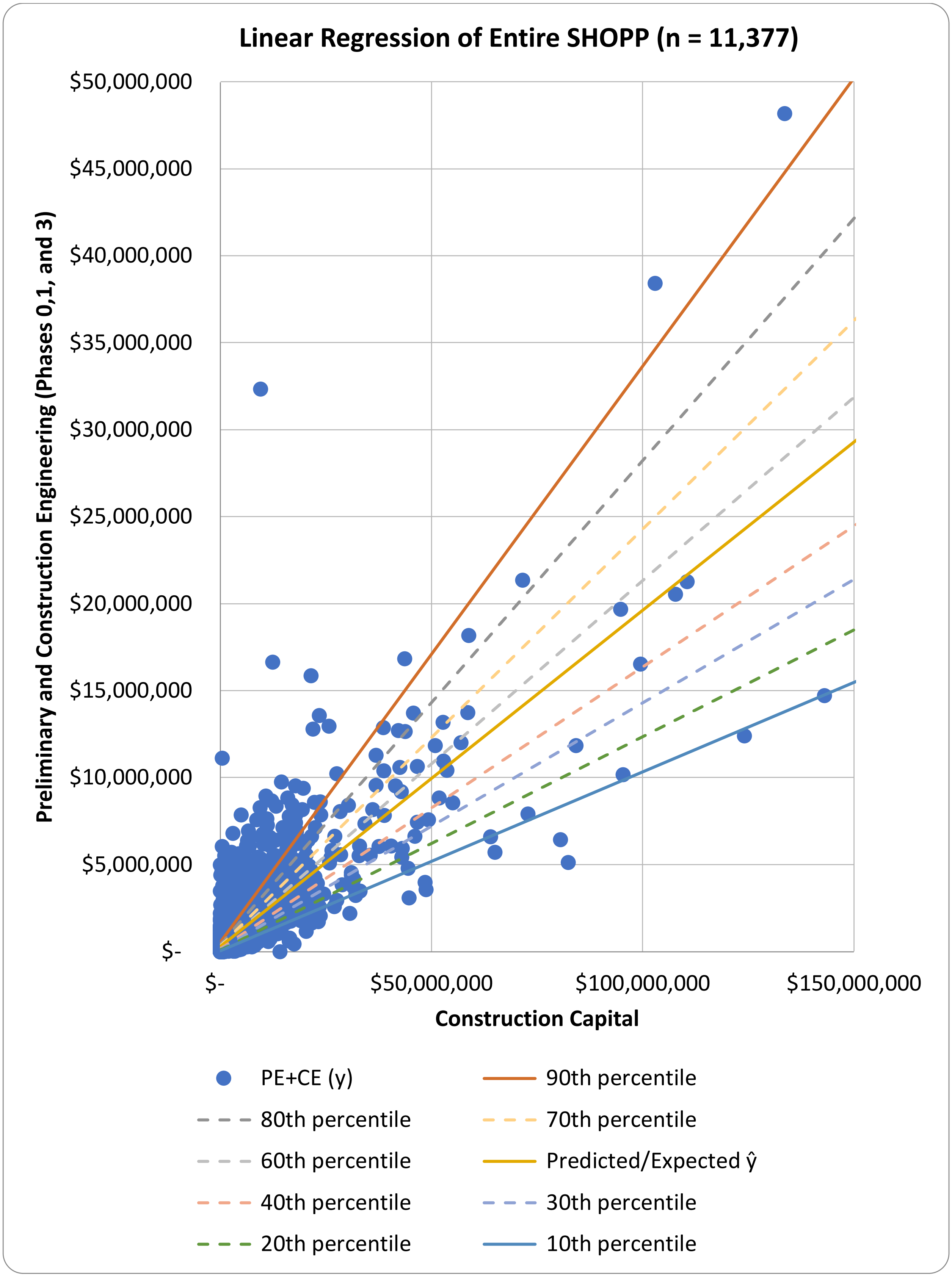
- Construction capital: Phase 4 and 5, i.e., construction capital, as the denominator.

An initial data set development found that 11,377 SHOPP projects were started on or after July 1, 1999, and completed on or before June 30, 2021. Linear regression was applied to a data set of these projects and moving deciles were found to produce Figure 1.

The best-fit linear regression, also called the predicted or expected value, in Figure 1 has a coefficient of determination of 0.669. This means that for a given construction capital cost, the best-fit line explains 66.9% of the variation in the required engineering cost. The line has a zero intercept of $320,927, meaning that there is a fixed engineering cost of that amount on the average SHOPP project. The line has a positive slope of 0.193, meaning that each dollar of construction capital incurs 19.3 cents of engineering cost in addition to the fixed engineering cost.

The decile lines are significant and are a deviation from earlier work. There has been a tendency to offer only single-point best-fit estimates, which are misleading. In practice, virtually all actual costs are above or below the best-fit line. It would be extremely rare for an actual cost to fall exactly on the best-fit line. Knowing this, there is a challenge to determine whether a particular estimate is unreasonably far above or below the best-fit line. The percentile lines provide an indication of that reasonableness.

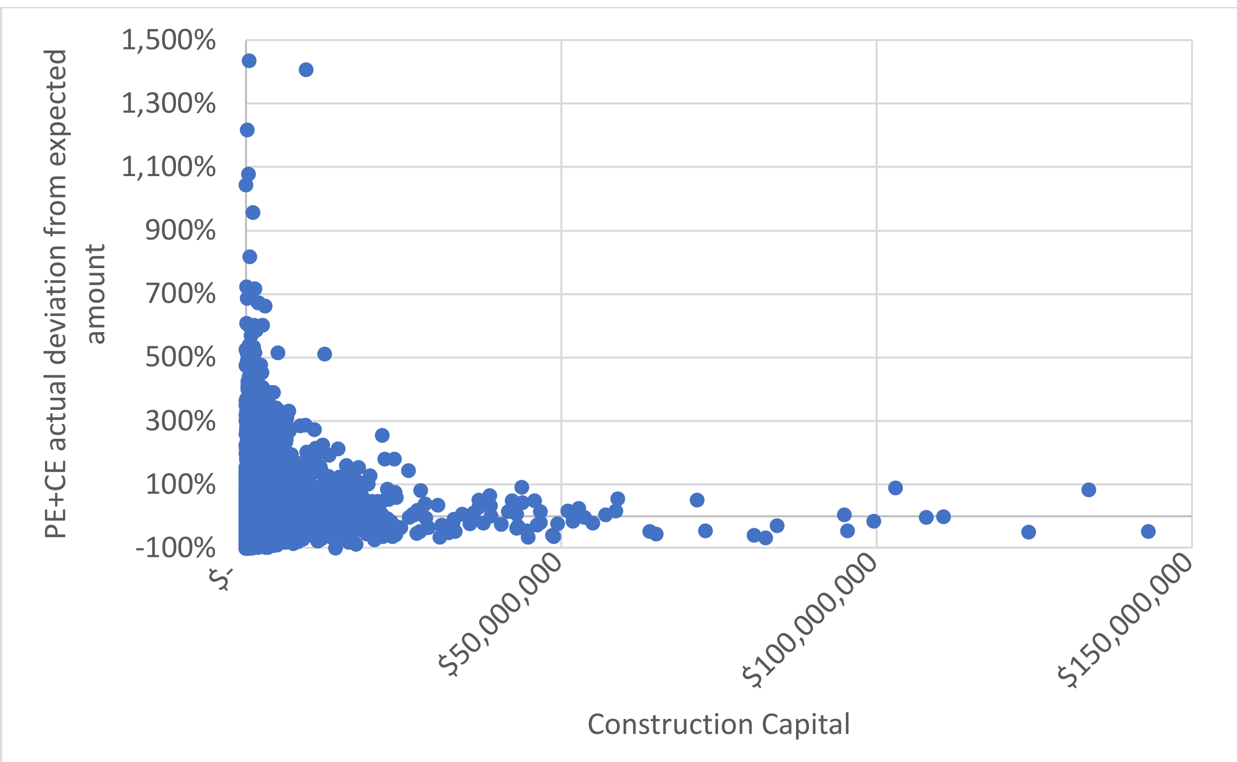
Figure 1. Linear Regression Lines for the Entire SHOPP



4.2.3 Standard Deviations

Later work in this report (Figures 5 to 12) uses standard deviations of the engineering cost rather than percentile lines. These standard deviations are fixed-dollar amounts above and below the expected values. When one considers the scatter of the actual engineering cost as a percentage of the expected engineering cost, one sees that the upper limits of the scatter are considerably further from the expected cost than are the lower limits. This is shown in Figure 2. The lowest possible engineering cost is zero, or close to zero. At the lower limit, one has to incur at least some engineering cost on a project unless one has a financial-contribution-only project (FCO), and we excluded FCOs from our data set. There is, however, no upper limit to costs. No matter how much one has spent, it is always possible to incur further expenses.

Figure 2. Scatter of the Actual Engineering Cost as a Percentage of the Expected Engineering Cost

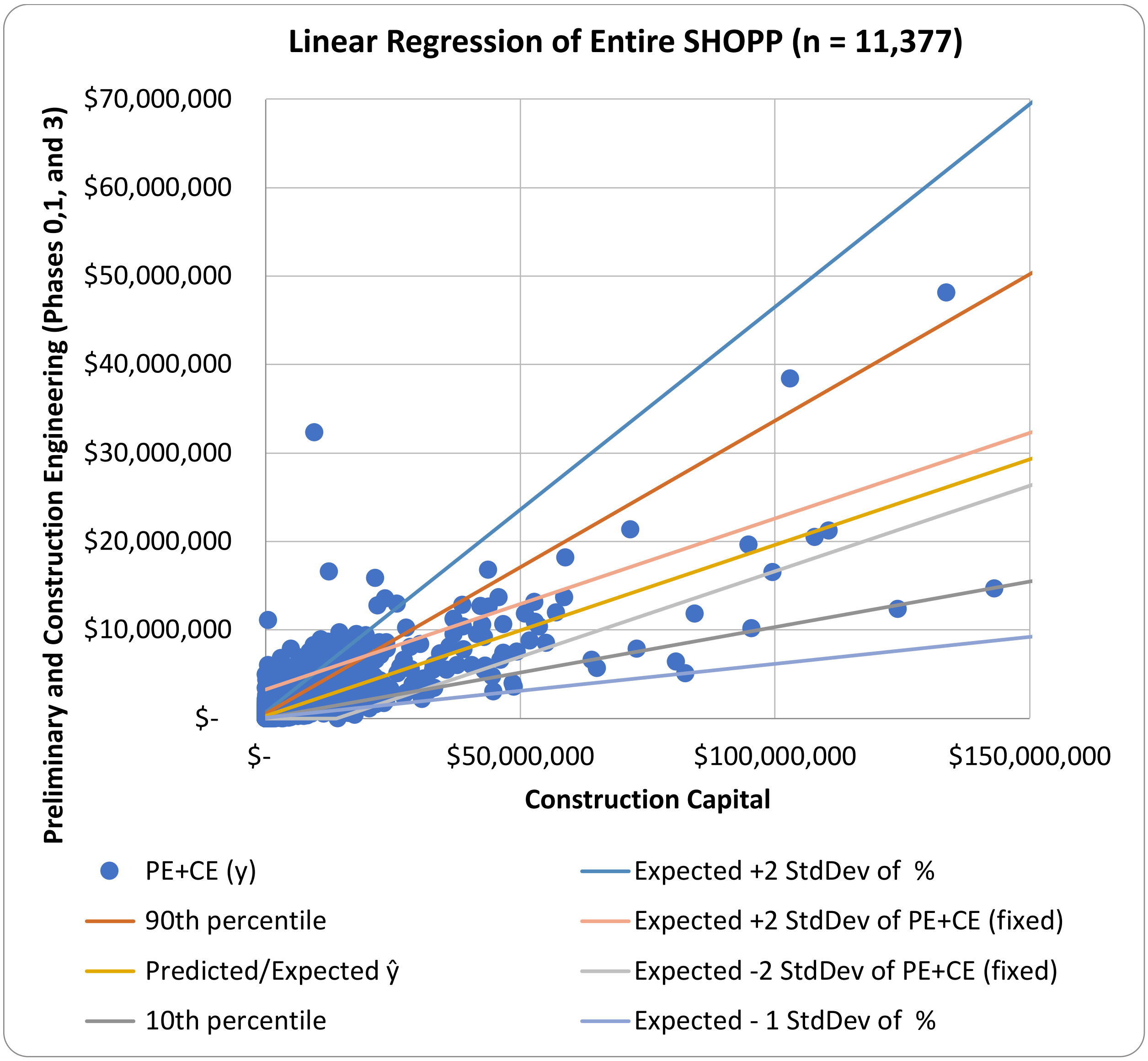


An engineering cost of zero would be 100% of the expected amount below the expected amount, resulting in a deviation of -100%, the lowest possible. A cost of, say, three times the expected amount would have a deviation of +200%. Figure 2 is truncated vertically at 1,500%. The actual highest recorded deviation on our data set is 2,685%, but the truncation avoids having a lot of fairly empty space at the top of the figure. High actual engineering costs of this proportion are not unusual in relatively small projects with complex and unusual engineering work.

Figure 2 illustrates a challenge created using fixed dollar standard deviations. The deviations of the data are not distributed about the mean according to a statistical normal function, but rather have a long-skewed tail toward the high costs combined with a hard low-cost truncation at the zero dollar limit.

Figure 3 illustrates both the challenge with fixed-dollar standard deviations and an alternative approach. This figure copies the 90th percentile, predicted/expected, and 10th percentile lines from Figure 1. It then adds 2-standard deviation fixed-dollar lines above and below the expected line, similar to those in Figure 5, which will be discussed later.

Figure 3. Alternative Approach to Addressing Standard Deviations



As an alternative to the fixed-dollar standard deviations, Figure 3 takes the standard deviations of the percentages from Figure 2 and adds them to the figure. This standard deviation is 68%. One cannot reasonably deduct 2 x 68% from the expected values because that would end with a negative lower limit, constantly become more negative as the capital cost increases. Instead, Figure 3 adds a 1-standard deviation line below the expected and a 2-standard deviation line above the expected. The lower 1-standard deviation line is lower than the 10th percentile line, while the upper 2-standard deviation line is higher than the 90th percentile.

In a normal distribution, a line at 1-standard deviation below expected would be at the 16th percentile, while a line at 2-standard deviations above expected would be at the 98th percentile. The 90th percentile line would correspond to 1.3 standard deviations above expected, while the 10th percentile would be at 1.3 standard deviations below expected. As suggested by Figure 2 and the discussion of limits, the data is skewed, with a long tail on the high side.

Further study of the upper and lower limit lines could provide a useful extension of this research. This is discussed in Section 6.

4.2.4 Sub-portfolios of the SHOPP

As noted in Section 1.4 of this report, an objective was to find subsets of the SHOPP portfolio that would provide Caltrans with useful tools for checking and validating SHOPP engineering cost estimates. Our initial tentative breakdown was to divide the SHOPP into six sub-portfolios, namely 1. Pavement, 2. Bridges, 3. Safety, 4. Major Damage, 5. Drainage, and 6. Roadside. These are strictly tentative breakdowns for the sole purpose of validating engineering cost estimates and under no circumstances should any suggestion be made that we are advocating that this should be a breakdown for the purposes of portfolio management. On the contrary, the Caltrans approach of adopting a single unified SHOPP portfolio for purposes of management, planning, programming, construction, and asset management appears to be well thought-out. As we have noted elsewhere in this report, it would be good to use some of the cost estimating and evaluation tools in our present study and previous study to test the effectiveness of this unified portfolio approach and possibly to suggest tweaks or improvements to the path that Caltrans has adopted.

As part of the present study, we developed datasets and models for the first four of the above sub-portfolios, namely 1. Pavement, 2. Bridges, 3. Safety, and 4. Major Damage. Appendix F provides figures similar to Figure 1 for each of these four sub-portfolios.

Table 2 summarizes the major parameters of Figure 1 and the four figures in Appendix F.

Table 2. Major Parameters of the Five SHOPP Linear Regression Charts

Portfolio or Sub-Portfolio	n	Regression R^2	a (fixed cost)	b (variable cost)
Entire SHOPP	11,377	0.669	$302,927	$0.193 / $ of capital
Pavement projects	1,640	0.731	$206,661	$0.171 / $ of capital
Bridge projects	542	0.638	$475,757	$0.364 / $ of capital
Safety projects	3,048	0.589	$291,488	$0.338 / $ of capital
Major damage projects	3,662	0.860	$161,833	$0.179 / $ of capital

The increased granularity from the entire SHOPP to major sub-portfolios improved the reliability of the result for pavement projects and major damage, but not for bridge and safety projects.

4.2.5 Sub-sub-portfolios of Pavement Projects

Having completed a linear regression of the above four sub-portfolios, a question arose about whether a more granular division of sub-portfolios might be more useful. This was asked particularly with regard to pavement projects, which include three sub-portfolios of increasing complexity, namely

- Sub-portfolio 20.201.120, also called 3R (Resurfacing, Restoration, and Rehabilitation);
- Sub-portfolio 20.201.121, also called 1R (Resurfacing), and widely known as Capital Preventive Maintenance (CAPM);
- Sub-portfolio 20.201.122, also called 2R (Resurfacing and Restoration).

In general, 1R requires the least engineering work, 2R more work, and 3R yet more.

The 1,640 pavement projects were divided into the three sub-portfolios and the analysis repeated. This produced the data in Table 3.

Table 3. Further Analysis for Pavement Projects

Portfolio or Sub-Portfolio	n	Regression R^2	a (fixed cost)	b (variable cost)
All of the below pavement projects	1,640	0.731	$206,661	$0.171 / $ of capital
20.201.121 pavement 1R, i.e. CAPM	797	0.658	$189.73	$0.153 / $ of capital
20.201.122 pavement 2R	30	0.783	$797,218	$0.132 / $ of capital
20.201.120 pavement 3R	813	0.716	$219,841	$0.174 / $ of capital

4.2.6 Effect of Future Inflation

Because the figures in this report are based upon dollar amounts, inflation is a concern. It would be reasonable to ask whether and when the figures would become obsolete due to inflation. It should be noted that the figures are based in large part upon proportions rather than absolute numbers. This is illustrated in Tables 2 and 3, both of which show fixed amounts, a, and variable, or proportional, amounts, b. The b factor will change with inflation only if the long-term inflation in construction capital cost is significantly different from the long-term inflation of engineering cost. Blampied (2018) found that, for Caltrans, the inflation of construction capital costs was 4% per annum from the second quarter of 2000 to the third quarter 2012, based on the Caltrans Highway Construction Cost Index (CHCCI). The engineering cost inflation over the same period was 4.7%. Based in this data, the slopes of the b factors in Tables 2 and 3 would need to be adjusted downward by 0.7% of the b amounts. That is, the 0.193 slope of the 'Entire SHOPP' line in Table 2 might need to become 0.193 x (1.04/1.047) = 0.192.

The fixed amounts, a, are directly affected by inflation and would need to be adjusted upwards by 4.7% compounded annually.

For both the a and the b numbers, it would be useful to analyze more recent inflation data than that from 2012.

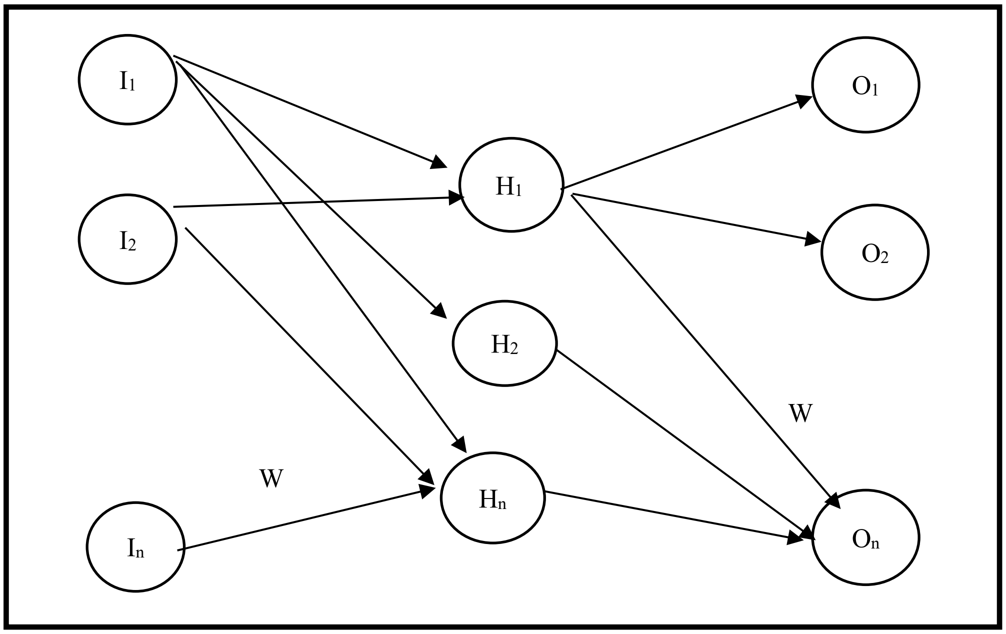
4.3 AI Models

4.3.1 Machine Learning Technique Applied

There are different machine learning techniques that have been used in many construction industry applications. Examples of these techniques are case-based reasoning, decision trees, and neural networks. Neural networks have been reported to be the most commonly used machine learning technique in cost estimation applications (Shehab et al., 2014; Adel et al., 2016; Barakchi et al., 2017; Wang et al., 2017; El-Kholy et al., 2020; Goodarzizad et al., 2021; Matel et al., 2022).

There are various paradigms of neural networks, each of which is considered applicable for certain applications. Backpropagation is considered to be the most commonly used paradigm in cost estimation applications. This paradigm consists of an input layer, output layer, and one or more hidden layers (Figure 4). As illustrated in Figure 4, each of these layers contains one or more neurons. The neurons in the input layer are connected to the neurons in the hidden layer(s) and the neurons in the hidden layer(s) are connected to the neurons in the output layer. These connections are associated with weighted factors (Ws), which represent the network's state of knowledge.

Figure 4. Structure of Backpropagation Neural Networks



4.3.2 Model Training and Evaluation

To train the backpropagation neural networks, the supervised training technique was used. In this technique, sample input data is fed into the input layer. The network processes this input data using parameters such as momentum, activation functions, and learning rates to calculate outputs. These calculated outputs are then compared to the actual ones and error factors are calculated. This process is repeated until the error factors are minimized. It should be noted that the input data used to develop these models was the project's total cost of Phases 0, 1, and 3 (Pre-construction & Engineering Costs), while the actual output was the project's total cost of Phases 4 and 5 (Construction Cost).

The above-described training process was used to develop eight Artificial Intelligence (AI) models. These AI models serve the same purpose as the statistical models presented in the previous chapter. To develop the AI models for the entire SHOPP projects—Pavement, Safety, Bridges, Major Damage, Pavement 1R, Pavement 2R, and Pavement 3R models—a total of 11,377; 1,640; 3,049; 542; 3,662; 797; 30; and 813 projects were used, respectively. The performance of the developed neural network models was evaluated using the Coefficient of Multiple Determination (R2). Table 4 presents the total number of projects used to develop these models and R2 values of each model.

To reflect on the variability of data used in developing these models, the standard deviation (σ) was calculated. For a normal distribution, 68% and 95% of data are within 1σ and 2σ range, respectively. One way to look at these standard deviations is if the developed models predict a cost "X" of a project, then the actual cost could be as high as $X+2\sigma$, as low as $X-2\sigma$, or anywhere in between. It should be noted that these standard deviations serve the same purpose of the 10th and 90th percentiles presented in Section 4.2, but with a slightly wider range. This is because 2σ is about the 98th percentile. Figures 5 to 12 depict the developed AI models along with their associated one and two standard deviation ranges (i.e., +/- σ and +/- 2σ).

Table 4. Performance of Developed AI Models

AI Model	Number of Projects	R^2
Entire SHOPP	11,377	0.71
Pavement	1,640	0.76
Safety	3,049	0.67
Bridge	542	0.68
Major damage	3662	0.88
Pavement 1R	797	0.71
Pavement 2R	30	0.91
Pavement 3R	813	0.77

Figure 5. AI Model for the Entire SHOPP

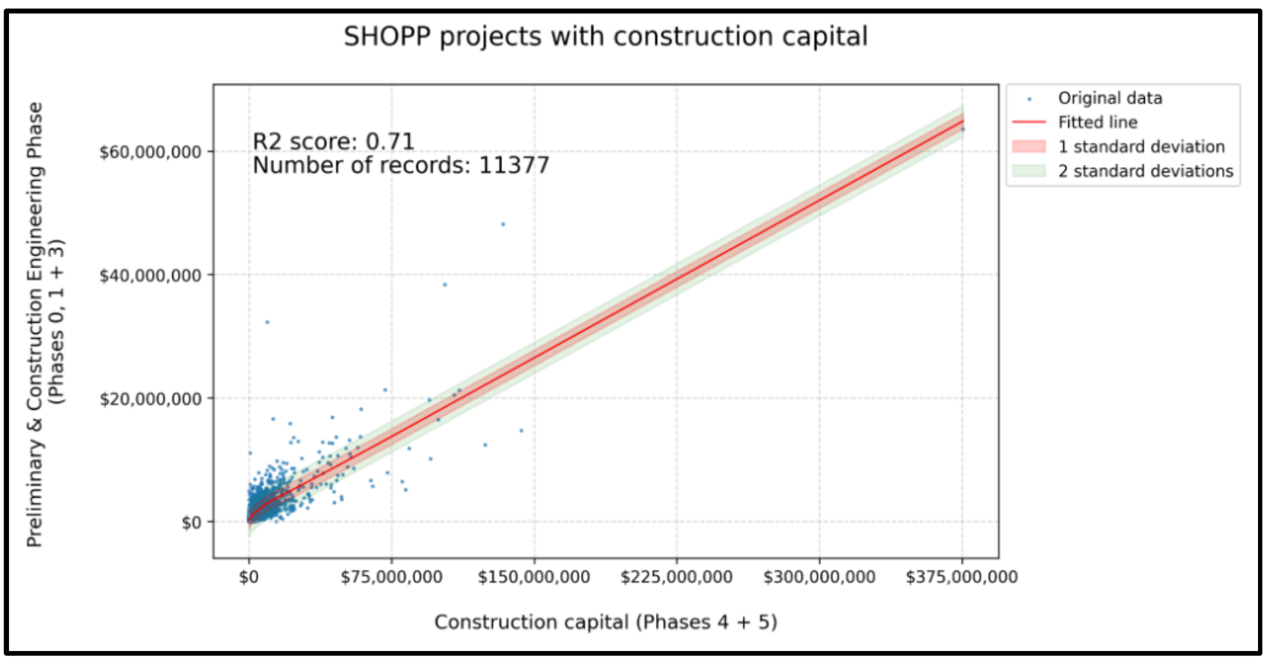


Figure 6. AI Model for Pavement Projects

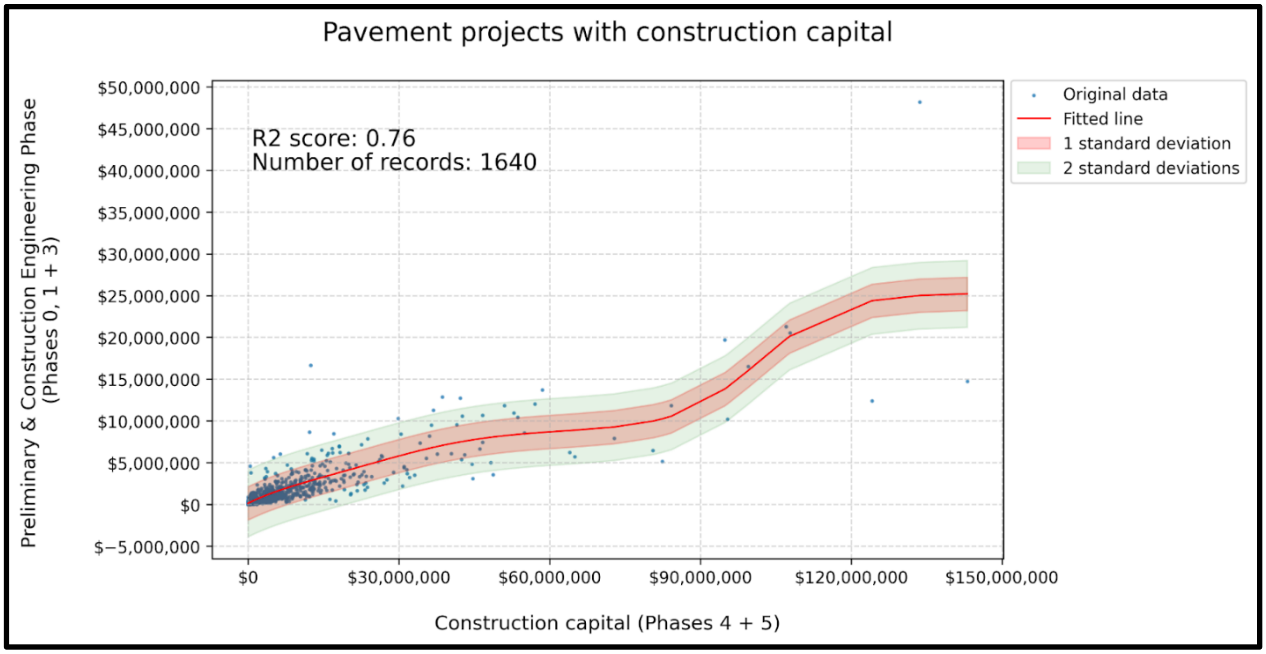


Figure 7. AI Model for Bridge Projects

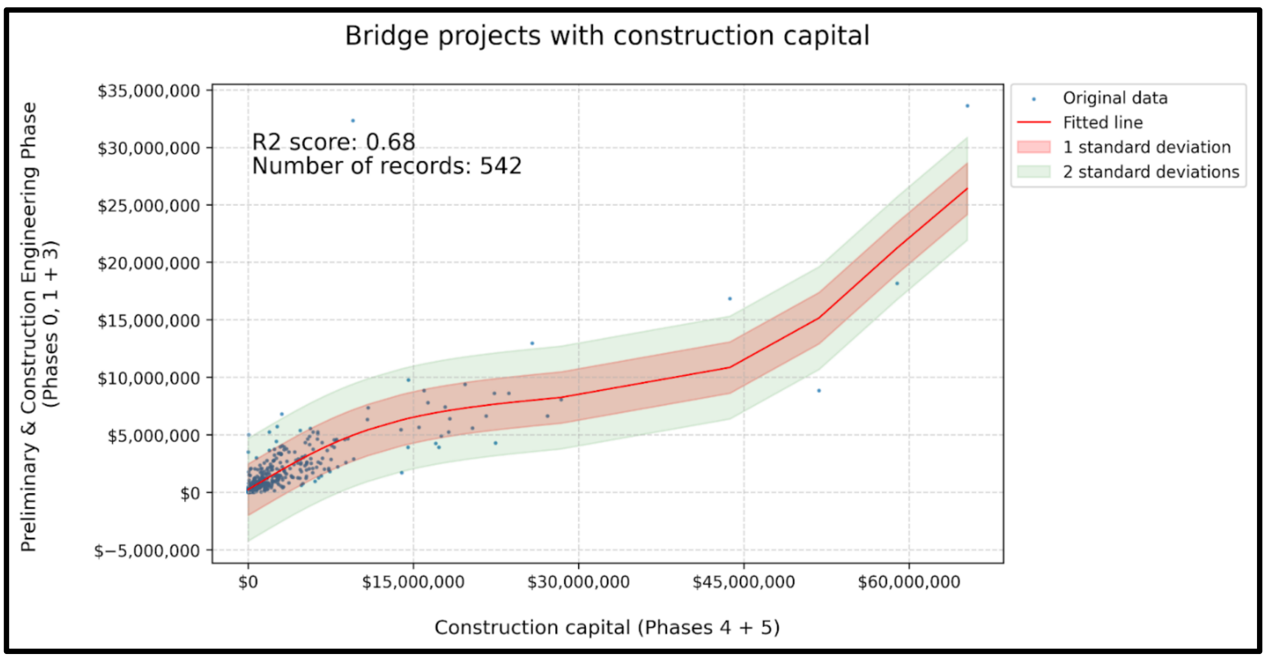


Figure 8. AI Model for Safety Projects

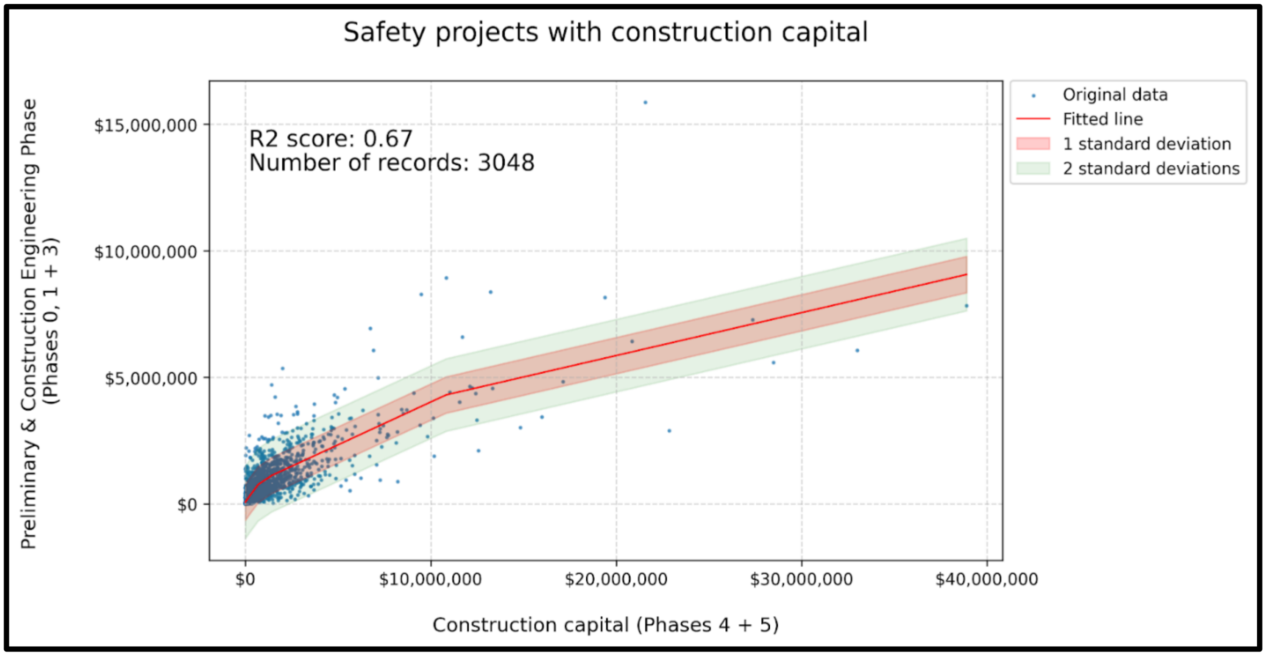


Figure 9. AI Model for Major Damage Projects

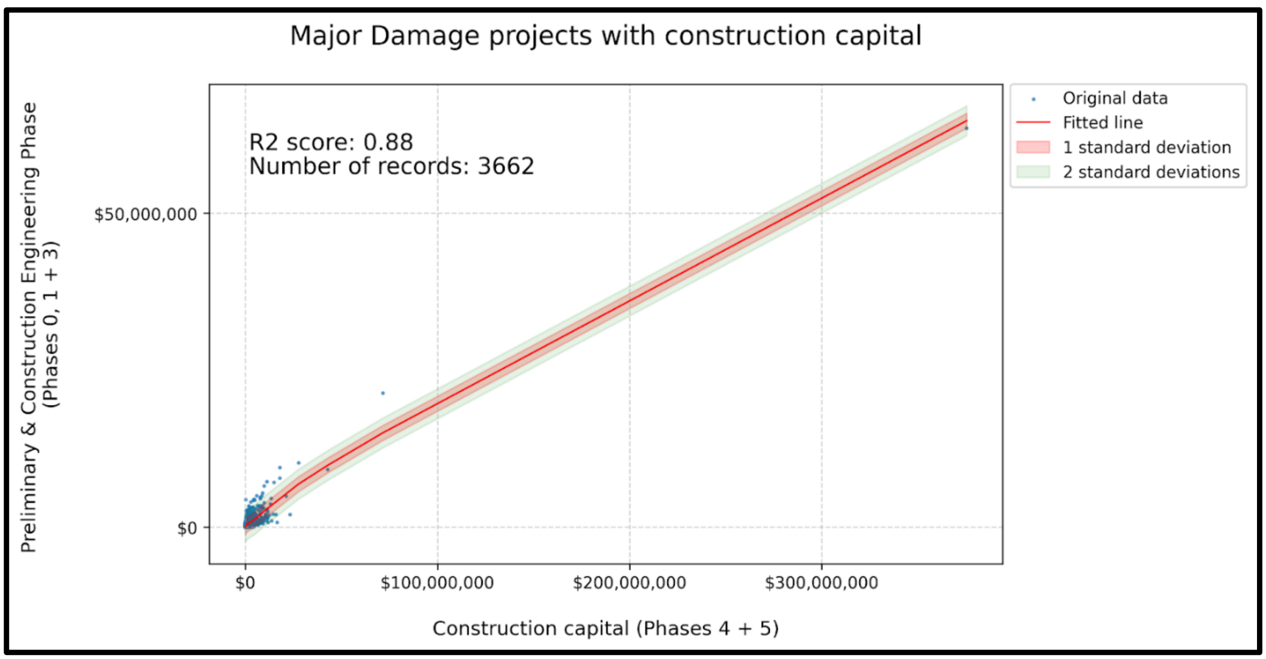


Figure 10. AI Model for 1R Pavement Projects

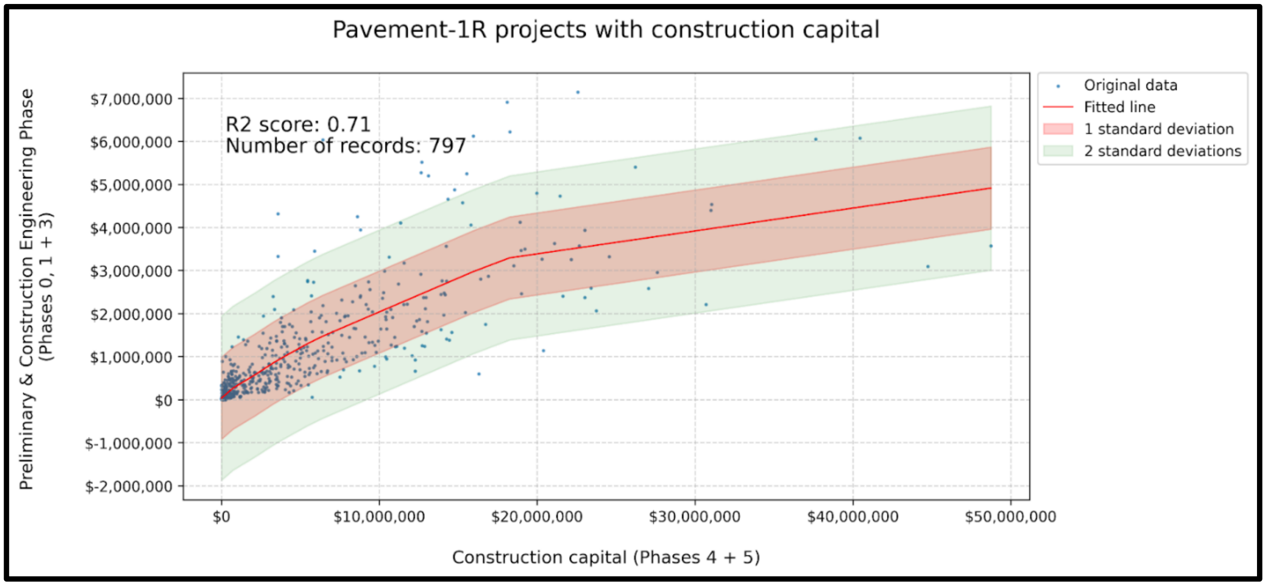


Figure 11. AI Model for 2R Pavement Projects

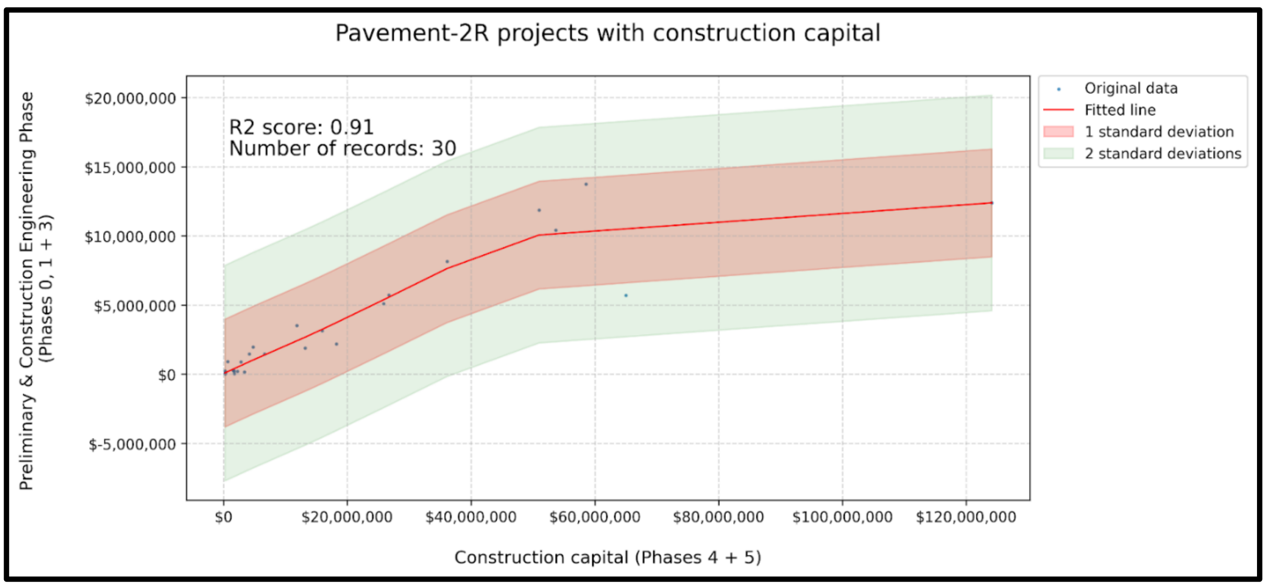
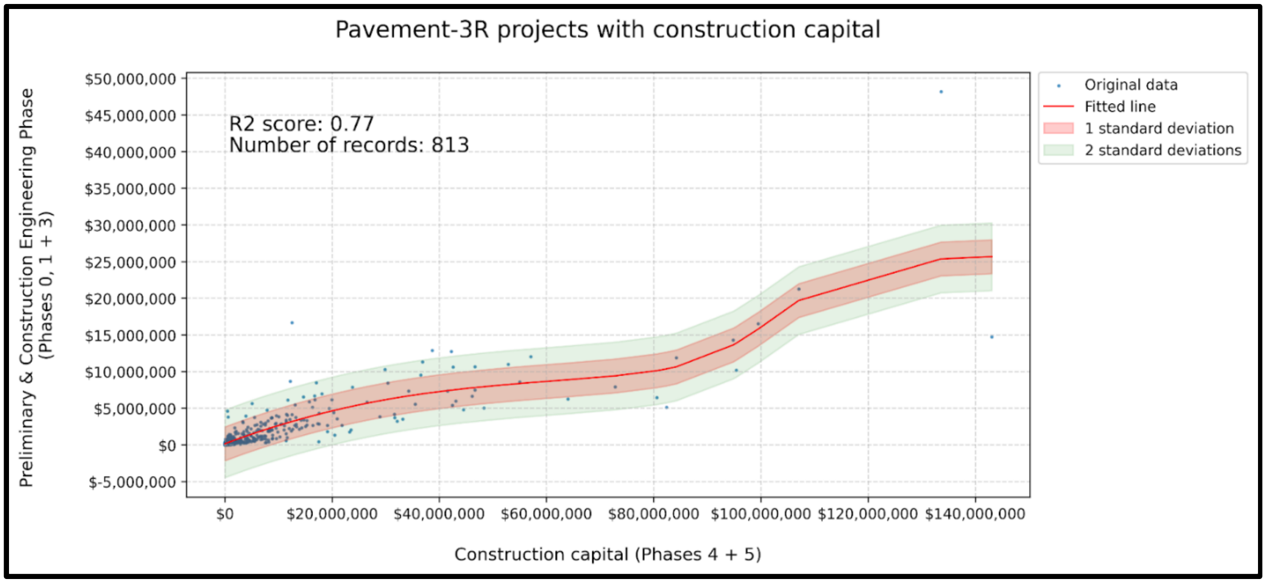


Figure 12. AI Model for 3R Pavement Projects



4.3.3 Comparison with Regression Models

Table 5 compares the results of statistical and AI modeling approaches. As demonstrated in Table 5, the performance of the AI models is slightly better (i.e., up to 13% higher) than the parametric models.

Table 5. Comparison of AI and Regression Models

Portfolio	Number of Projects	Regression R^2	AI R^2
Entire SHOPP	1.1377	0.669	0.71
Pavement	1.640	0.731	0.76
Safety	3.049	0.638	0.67
Bridge	542	0.589	0.68
Major damage	3662	0.860	0.88
Pavement 1R	797	0.658	0.71
Pavement 2R	30	0.783	0.91
Pavement 3R	813	0.716	0.77

5. Summary and Conclusions

5.1 General Overview

The SHOPP is a vital initiative administered by Caltrans aimed at maintaining and preserving the state's highway system. The program addresses crucial aspects like ADA compliance and stormwater control, thereby sustaining and improving the state's transportation network. As a result, accurate cost estimation plays a fundamental role in the successful execution of SHOPP projects. It ensures proper budgeting, resource allocation, and financial planning, which are critical for completing projects within their allocated budgets and timelines. Precise cost estimates help avoid cost overruns, project delays, and inefficient resource utilization, thereby maintaining the reliability and safety of the transportation infrastructure.

The primary objective of this study has been to develop robust portfolio-level cost-estimating models for the SHOPP. This has involved establishing cost norms, comparing regression models and neural network models, and enhancing tools for Caltrans to assess project costs. By achieving these objectives, the study aims to improve the accuracy and reliability of cost estimates for SHOPP, supporting better financial management and decision-making processes for Portfolio of Projects - Portfolio Management.

5.2 Data Collection, Analysis, and Set Development

The data collection and analysis for this project involved a detailed process of compiling, standardizing, and verifying extensive datasets from Caltrans. The primary sources of data included: annual data sets of all Caltrans State Highway project expenditures from 1983 to 2021; detailed Caltrans project bid item data on over 1,000 projects from 2016 to 2021; and data on the primary outputs from each Caltrans project in the SHOPP. The project initially focused on transforming these disparate datasets into a cohesive, usable format.

The annual dataset, consisting of 89 files in various formats, was consolidated into a single expenditure history dataset with 2,396,747 records. A quality assurance check was performed to ensure data integrity. The datasets were developed by aggregating costs across different project phases, focusing on projects active between 2000 and 2022 that included both construction and support costs. These datasets included pavement, major damage, safety, and bridge projects.

A very important aspect of the process was Data Quality Measures, which included Quality Control (QC) and Quality Assurance (QA). QC involved verifying the accuracy and completeness of the combined expenditure history file by comparing sums from the original source files with those in the combined file. This was done using pivot tables for each source file and Python code for the combined file. The checks ensured that the record counts, expenditure amounts, and labor hours matched between the source files and the combined file, confirming the integrity of the data merging process. While QA entailed a secondary, sample-based audit to confirm the QC results.

This involved creating pivot tables for a selection of original expenditure files and a Power Pivot table for the entire combined file, comparing key fields across both datasets. The QA results showed consistent matches in record counts, expenditure amounts, and labor hours, reinforcing the accuracy of the combined dataset. The rigorous QC and QA processes demonstrated that the combined expenditure history file accurately reflects the original expenditure data. Despite minor discrepancies in labor hour totals, which were within acceptable limits of measurement accuracy, the combined file is reliable for future research and reporting. The successful standardization and verification of this large dataset provide a robust foundation for analyzing Caltrans project expenditures and performance over an extended period, supporting ongoing and future transportation infrastructure studies.

5.3 Model Development Process

The development of cost estimation models for the SHOPP portfolio was comprehensive, involving multiple steps to ensure robustness and reliability. Key phases included the engagement of various subject matter experts (SMEs) and ongoing discussions with Caltrans representatives at both district and headquarters levels.

The involvement of industry SMEs was crucial for the model development process, providing essential insights and guidance. Their participation included consultations, which are regular meetings with experienced transportation professionals who offered practical insights into highway project cost estimation, highlighting common industry challenges as well as review and feedback. Draft models and preliminary findings were shared with SMEs for critique, and their feedback was instrumental in refining the models to ensure practical applicability.

Engaging with Caltrans districts and headquarters was key to tailoring the cost estimation models to the specific needs of SHOPP projects. This involved needs assessment, initial discussions to help identify the unique requirements and challenges faced by Caltrans, ensuring the models addressed project cost variations, and data collection and validation. Data on past project expenditures and outcomes were collected and validated through direct communication with Caltrans personnel, providing a solid foundation for model development and feedback on model prototypes. Preliminary versions of the models were shared with Caltrans for feedback. Their practical insights and suggestions were crucial in refining the models for accuracy and usability. Regular updates and ongoing communication ensured that the models remained aligned with evolving needs and priorities, incorporating real-time data and emerging trends.

5.4 Parametric Models

The development of parametric models, also known as statistical, stochastic, or regression models, was guided by established methods in the literature. Three primary forms of parametric functions were considered: (1) additive exponential regression, which includes models where costs are expressed as a sum of exponential terms; (2) linear regression, which is a special case of the additive exponential function with all exponents set to 1; and (3) multiplicative exponential regression, in which costs are expressed as products of terms, which could also be transformed into a logarithmic form.

The initial data set development identified 11,377 SHOPP projects that were analyzed using linear regression. The resulting model demonstrated a coefficient of determination of 0.669, indicating that 66.9% of the variation in engineering costs could be explained by the construction capital cost. The best-fit model highlighted both a fixed engineering cost and a variable component dependent on construction capital.

5.5 AI Model

We developed advanced AI models designed to enhance the accuracy and efficiency of project cost forecasts. These models leverage state-of-the-art machine learning techniques to transform traditional cost estimation practices, ensuring more reliable financial planning and project execution.

The selected models were trained on a large dataset, with a portion set aside for validation to prevent overfitting. We used fine-tuning techniques to ensure the models' robustness. The models' performances were assessed on the basis of metrics such as the Coefficient of Multiple Determination (R2). Once the models were trained and validated, they were implemented into our cost estimation framework.

The AI models demonstrated a significant improvement in the accuracy of cost estimates, as evidenced by relatively high R2 values. This has led to more reliable budgeting, minimizing the risk of cost overruns and financial discrepancies. Additionally, the models have streamlined the estimation process, reducing the time and effort required for project planning. Recognizing the dynamic nature of the construction industry, we suggest establishing mechanisms for the continuous improvement of the AI models. Feedback from completed projects should be systematically incorporated, and the models are regularly updated to reflect new data and changing market conditions. This iterative approach ensures that our cost estimation practices remain cutting-edge and highly effective.

5.6 Key Findings

Enhanced Accuracy: The implementation of advanced AI and statistical cost-estimating models, leveraging Caltrans historical data, boosts forecast accuracy and efficiency for SHOPP projects. By employing sophisticated modeling techniques and integrating comprehensive historical data, the variance between predicted and actual project costs was notably reduced. This improvement minimized budget overruns and financial discrepancies, demonstrating the models' effectiveness in producing accurate estimates.

Efficiency in Estimation: The new cost estimating models significantly streamlined the estimation process. Traditional methods, which were often time-consuming and labor-intensive, were replaced with automated, data-driven approaches. This efficiency gain is critical for meeting project deadlines, optimizing resource allocation, and ensuring that project milestones are met without delays.

Data Utilization: Leveraging historical project data and integrating it with current market trends provided a robust foundation for cost estimation. The use of big data analytics was instrumental in refining the accuracy of the models, allowing for the extraction of valuable insights from large datasets and enhancing the models' predictive power.

Stakeholder Confidence: The improved accuracy and efficiency of the cost estimates increases stakeholder confidence in project budgeting and financial planning. Project stakeholders can rely on these models for more dependable budgeting and financial planning. This increased trust is essential for securing funding, gaining approvals, and maintaining project momentum.

5.7 Contributions to the Field

Innovative Methodologies: This study introduced groundbreaking methodologies in cost estimation for large-scale infrastructure projects. Leveraging modern technological advancements such as machine learning and big data analytics, these methodologies significantly transform traditional estimation practices, contributing to the body of knowledge in transportation project management.

Model Development: The development and rigorous validation of cost estimating models tailored specifically for the SHOPP portfolio represent a significant contribution. These models can serve as a valuable reference for other state highway programs and transportation departments, providing a blueprint for similar initiatives and enhancing the precision and reliability of cost forecasts in the field.

Integration of Technology: By incorporating advanced data analytics and machine learning techniques, this study showcases the transformative potential of technology in traditional cost estimation practices. The successful application of these technologies demonstrates their capability to revolutionize the construction industry's approach to cost estimation, paving the way for future innovations and improvements in accuracy and efficiency.

5.8 Limitations of the Study

Data Limitations: Despite the extensive use of historical data, the study faced significant challenges due to the quality and availability of this data. Inconsistencies, missing entries, and outdated records posed obstacles to achieving the highest possible accuracy in the models.

Scope of Application: The cost estimating models developed are specifically tailored to the SHOPP portfolio. While they offer valuable insights, their applicability to other contexts or regions may be limited without further adaptation and customization, thus restricting the generalizability of the findings to other transportation projects.

Dynamic Market Conditions: The models, while robust, may not fully account for sudden and rapid changes in market conditions. Factors such as unexpected spikes in material costs, labor shortages, or economic downturns can impact cost estimates, necessitating continuous updates and adjustments to maintain accuracy.

5.9 Implications for Practice and Policy

Policy Development: The findings of this study support the creation and implementation of policies that advocate for the adoption of advanced cost estimation models in transportation project planning and budgeting. These policies can enhance financial planning accuracy, reduce the risk of budget overruns, and improve project outcomes at both state and national levels.

Training and Education: To fully leverage the benefits of these advanced models, comprehensive training programs are essential. Project managers, estimators, and other relevant personnel should be equipped with the necessary knowledge and skills to effectively utilize these tools. Such training programs will ensure that the workforce is proficient in modern estimation techniques and can apply them to achieve optimal results.

Continuous Improvement: The study highlights the importance of establishing mechanisms for the continuous improvement of cost estimation models. Agencies should systematically incorporate feedback from completed projects and regularly update the models to adapt to changing market conditions and new data, ensuring their ongoing relevance and accuracy.

The study provides significant insights into improving cost estimation practices for transportation infrastructure projects. It makes substantial contributions to the field of transportation infrastructure project management, particularly in developing and validating advanced cost estimating models for the SHOPP portfolio. By addressing the identified limitations and implementing the recommended practices and policies, transportation agencies can enhance their project planning, budgeting, and execution processes. This will lead to more successful, cost-effective, and timely completion of infrastructure projects, ultimately benefiting the broader community and economy.

6. Next Steps – Recommendations for Additional Research

Considering the findings from this study, there are several opportunities for further research and development. These next steps are aimed at refining the current models, expanding their applicability, and fostering collaboration to advance the field of cost estimation in transportation infrastructure.

- User Friendly Interface Tool: Development of a user-friendly interface to enhance the accuracy and efficiency of cost estimation models specific to the SHOPP portfolio that allows project managers and estimators to input project parameters and receive accurate cost forecasts. The interface also would provide detailed insights into the factors influencing the estimates, helping stakeholders make informed decisions. This includes integration of additional methodologies and technologies to refine portfolio cost estimation processes as well as further evaluation of current models and identification of areas for improvement.

- Optimization Approach: Caltrans performs continuous rehabilitation, replacement, restoration and improvement programs to improve its services. The current research has proposed some sort of a WBS hierarchy structure to form portfolios and develop cost estimation models for each. These portfolios serve six types of projects: (1) pavement; (2) bridges; (3) safety; (4) buildings; (5) roadsides, and (6) drainage. Despite the valuable contribution of the developed cost estimation models for each type of project, the researchers believe they can go a step further and propose additional sets of portfolios that combine multiple types of projects. These portfolios will be formed using optimization techniques that maximize the performance measures of proposed projects and minimize expenditures in limited budget environments. The proposed research work will help Caltrans make the best of its money and provide better services to the public.

- Upper and lower confidence limits: The discussion begun in Sections 4.2.2 and 4.2.3 could be continued in further research on the upper and lower cost confidence limits. The data provided by Caltrans could be used for both a statistical analysis of cost estimate ranges at the bid stage and for further analysis of the ranges of engineering to capital ratios. The usefulness of such ranges has been attested in several sources, including AASHTO (2013).

- Incorporating Real-Time Data: Future research should explore the integration of real-time data into cost estimation models. This can include live market data, weather forecasts, and on-site conditions that can impact project costs. Real-time data integration can enhance the responsiveness and accuracy of the models.

- Refinement of Predictive Algorithms: Continued research should aim at refining the predictive algorithms used in the models. This can involve testing different machine learning approaches, optimizing existing algorithms, and incorporating new data sources to improve prediction accuracy.

- Scenario Analysis and Simulation: Enhancements to the models can include the capability to perform scenario analysis and simulation. This would allow project managers to assess various what-if scenarios, evaluate the impact of different variables, and make informed decisions based on potential outcomes.

- Development of Universal Models: A long-term goal is to develop universal cost estimation models that can be adapted to different types of infrastructure projects and geographical regions. This would involve creating flexible, scalable models that can be customized to meet specific project requirements.

- Automation and Artificial Intelligence: Advancing the automation of cost estimation processes through artificial intelligence is a key long-term goal. Research should focus on developing self-learning models that continuously improve based on new data and project outcomes.

- Comprehensive Cost Management Framework: Developing a comprehensive cost management framework that integrates cost estimation with budgeting, financial planning, and project control is essential. This framework should provide end-to-end support for project managers, ensuring that cost estimates are seamlessly incorporated into overall project management processes.

Appendix A. List of Data Sets of All Caltrans State Highway Project Expenditures

File Name	Transaction Year	Districts	Format	Number of data rows	Total Hours	Total Dollars	Source
ExpenditureHistory1983-84D1-4.txt	1984	1 to 4	Fixed width text	12,117	3,287,698	$ 279,854,082.00	TRAMS
ExpenditureHistory1983-84D5-8.txt	1984	5 to 8	Fixed width text	9,925	3,408,615	$ 344,568,333.00	TRAMS
ExpenditureHistory1983-84D59+.txt	1984	> 59	Fixed width text	586	408,269	$ 49,627,971.00	TRAMS
ExpenditureHistory1983-84D9-58.txt	1984	9 to 58	Fixed width text	28,636	3,618,171	$ 169,363,592.00	TRAMS
ExpenditureHistory1984-85D1-4.txt	1985	1 to 4	Fixed width text	12,478	3,359,867	$ 346,180,497.00	TRAMS
ExpenditureHistory1984-85D44-58.txt	1985	44 to 58	Fixed width text	24,217	2,550,555	$ 74,531,087.00	TRAMS
ExpenditureHistory1984-85D5-8.txt	1985	5 to 8	Fixed width text	10,714	3,690,191	$ 413,659,791.00	TRAMS
ExpenditureHistory1984-85D59+.txt	1985	> 59	Fixed width text	613	593,645	$ 52,121,417.00	TRAMS
ExpenditureHistory1984-85D9-43.txt	1985	9 to 43	Fixed width text	7,893	1,410,056	$ 167,086,068.00	TRAMS
ExpenditureHistory1985-86D1-4.txt	1986	1 to 4	Fixed width text	12,356	3,164,912	$ 401,237,992.00	TRAMS
ExpenditureHistory1985-86D44-58.txt	1986	44 to 58	Fixed width text	22,796	1,316,380	$ 48,064,858.00	TRAMS
ExpenditureHistory1985-86D5-8.txt	1986	5 to 8	Fixed width text	10,368	3,308,220	$ 418,323,546.00	TRAMS
ExpenditureHistory1985-86D59+.txt	1986	> 59	Fixed width text	4,486	1,767,298	$ 73,994,701.00	TRAMS
ExpenditureHistory1985-86D9-43.txt	1986	9 to 43	Fixed width text	9,463	1,296,603	$ 149,520,199.00	TRAMS
ExpenditureHistory1986-87D1-4.txt	1987	1 to 4	Fixed width text	11,833	2,980,191	$ 427,725,707.00	TRAMS
ExpenditureHistory1986-87D5-8.txt	1987	5 to 8	Fixed width text	10,533	3,156,911	$ 496,766,376.00	TRAMS
ExpenditureHistory1986-87D59+.txt	1987	> 59	Fixed width text	5,789	1,840,059	$ 85,718,646.00	TRAMS
ExpenditureHistory1986-87D9-58.txt	1987	9 to 58	Fixed width text	26,239	2,521,236	$ 268,300,645.00	TRAMS
ExpenditureHistory1987-88D1-4.txt	1988	1 to 4	Fixed width text	11,647	3,473,576	$ 440,473,574.00	TRAMS
ExpenditureHistory1987-88D5-8.txt	1988	5 to 8	Fixed width text	10,106	3,424,820	$ 543,589,870.00	TRAMS
ExpenditureHistory1987-88D59+.txt	1988	> 59	Fixed width text	5,949	1,913,424	$ 90,080,873.00	TRAMS
ExpenditureHistory1987-88D9-58.txt	1988	9 to 58	Fixed width text	23,699	2,766,297	$ 298,318,355.00	TRAMS
ExpenditureHistory1988-89D1-4.txt	1989	1 to 4	Fixed width text	12,580	3,911,800	$ 584,635,435.00	TRAMS
ExpenditureHistory1988-89D5-8.txt	1989	5 to 8	Fixed width text	10,567	3,356,503	$ 532,087,957.00	TRAMS
ExpenditureHistory1988-89D59+.txt	1989	> 59	Fixed width text	12,643	2,533,997	$ 121,401,300.00	TRAMS
ExpenditureHistory1988-89D9-58.txt	1989	9 to 58	Fixed width text	21,964	3,273,249	$ 291,635,968.00	TRAMS
ExpenditureHistory1989-90D1-4.txt	1990	1 to 4	Fixed width text	13,225	3,969,942	$ 806,890,427.00	TRAMS
ExpenditureHistory1989-90D5-8.txt	1990	5 to 8	Fixed width text	10,853	3,243,174	$ 606,107,233.00	TRAMS
ExpenditureHistory1989-90D59+.txt	1990	> 59	Fixed width text	14,467	2,604,324	$ 138,737,863.00	TRAMS
ExpenditureHistory1989-90D9-58.txt	1990	9 to 58	Fixed width text	20,177	3,148,716	$ 417,813,135.00	TRAMS

File Name	Transaction Year	Districts	Format	Number of data rows	Total Hours	Total Dollars	Source
ExpenditureHistory1990-91D1-4.txt	1991	1 to 4	Fixed width text	13,881	4,226,938	$ 740,081,567.00	TRAMS
ExpenditureHistory1990-91D5-8.txt	1991	5 to 8	Fixed width text	11,137	3,713,396	$ 802,776,761.00	TRAMS
ExpenditureHistory1990-91D59+.txt	1991	> 59	Fixed width text	18,003	2,817,193	$ 161,755,480.00	TRAMS
ExpenditureHistory1990-91D9-58.txt	1991	9 to 58	Fixed width text	21,700	3,530,558	$ 500,309,900.00	TRAMS
ExpenditureHistory1991-92D1-4.txt	1992	1 to 4	Fixed width text	15,195	4,643,859	$ 739,549,322.00	TRAMS
ExpenditureHistory1991-92D5-8.txt	1992	5 to 8	Fixed width text	12,418	4,279,000	$ 882,812,265.00	TRAMS
ExpenditureHistory1991-92D59+.txt	1992	> 59	Fixed width text	20,626	3,100,052	$ 213,539,003.00	TRAMS
ExpenditureHistory1991-92D9-58.txt	1992	9 to 58	Fixed width text	24,944	4,255,904	$ 607,261,934.00	TRAMS
ExpenditureHistory1992-93D1-4.txt	1993	1 to 4	Fixed width text	14,809	4,544,305	$ 894,257,490.00	TRAMS
ExpenditureHistory1992-93D44-58.txt	1993	44 to 58	Fixed width text	13,656	1,201,201	$ 68,870,461.00	TRAMS
ExpenditureHistory1992-93D5-8.txt	1993	5 to 8	Fixed width text	13,425	4,210,132	$ 772,203,395.00	TRAMS
ExpenditureHistory1992-93D59+.txt	1993	> 59	Fixed width text	22,303	3,081,615	$ 206,610,977.00	TRAMS
ExpenditureHistory1992-93D9-43.txt	1993	9 to 43	Fixed width text	15,953	3,000,093	$ 502,587,028.00	TRAMS
ExpenditureHistory1993-94D1-4.txt	1994	1 to 4	Fixed width text	15,658	4,257,767	$ 922,331,084.00	TRAMS
ExpenditureHistory1993-94D44-58.txt	1994	44 to 58	Fixed width text	13,872	1,129,437	$ 55,415,866.00	TRAMS
ExpenditureHistory1993-94D5-8.txt	1994	5 to 8	Fixed width text	14,483	4,294,642	$ 965,694,630.00	TRAMS
ExpenditureHistory1993-94D59+.txt	1994	> 59	Fixed width text	23,737	3,054,225	$ 202,298,243.00	TRAMS
ExpenditureHistory1993-94D9-43.txt	1994	9 to 43	Fixed width text	18,507	2,899,122	$ 553,794,732.00	TRAMS
ExpenditureHistory1994-95D1-4.txt	1995	1 to 4	Fixed width text	16,643	3,883,192	$ 854,006,124.00	TRAMS
ExpenditureHistory1994-95D44-58.txt	1995	44 to 58	Fixed width text	14,416	780,465	$ 62,451,611.00	TRAMS
ExpenditureHistory1994-95D5-8.txt	1995	5 to 8	Fixed width text	15,095	3,902,636	$ 844,471,953.00	TRAMS
ExpenditureHistory1994-95D59+.txt	1995	> 59	Fixed width text	27,503	3,355,543	$ 198,569,583.00	TRAMS
ExpenditureHistory1994-95D9-43.txt	1995	9 to 43	Fixed width text	19,806	2,624,587	$ 508,992,568.00	TRAMS
ExpenditureHistory1995-96D1-4.txt	1996	1 to 4	Fixed width text	16,925	3,309,701	$ 1,005,458,697.00	TRAMS
ExpenditureHistory1995-96D5-8.txt	1996	5 to 8	Fixed width text	14,804	3,226,232	$ 732,337,404.00	TRAMS
ExpenditureHistory1995-96D59+.txt	1996	> 59	Fixed width text	23,260	3,398,384	$ 232,724,035.00	TRAMS
ExpenditureHistory1995-96D9-58.txt	1996	9 to 58	Fixed width text	25,484	2,964,347	$ 531,428,986.00	TRAMS
ExpenditureHistory1996-97D1-4.txt	1997	1 to 4	Fixed width text	19,396	3,474,542	$ 1,093,594,453.00	TRAMS
ExpenditureHistory1996-97D5-8.txt	1997	5 to 8	Fixed width text	19,425	3,753,446	$ 891,472,810.00	TRAMS
ExpenditureHistory1996-97D59+.txt	1997	> 59	Fixed width text	24,647	3,050,188	$ 209,196,676.00	TRAMS

File Name	Transaction Year	Districts	Format	Number of data rows	Total Hours	Total Dollars	Source
ExpenditureHistory1996-97D9-58.txt	1997	9 to 58	Fixed width text	25,540	2,551,108	$ 596,981,514.00	TRAMS
ExpenditureHistory1997-98D1-4.txt	1998	1 to 4	Fixed width text	20,481	3,682,659	$ 849,084,871.00	TRAMS
ExpenditureHistory1997-98D5-8.txt	1998	5 to 8	Fixed width text	20,414	3,852,282	$ 904,535,846.00	TRAMS
ExpenditureHistory1997-98D59+.txt	1998	> 59	Fixed width text	24,263	2,777,950	$ 253,774,933.00	TRAMS
ExpenditureHistory1997-98D9-58.txt	1998	9 to 58	Fixed width text	27,480	2,735,385	$ 674,661,615.00	TRAMS
ExpenditureHistory1998-99.txt	1999	All	Comma separated value text	105,721	16,996,909	$ 2,866,431,927.37	TRAMS
ExpenditureHistory1999-00.txt	2000	All	Comma separated value text	89,182	17,565,272	$ 2,923,152,922.88	TRAMS
ExpenditureHistory2000-01.txt	2001	All	Comma separated value text	88,917	19,544,140	$ 3,139,952,653.33	TRAMS
ExpenditureHistory2001-02.txt	2002	All	Comma separated value text	87,600	20,482,521	$ 3,592,871,325.51	TRAMS
ExpenditureHistory2002-03.txt	2003	All	Comma separated value text	88,879	19,132,037	$ 3,958,320,084.28	TRAMS
ExpenditureHistory2003-04.xlsx	2004	All	Excel	60,424	17,848,571	$ 3,883,558,984.71	TRAMS
ExpenditureHistory2004-05.xls	2005	All	Excel	52,809	17,418,912	$ 3,775,112,925.67	TRAMS
ExpenditureHistory2005-06.xls	2006	All	Excel	52,965	17,524,935	$ 4,300,258,431.79	TRAMS
ExpenditureHistory2006-07.xls	2007	All	Excel	55,014	18,304,721	$ 5,303,229,274.88	TRAMS
ExpenditureHistory2007-08.xls	2008	All	Excel	54,080	18,167,921	$ 5,711,968,145.34	TRAMS
ExpenditureHistory2008-09.xls	2009	All	Excel	56,881	18,034,690	$ 5,483,176,279.86	TRAMS
Caltrans ROE FY 2010.xls	2010	All	Excel	2	-	$ (852,270.24)	Advantage
Caltrans ROE FY 2011.xls	2011	All	Excel	50,702	10,717,361	$ 4,425,761,516.06	Advantage
Caltrans ROE FY 2012.xls	2012	All	Excel	51,792	11,011,409	$ 5,155,326,139.19	Advantage
Caltrans ROE FY 2013.xls	2013	All	Excel	47,867	10,133,366	$ 4,986,462,314.16	Advantage
Caltrans ROE FY 2014.xls	2014	All	Excel	40,746	9,992,637	$ 5,120,253,085.42	Advantage
Caltrans ROE FY 2015.xls	2015	All	Excel	40,488	10,429,728	$ 4,331,385,370.32	Advantage
Caltrans ROE FY 2016.xls	2016	All	Excel	43,182	10,113,642	$ 4,514,758,092.73	Advantage
Caltrans ROE FY 2017.xls	2017	All	Excel	44,898	9,829,586	$ 4,704,849,199.66	Advantage
Caltrans ROE FY 2018.xls	2018	All	Excel	49,067	9,951,215	$ 4,514,073,312.77	Advantage
Caltrans ROE FY 2019.xls	2019	All	Excel	51,967	10,811,488	$ 4,442,606,637.53	Advantage
Caltrans ROE FY 2020.xls	2020	All	Excel	48,685	11,447,924	$ 5,137,966,455.30	Advantage
Caltrans ROE FY 2021.xls	2021	All	Excel	49,019	11,700,429	$ 5,560,897,402.97	Advantage
Caltrans ROE FY 2022.xls	2022	All	Excel	37,052	5,577,969	$ 2,926,841,781.01	Advantage
			Total records	**2.396.747**			

Appendix B. Example of Field Names and Formats for the Expenditure Records

Field name in final data set	Field name in 1982-1998 data files	Field name 1998-2009 data files	Field name in 2009-2022 data files	Description	Data format (original files)	Data format (final data set)	Required field (i.e., never null)?	Modifications
TransactionYear	TYR	Trans Year Code	Fiscal Year	through June 30, 2009 Trans(action) Year = post July 1, 2009 Fiscal Year. The actual year in which a cost was incurred, years are measured from July 1 to June 30, and listed as the later year (e.g., 2009 = July 1, 2008 to June 30, 2009). In the proposed database, 2-character years need to be converted to 4 characters (e.g. 84 becomes 1984). Some of the newer files have no year field, but there is a year in the file name.	TYR - 2 characters Trans Year Code - 2 characters Fiscal Year - 4 characters	4 characters	Yes	For 1984-1998 data, TYR is converted to 4 digit numeric from 2 characters by prepending 19 if year>80 and year<100 else 20 For 1988-2009 data, Trans Year Code is converted to 4 digit numeric from 2 characters by prepending 19 if year>80 and year<100 else 20 For 2009-2022 data, the Fiscal Year will be imported with no change
FundingFiscalYear	FFY	Funding Fiscal Year Code	n/a	Pre 2009 Funding Fiscal Year = the annual budget that paid for the cost (y=ears are measured from July 1 to June 30). Not provided in the post 2009 data.	FFY - 2 characters Funding Fiscal Year - 2 characters	4 characters	No	For 1984-1998 data, FFY - is converted to 4 digit numeric from 2 characters by prepending 19 if year>80 and year<100 else 20. For 1998-2009 data, Funding Fiscal Year is converted to 4 digit numeric from 2 characters by prepending 19 if year>80 and year<100 else 20 For 2009-2022 data, there is no Funding Fiscal Year in the data. This field will therefore remain null.
SourceDistrict	SD	Source District Code	District	through June 30, 2009 Source District = post July 1, 2009 District. The district that performed the work.	SD - 2 characters Source District Code - 2 characters District - 2 characters	2 characters	Yes	Keep as is

Appendix C. Quality Control: The Records for FFY 1977 in Each Original File

Source File	FFY	Count of FFY
ExpenditureHistory1983-84D1-4.txt	77	4
ExpenditureHistory1983-84D5-8.txt	77	24
ExpenditureHistory1983-84D59+.txt	77	0
ExpenditureHistory1983-84D9-58.txt	77	8
ExpenditureHistory1984-85D1-4.txt	77	3
ExpenditureHistory1984-85D44-58.txt	77	13
ExpenditureHistory1984-85D5-8.txt	77	6
ExpenditureHistory1984-85D59+.txt	77	0
ExpenditureHistory1984-85D9-43.txt	77	0
ExpenditureHistory1985-86D1-4.txt	77	12
ExpenditureHistory1985-86D44-58.txt	77	24
ExpenditureHistory1985-86D5-8.txt	77	6
ExpenditureHistory1985-86D59+.txt	77	0
ExpenditureHistory1985-86D9-43.txt	77	0
ExpenditureHistory1986-87D1-4.txt	77	5
ExpenditureHistory1986-87D5-8.txt	77	14
ExpenditureHistory1986-87D59+.txt	77	2
ExpenditureHistory1986-87D9-58.txt	77	0
ExpenditureHistory1987-88D1-4.txt	77	5
ExpenditureHistory1987-88D5-8.txt	77	10
ExpenditureHistory1987-88D59+.txt	77	0
ExpenditureHistory1987-88D9-58.txt	77	0
ExpenditureHistory1988-89D1-4.txt	77	11
ExpenditureHistory1988-89D5-8.txt	77	9
ExpenditureHistory1988-89D59+.txt	77	0
ExpenditureHistory1988-89D9-58.txt	77	0
	Subtotal	156

Source File	FFY	Count of FFY
ExpenditureHistory1989-90D1-4.txt	77	1
ExpenditureHistory1989-90D5-8.txt	77	0
ExpenditureHistory1989-90D59+.txt	77	3
ExpenditureHistory1989-90D9-58.txt	77	2
ExpenditureHistory1990-91D1-4.txt	77	2
ExpenditureHistory1990-91D5-8.txt	77	4
ExpenditureHistory1990-91D59+.txt	77	0
ExpenditureHistory1990-91D9-58.txt	77	0
ExpenditureHistory1991-92D1-4.txt	77	2
ExpenditureHistory1991-92D5-8.txt	77	1
ExpenditureHistory1991-92D59+.txt	77	2
ExpenditureHistory1991-92D9-58.txt	77	0
ExpenditureHistory1992-93D1-4.txt	77	0
ExpenditureHistory1992-93D44-58.txt	77	1
ExpenditureHistory1992-93D5-8.txt	77	0
ExpenditureHistory1992-93D59+.txt	77	0
ExpenditureHistory1992-93D9-43.txt	77	0
ExpenditureHistory1993-94D1-4.txt	77	0
ExpenditureHistory1993-94D44-58.txt	77	0
ExpenditureHistory1993-94D5-8.txt	77	5
ExpenditureHistory1993-94D59+.txt	77	0
ExpenditureHistory1993-94D9-43.txt	77	0
ExpenditureHistory1994-95D1-4.txt	77	0
ExpenditureHistory1994-95D44-58.txt	77	0
ExpenditureHistory1994-95D5-8.txt	77	3
ExpenditureHistory1994-95D59+.txt	77	0
ExpenditureHistory1994-95D9-43.txt	77	0
	Subtotal	26
	Total of both columns:	**182**

Appendix D. Example of an Original Expenditure File vs. the Single Expenditure History File

TYR	SD	P	E	C	T	Count of M	Sum of M_DOLLARS	Sum of M_HOURS	Transactic	SourceDis	Program	Element	Compone	Task	Count of E	Sum of ExpeditureAmount	Sum of Lal
84	1	20	10	10	0	78	678744	1586.2	1984	01	20	10	010	000	78	678744	1586.2
84	1	20	10	20	100	155	4129379	23428	1984	01	20	10	020	100	155	4129379	23428
84	1	20	10	20	200	291	5100273	62062	1984	01	20	10	020	200	291	5100273	62062
84	1	20	10	20	300	439	6704914	81530.9	1984	01	20	10	020	300	439	6704914	81530.9
84	1	20	10	20	500	22	13115	217	1984	01	20	10	020	500	22	13115	217
84	1	20	10	20	600	24	60297	588.1	1984	01	20	10	020	600	24	60297	588.1
84	1	20	10	40	0	187	1576341	13124.7	1984	01	20	10	040	000	187	1576341	13124.7
84	1	20	20	10	0	190	1253964	21737.9	1984	01	20	20	010	000	190	1253964	21737.9
84	1	20	20	20	110	10	2197	101.6	1984	01	20	20	020	110	10	2197	101.6
84	1	20	20	20	120	7	9	0.3	1984	01	20	20	020	120	7	9	0.3
84	1	20	20	20	200	18	6602	295	1984	01	20	20	020	200	18	6602	295
84	1	20	20	20	300	13	2348	106.8	1984	01	20	20	020	300	13	2348	106.8
84	1	20	20	20	400	27	26693	495.5	1984	01	20	20	020	400	27	26693	495.5
84	1	20	20	30	100	44	135250	2269.2	1984	01	20	20	030	100	44	135250	2269.2
84	1	20	20	30	200	19	6154	261.1	1984	01	20	20	030	200	19	6154	261.1
84	1	20	20	30	300	17	122	4.2	1984	01	20	20	030	300	17	122	4.2
84	1	20	20	30	400	202	2275456	34074.1	1984	01	20	20	030	400	202	2275456	34074.1
84	1	20	50	10	0	2	0	0	1984	01	20	50	010	000	2	0	0
84	1	20	50	10	100	35	726655	6207.8	1984	01	20	50	010	100	35	726655	6207.8
84	1	20	50	10	200	37	65102	2790.6	1984	01	20	50	010	200	37	65102	2790.6
84	1	20	50	10	300	10	91	1.8	1984	01	20	50	010	300	10	91	1.8
84	1	20	50	10	400	118	2636430	63846.9	1984	01	20	50	010	400	118	2636430	63846.9
84	1	20	50	30	0	46	163909	2440.9	1984	01	20	50	030	000	46	163909	2440.9
84	2	20	10	10	0	136	405349	6901.5	1984	02	20	10	010	000	136	405349	6901.5
84	2	20	10	20	100	208	3066594	18220.8	1984	02	20	10	020	100	208	3066594	18220.8
84	2	20	10	20	200	311	12772139	88319.5	1984	02	20	10	020	200	311	12772139	88319.5
84	2	20	10	20	300	239	2277130	29668.8	1984	02	20	10	020	300	239	2277130	29668.8
84	2	20	10	20	500	68	163883	2187.1	1984	02	20	10	020	500	68	163883	2187.1
84	2	20	10	20	600	92	37366	2431.3	1984	02	20	10	020	600	92	37366	2431.3
84	2	20	10	40	0	229	1550332	17617.8	1984	02	20	10	040	000	229	1550332	17617.8
84	2	20	20	10	0	277	1783956	34631.9	1984	02	20	20	010	000	277	1783956	34631.9
84	2	20	20	20	110	25	2449	108.6	1984	02	20	20	020	110	25	2449	108.6
84	2	20	20	20	120	23	570	19.8	1984	02	20	20	020	120	23	570	19.8
84	2	20	20	20	200	78	45778	1922	1984	02	20	20	020	200	78	45778	1922
84	2	20	20	20	300	64	56807	2385.6	1984	02	20	20	020	300	64	56807	2385.6
84	2	20	20	20	400	66	38851	904.5	1984	02	20	20	020	400	66	38851	904.5
84	2	20	20	30	100	47	62851	1263.6	1984	02	20	20	030	100	47	62851	1263.6
84	2	20	20	30	200	79	105766	6056.4	1984	02	20	20	030	200	79	105766	6056.4
84	2	20	20	30	300	68	86089	3291.7	1984	02	20	20	030	300	68	86089	3291.7
84	2	20	20	30	400	179	802593	28616	1984	02	20	20	030	400	179	802593	28616
84	2	20	50	10	0	1	1686	0	1984	02	20	50	010	000	1	1686	0
84	2	20	50	10	100	100	1656510	20741.1	1984	02	20	50	010	100	100	1656510	20741.1
84	2	20	50	10	200	87	1225077	9477.7	1984	02	20	50	010	200	87	1225077	9477.7

Appendix E. Original Nine Files vs. the Single Expenditure History File

Original expenditure file	Totals from each original expenditure file			Totals from the combined expenditure file		
	Number of records	Total expenditures	Total labor hours	Number of records	Total expenditures	Total labor hours
ExpenditureHistory1983-84D1-4.txt	12,117	279,854,082	3,287,697.5	12,117	279,854,082	3,287,697.5
ExpenditureHistory1987-88D5-8.txt	10,106	543,589,870	3,424,820.3	10,106	543,589,870	3,424,820.3
ExpenditureHistory1991-92D9-58.txt	24,944	607,261,934	4,255,903.6	24,944	607,261,934	4,255,903.6
ExpenditureHistory1995-96D59+.txt	23,260	232,724,035	3,398,384.1	23,260	232,724,035	3,398,384.1
ExpenditureHistory2000-01.txt	88,917	3,139,952,653.33	19,544,140.18	88,917	,139,952,653.33	19,544,140.18
ExpenditureHistory2005-06.xlsx	52,965	4,300,258,431.75	17,524,934.81	52,965	4,300,258,431.79	17,524,934.81
Caltrans ROE FY 2011.xls	50,702	4,425,761,516.06	10,717,360.51	50,702	4,425,761,516.06	10,717,360.52
Caltrans ROE FY 2016.xls	43,182	4,514,758,092.73	10,113,641.81	43,182	4,514,758,092.73	10,113,641.87
Caltrans ROE FY 2021.xls	49,019	5,560,897,402.97	11,700,428.83	49,019	5,560,897,402.97	11,700,428.78

Appendix F. Linear Regression Lines for Four SHOPP Sub-portfolios

Figure 13. Linear Regression Lines for Pavement Projects

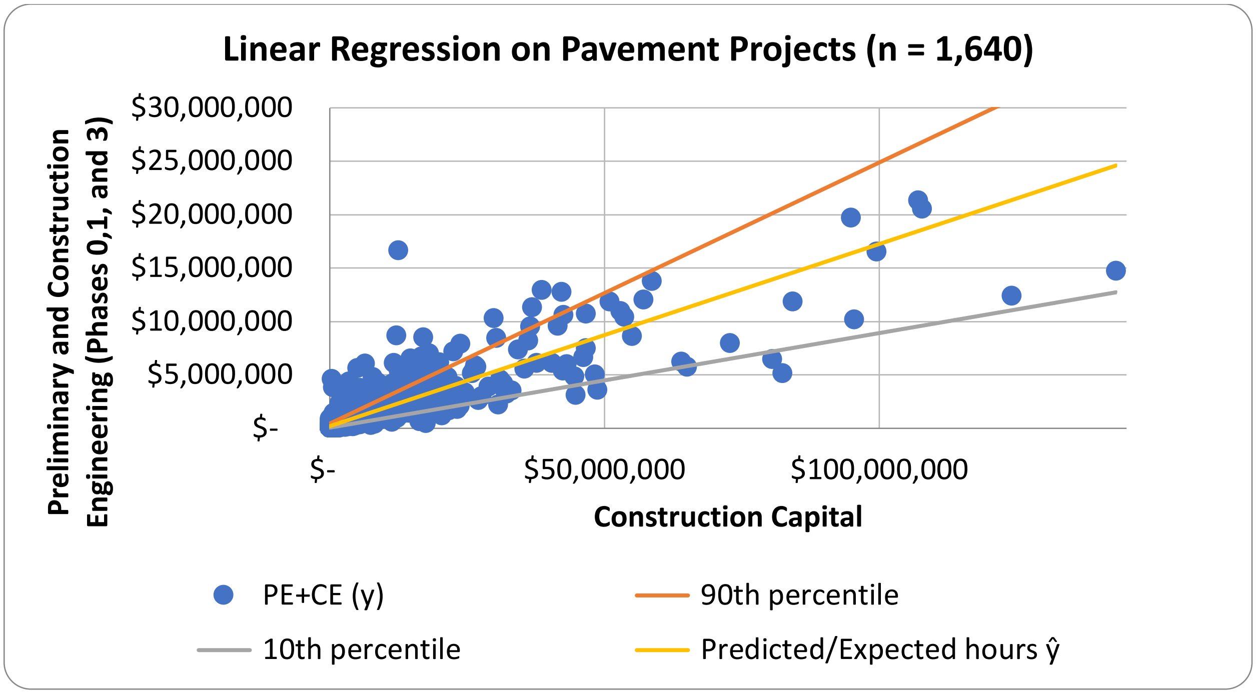


Figure 14. Linear Regression Lines for Bridge Projects

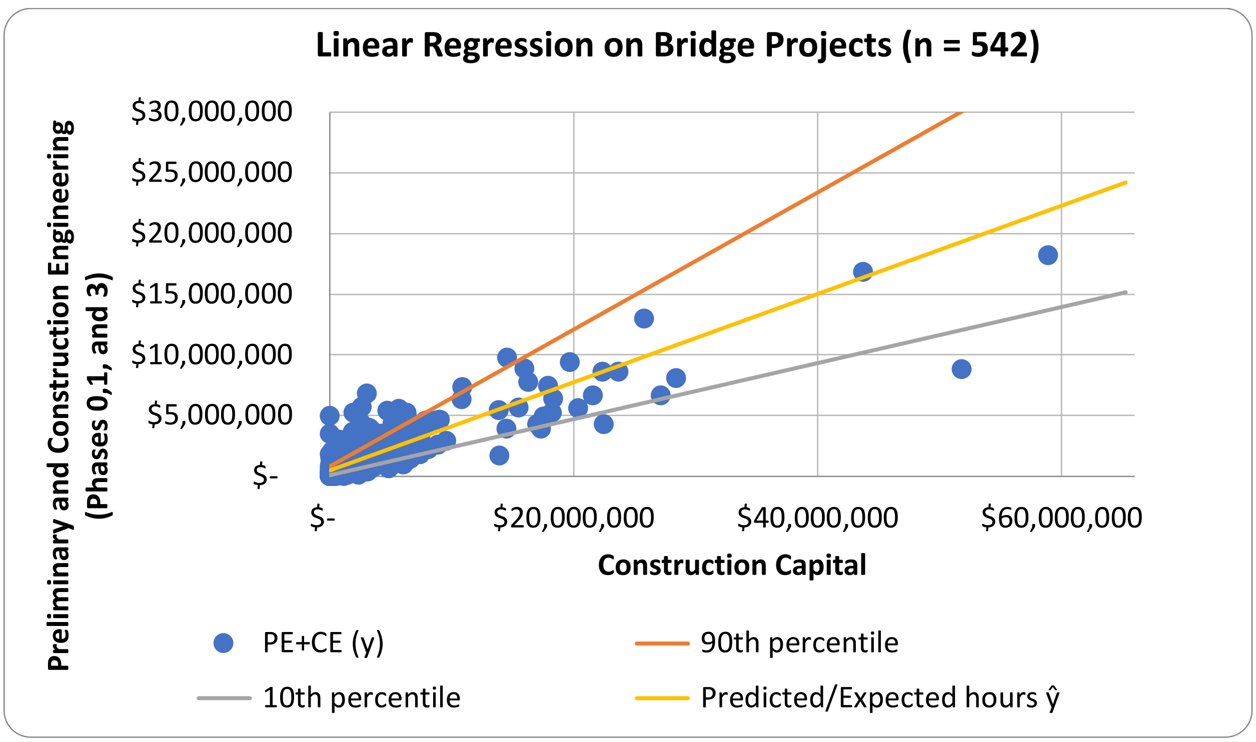


Figure 15. Linear Regression Lines for Safety Projects

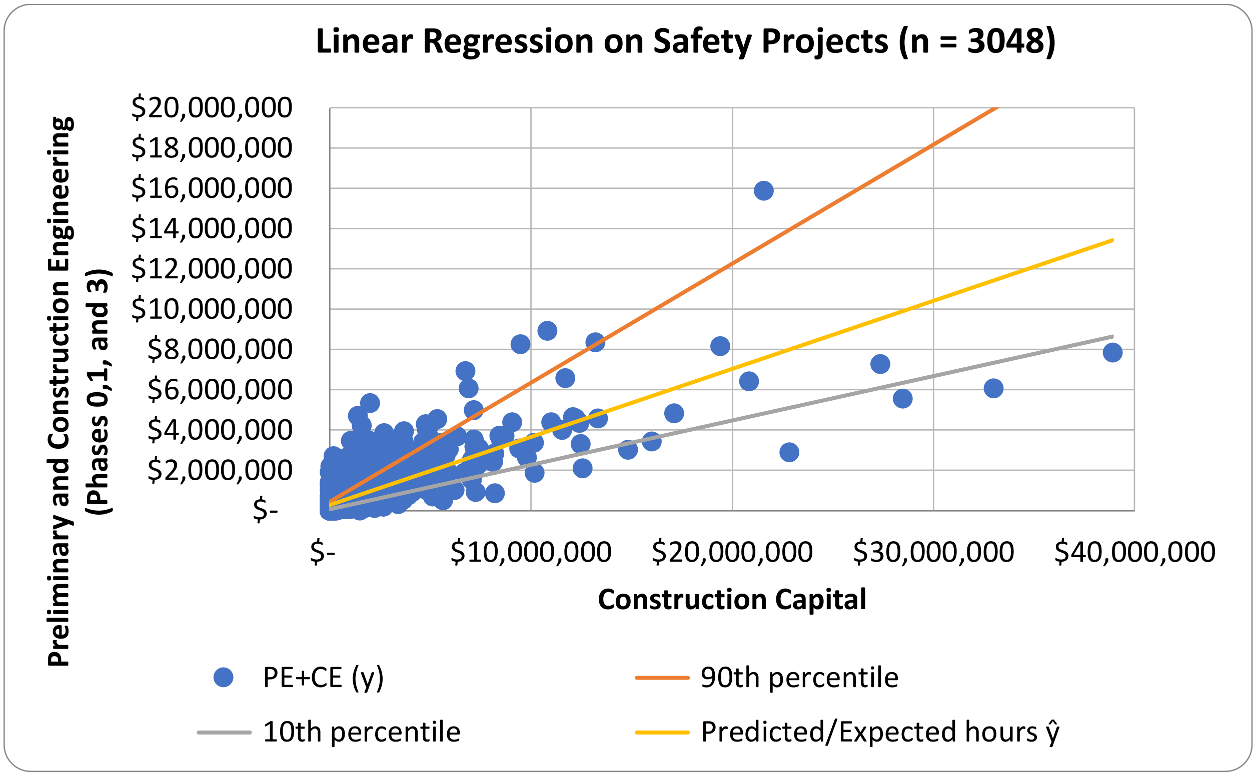
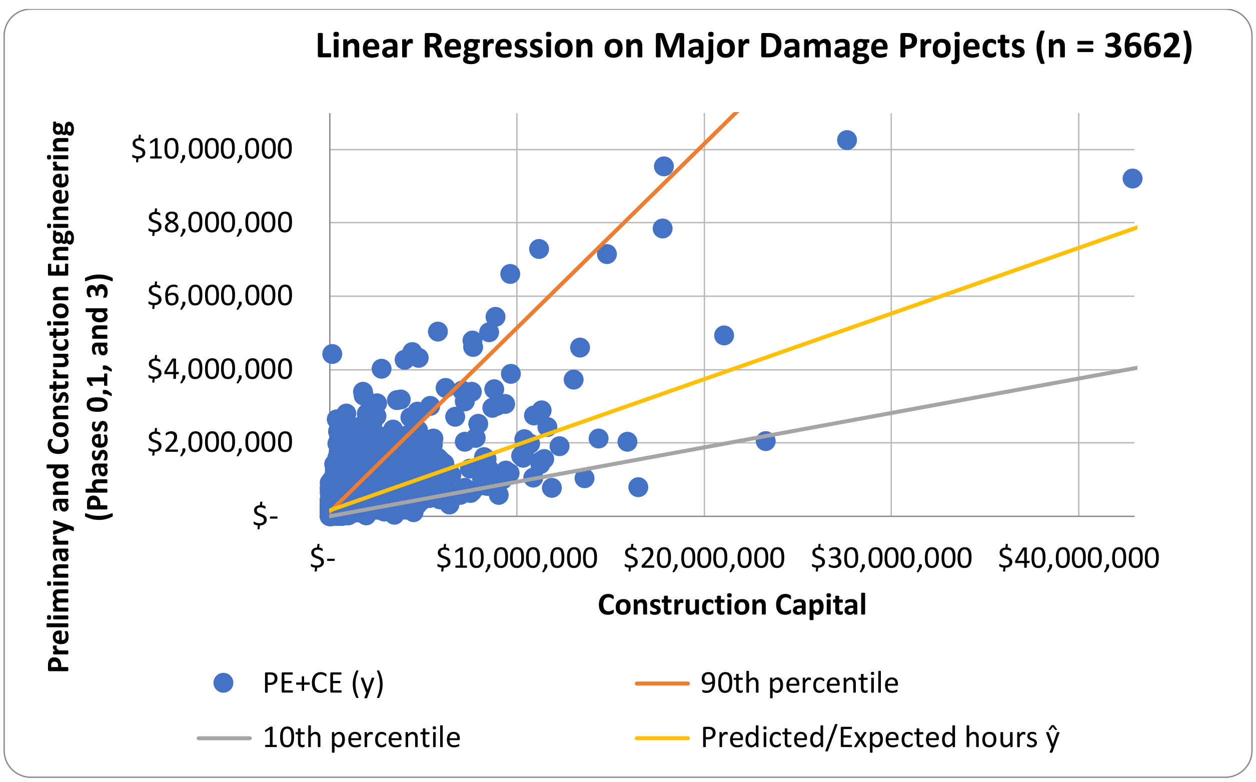


Figure 16. Linear Regression Lines for Major Damage Projects



Bibliography

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

Action TU1003. (2013). *COST*. Retrieved on June 16, 2024, from https://www.cost.eu/actions/TU1003/

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

Akeel, N. F. (1989). *A database tool for statistically based construction cost estimate* (Doctoral dissertation, University of Colorado), Boulder, CO.

Akintoye, A. (2000). Analysis of factors influencing project cost estimating practice. *Construction Management and Economics*, 18(1), 77-89.

Anderson, S., Molenaar, K., and Schexnayder, C. (2007). *Guidance for Cost Estimation and Management for Highway Projects During Planning, Programming, and Preconstruction*, National Cooperative Highways Research Program Report 574, Transportation Research Board, Washington, DC.

Anderson, S., Molenaar, K., and Schexnayder, C. (2009). *Procedures Guide for Right-of-Way Cost Estimation and Cost Management, National Cooperative Highways Research Program Report 625,* Transportation Research Board, Washington, DC.

Araya, F., Faust, K., Khwaja, N., O'Brien, W. J., Liang, X., and Bur, M. K. (2020). Exploring a quantitative and qualitative mixed approach for estimating preliminary engineering efforts of bridge replacement projects. *Transportation Research Record*, 2674(6), 13-22.

Avineri, E. and Cohen, Y. (2018). Optimal road project portfolio selection under limited local authorities' budgets in polynomial time. *14th International Conference on Industrial Logistics*, 29-34.

Barakchi M, Torp O, Belay A. (2017). Cost estimation methods for transport infrastructure: A systematic literature review. *Procedia Engineering*. 196:270–277.

Beşikci, U., Bilge, Ü., and Ulusoy, G. (2015). Multi-mode resource constrained multi-project scheduling and resource portfolio problem. *European Journal of Operational Research*, 240(1), 22-31

Bhuiyan, N. (2011). A framework for successful new product development. *Journal of Industrial Engineering and Management* (JIEM), 4(4), 746-770.

Blampied, N. B. (2018). *Parametric functions for conceptual and feasibility estimating in public highway project portfolios*. UC Berkeley. Retrieved from https://escholarship.org/uc/item/5ks29694

Blampied, N., Shehab, T., Nasr, E., and Samudrala, L. S. (2023). *Preconstruction Support Cost Hours Estimating on Caltrans Pavement Rehabilitation Projects*. Report 23-07, May, 2023. Mineta Transportation Institute, San José State University. DOI: 10.31979/mti.2023.2148

Caltrans. (2024a). *2024 SHOPP State Highway Operation and Protection Program Fiscal Years 2024-25 through 2027-28.* California Department of Transportation. Retrieved on July 30, 2024 from https://catc.ca.gov/-/media/ctc-media/documents/programs/shopp/2024/2024-shopp-document-final.pdf.

Caltrans. (2024b). *Project Development Procedures Manual (PDPM)*. California Department of Transportation. Retrieved on July 1, 2024, from https://dot.ca.gov/programs/design/manual-project-development-procedures-manual-pdpm

Carr, R. I. (1989). Cost-estimating principles. *Journal of Construction Engineering and Management*, 115(4), 545-551.

Christiansen, J. K., and Varnes, C. (2008). From models to practice: decision making at portfolio meetings. *International Journal of Quality and Reliability Management*, 25(1), 87-101.

Doll, C., and van Essen, H. (2008). Road infrastructure cost and revenue in Europe. Produced within the study *Internalisation Measures and Policies for all external cost of Transport* (IMPACT) P Deliverable, 2, 1435.

Elbok, G., and Berrado, A. (2017). Towards an effective project portfolio selection process. In *Proceedings of the international conference on industrial engineering and operations management* (pp. 2158-2169).

El-Kholy A, Tahwia A, Elsayed M. 2020. Prediction of simulated cost contingency for steel reinforcement in building projects: ANN versus regression-based models. *Int J of Construction Managemen*t. 22(9):1675–1689.

ElMousalami, H. H., Elyamany, A. H., and Ibrahim, A. H. (2018). Predicting Conceptual Cost for Field Canal Improvement Projects. *Journal of Construction Engineering and Management,* 144(11), 04018102.

Farshchian, M., Heravi, G. and AbouRizk, S. (2017). Optimizing the Owner's Scenarios for Budget Allocation in a Portfolio of Projects Using Agent-Based Simulation. *Journal of Construction Engineering and Management*, 143(7), 04017022.

Forster, L., Bardos, C. P., and Southgate, D. D. (1987). "Soil erosion and water treatment costs." *J. Soil Water Conserv.*, 42, 349–352.

Frej, E., Ekel, P. and Almeida, A. (2021). A benefit-to-cost ratio based approach for portfolio selection under multiple criteria with incomplete preference information, *Journal of Information Sciences*, 545, 487-498.

Gabriel, S., Ordóñez, J., and Faria, J. (2006). Contingency planning in project selection using multiobjective optimization and chance constraints. *Journal of Infrastructure Systems*, 12(2), 112–120.

Gajpal, Y., and Elazouni, A. (2015). Enhanced heuristic for finance-based scheduling of construction projects. *Construction Management and Economics*, 33(7), 531-553.

Goodarzizad P, Golafshani E, Arashpour M. 2021. Predicting the construction labour productivity using artificial neural network and grasshopper optimization algorithm. *Int J of Construction Management*. 21(5):1–17.

Gransberg, D., Jeong, H. D., Craigie, E. K., Rueda-Benavides, J. A., and Shrestha, K. J. (2016). *Estimating Highway Preconstruction Services Costs. Volume 1: Guidebook*. National Academies of Sciences, Engineering, and Medicine.

Hadjinicolaou, N., and Dumrak, J. (2017). Investigating association of benefits and barriers in project portfolio management to project success. *Procedia Engineering*, 182, 274-281.

Hamaker, J. (1987). Parametric Estimating, in Stewart, R. D., and Wyskida, R. M., (Eds.). *Cost estimator's reference manual, First Edition*. John Wiley and Sons.

Hessami, A. R., Sun, D., Odreman, G. J., Nejat, A., and Saeedi, M. (2017). *Project Scoping Guidebook for Metropolitan Planning Organization Transportation Projects* (No. 0-6929-P1). Texas A & M University.

Hollar, D. A., and Rasdorf, W. (2013). Predictive modeling for transportation infrastructure preconstruction costs. *Public Works Management and Policy*, 18(4), 304-321.

Irfan, M., Khurshid, M. B., Ahmed, A., and Labi, S. (2012). "Scale and Condition Economies in Asset Preservation Cost Functions: Case Study involving Flexible Pavement Treatments." *Journal of Transportation Engineering*, 138(2), 218-228.

Jeong, H. S., and Woldesenbet, A. (2012). *Procedures and models for estimating preconstruction costs of highway projects* (No. OTCREOS10. 1-19-F). Oklahoma Transportation Center.

Klingebiel, R., and Rammer, C. (2014). Resource allocation strategy for innovation portfolio management. *Strategic management journal*, 35(2), 246-268.

Kouskoulas, V. (1984). Cost functions for transportation systems. *Journal of transportation engineering*, 110(1), 129-142.

Lei, T., Xu, J., Jia, X., and Meng, W. (2015). "Spatial Spillover Effects of Highway Transport Infrastructure on Economic Growth in Northwest China" *ICTE 2015*.

Liu, S. S., and Wang, C. J. (2011). Optimizing project selection and scheduling problems with time-dependent resource constraints. *Journal of Automation in Construction*, 20(8), 1110–1119.

Martinsuo, M. (2013). Project portfolio management in practice and in context. *International journal of project management*, 31(6), 794-803.

Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T., and Voordijk, H. (2022). An artificial neural network approach for cost estimation of engineering services. *International journal of construction management*, 22(7), 1274-1287.

Membah, J. F. (2016). *Parametric Cost Estimating and Risk Analysis of Transportation Tunneling Projects* (Doctoral dissertation, North Dakota State University), Fargo, ND.

Müller, R., Martinsuo, M., and Blomquist, T. (2008). Project portfolio control and portfolio management performance in different contexts. *Project management journal*, 39(3), 28-42.

Nevett, R., and Goodrum, P.M. (2022). *Assessment of Federal Highway Administration Highway Project Cost Estimation Tools*. Federal Highway Administration, Publication No. FHWA-HRT-22-075, July 2022. https://doi.org/10.21949/1521882

PMI (2013). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide),* Fifth Edition. Project Management Institute. ANSI/PMI 99-001-2013.

PMI. (2017). *The Standard for Portfolio Management*, Fourth Edition. Project Management Institute. ANSI/PMI 08-003-2017

PMI. (2019). *Practice Standard for Project Estimating*, Second Edition. Project Management Institute.

PMI. (2021). *A guide to the project management body of knowledge (PMBOK® Guide)*—Seventh edition. Project Management Institute. ANSI/PMI 99-001-2021.

Prencipe, A., and Tell, F. (2001). Inter-project learning: processes and outcomes of knowledge codification in project-based firms. *Research policy*, 30(9), 1373-1394.

Shakhsi-Niaei, M., Shiripour, M., and Iranmanesh, S. H. (2015). Application of genetic and differential evolution algorithms on selecting portfolios of projects with consideration of interactions and budgetary segmentation.*International Journal of Operation Research*., 22(1), 106–128.

Shehab T, Nasr E, Farooq M. 2014. Conceptual cost-estimating model for water and sewer projects. In *Pipelines 2014: From Underground to the Forefront of Innovation and Sustainability*. Portland, Oregon. 367–373.

Son, J. and Khwaja, N. (2022). Developing a Preliminary Engineering Cost Estimation Method for a Portfolio of Bridge Construction Projects in Project Planning Phase, *Transportation Research Record*, 2676 (8), 412-422.

Son, J., Khwaja, N. and Milligan, D. (2019). Planning-Phase Estimation of Construction Time for a Large Portfolio of Highway Projects. *Journal of Construction Engineering and Management*, 145(4), 0401918.

State of California (2021). *Statutes of 2021: Budget Act of 2021, Chapter 21* (AB 128). Sacramento, CA: Secretary of State.

Sun, H., Liang, Y., Wang, Y., Liu, X. and Chen, J. (2021). An Optimal Investment Portfolio for BOT Freeway Projects from the Perspective of Transportation and Construction Enterprises, *Korean Society of Civil Engineers*, 25(10), 3636-3649.

Tadeu de Oliveira Lacerda, R., Ensslin, L., and Rolim Ensslin, S. (2011). A performance measurement framework in portfolio management: A constructivist case. *Management Decision*, 49(4), 648-668.

Taghaddos, H., AbouRizk, S., Mohamed, Y., and Hermann, U. (2012). Simulation-based auction protocol for resource scheduling problems. *Journal of construction engineering and management*, 138(1), 31-42.

Tijanić, K., Car-Pušić, D., and Šperac, M. (2020). Cost estimation in road construction using artificial neural network. *Neural Computing and Applications*, 32(13), 9343-9355.

van Niekerk, M, and Bekker, J. (2014). Developing a Tool for Project Contingency Estimation in a Large Portfolio of Construction Projects. *South African Journal of Industrial Engineering*, November, 2014. Vol. 25(3). Pp. 96-111.

Wallace, D.M. (1978). Statistical Analysis for Cost Data. *Transactions of the American Association of Cost Engineers.*

Wang, W.C., Bilozerov, T., Dzeng, R.-J., Hsiao, F.-Y., and Wang, K.C. (2017). Conceptual cost estimations using neuro-fuzzy and multi-factor evaluation methods for building projects, *Journal of Construction Engineering and Management,* 23:1, 1-14.

Washington State DOT. (2023). *Cost Estimating Manual for Projects*, M 30-34.05 January 2023, , Washington State Department of Transporation.

Watanatada, Thawat; Harral, Clell G.; Paterson, William D. O.; Dhareshwar, Ashok M.; Bhandari, Anil; Koji Tsunokawa. (1987) *The highway design and maintenance standards model : Description of the HDM-III model (English*). The Highway design and maintenance standards series Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/129721468326428758/Description-of-the-HDM-III-model

Wei, W., Blampied, N., and Sreevijay, R. (2023). *Evaluation, Comparison, and Improvement Recommendations for Caltrans Financial Programming Processes and Tools*. Mineta Transportation Institute Report 2058. DOI: 10.31979/mti.2023.2058

Xue, X., Jia, Y., and Tang, Y. (2020). Expressway project cost estimation with a convolutional neural network model. *IEEE Access*, 8, 217848-217866.

Zhang, Y., Zhao, P., Wu, Q. Li, B. Huang, J. and Tan, M. (2023). Cost-Sensitive Portfolio Selection via Deep Reinforcement Learning, *IEEE International Conference on Data Engineering*, 3771-3772.

About the Authors

Elhami Nasr, PhD

Dr. Nasr is a Subject Matter Expert in Construction and Project Management. He built his career of over three decades with dual careers in academia and industry. He is a Professor of Construction Engineering Management at California State University, Long Beach. Previously, he served as Vice Provost, Interim Provost, and Vice President of Academic Affairs at Florida Polytechnic University. He also worked in industry (Caltrans) and in various national and international consulting roles, where he initiated, developed, and delivered hundreds of specialized and customized companywide project management training programs for large-scale professional organizations, in both private and government sectors.

Tariq Shehab, PhD

Dr. Shehab is an expert in developing cost estimating applications and the field of artificial intelligence. He has more than 25 years of experience in developing intelligent cost estimating applications, during which he developed systems for highway projects, utility pipe networks, and school buildings. His contributions also included the determination of key factors that attribute to the maintenance costs of major infrastructure facilities. He is the author and co-author of numerous articles published in many ASCE journals and other equally prestigious periodicals.

Nigel Blampied, PhD

Nigel Blampied's research and teaching focuses on project management in public transportation agencies. He teaches in the Master of Transportation Management program at San José State University and is a Research Associate at the Mineta Transportation Institute. His doctoral dissertation at the University of California, Berkeley, discussed parametric cost estimating at the early conceptual phase of projects, the subject, in part, of this report, and used data from the SHOPP, as does this report.

Vinit Kanani

Vinit Kanani earned his master's degree in artificial intelligence from San José State University, where he excelled academically and developed a solid foundation in machine learning and data science. He is currently working as a Data Discovery ETL engineer for Sikka Software, where his responsibilities include designing and implementing data pipelines, ensuring data quality, and applying machine learning techniques to extract valuable insights from complex datasets. His professional interests are centered on machine learning and artificial intelligence, and he is passionate about using these technologies to solve real-world problems.

MTI FOUNDER

Hon. Norman Y. Mineta

MTI BOARD OF TRUSTEES

Directors

Karen Philbrick, PhD
Executive Director

Hilary Nixon, PhD
Deputy Executive Director

Asha Weinstein Agrawal, PhD
Education Director
National Transportation Finance Center Director

Brian Michael Jenkins
Allied Telesis National Transportation Security Center

MINETA TRANSPORTATION INSTITUTE