

NOVEL DATA-DRIVEN APPROACHES FOR ENHANCED PROJECT DURATION  
AND COST-RELATED DECISION MAKING

A Dissertation

by

CHAU HAI LE

Submitted to the Graduate and Professional School of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ivan Damnjanovic
Co-Chair of Committee,	H. David Jeong
Committee Members,	Satish Bukkapatnam
	Kunhee Choi
Head of Department,	Timothy J. Jacobs

August 2021

Major Subject: Interdisciplinary Engineering

Copyright 2021 Chau Hai Le

## **ABSTRACT**

Proper and effective decision-making by project owners during pre-construction phases is highly critical to the successful completion of construction projects. Due to the lack or uncertainty of project information during the project development phases, many essential decisions are typically made under significant uncertainty with decision-makers' assumptions. Ideally, such assumptions need to be validated at the end of construction to improve the decision-making process of future projects. However, post-construction evaluations are currently not actively used in highway projects, and feedback loops to improve early decision making rarely exist. The purpose of this study is to develop alternative approaches to overcome the limitations above and enable continuous improvements in data-driven decision-making for project owners.

Specifically, this study presents novel data-driven approaches for enhancing project time and cost performances via three crucial topics, i.e., contractor evaluation, contract time estimation, and cost estimation, by leveraging pre-existing but underutilized historical project data. Regarding the first topic, a framework was developed to determine actual production rates of controlling activities in a project and enable the objective evaluation of contractors' past production performance using daily work report (DWR) data, considering the influence of common contractor-independent factors on production rates. Concerning the second topic, alternative approaches to construction sequencing were developed by extracting sequential patterns among construction activities from DWR data for different project work types and applying the

extracted patterns for the sequencing of new projects using sequential pattern mining algorithms, statistical tests, and the network theory. Last, a multi-objective optimization-driven approach was designed to find optimal major work items and discover new knowledge and insights for cost estimating in the scoping phase for budget authorization. Each proposed framework or approach was illustrated or validated by a case study using state highway agencies' historical data.

The research findings not only contribute to the body of knowledge but also provide practical approaches to highway agencies to enhance corresponding practices with data-driven systems without collecting any additional data. Furthermore, by periodically applying the proposed approaches to more extensive or newer datasets, the systems can be updated or improved for continuous improvements.

## **DEDICATION**

To my wife and parents for their overwhelming support.

## ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to express my deepest gratitude to my advisor, Dr. H. David Jeong, for his guidance and support throughout this research. It has been a privilege to conduct research under his supervision in the past four years. Regular discussions with and encouragements from him have guided and motivated me in this amazing and challenging journey and prepared me with the necessary skills to succeed in my future professional career.

I would also like to thank my co-advisor and committee chair, Dr. Ivan Damnjanovic, for his guidance and support in the past two years. My appreciation also goes to the committee members, Drs. Satish Bukkapatnam and Kunhee Choi, for their efforts and contributions to this work. Also, I am grateful to Dr. Bukkapatnam for his support during my transfer from Iowa State University to Texas A&M University and my change of majors.

Finally, thanks to my parents for their encouragement and to my wife for her patience and love.

## **CONTRIBUTORS AND FUNDING SOURCES**

### **Contributors**

This work was supervised by a dissertation committee consisting of Dr. Ivan Damnjanovic of the Zachry Department of Civil Engineering (Chair & Co-advisor), Dr. H. David Jeong of the Department of Construction Science (Co-Chair & Advisor), Dr. Satish Bukkapatnam of the Department of Industrial and Systems Engineering (Member), and Dr. Kunhee Choi of the Department of Construction Science (Member).

All work conducted for the dissertation was completed by the student independently.

### **Funding Sources**

This work was made possible in part by the National Cooperative Highway Research Program (NCHRP) under Grant Number 08-114. Its contents are solely the responsibility of the author and do not necessarily represent the official views of the Transportation Research Board.

## TABLE OF CONTENTS

	Page
ABSTRACT .....	ii
DEDICATION .....	iv
ACKNOWLEDGEMENTS .....	v
CONTRIBUTORS AND FUNDING SOURCES.....	vi
TABLE OF CONTENTS .....	vii
LIST OF FIGURES.....	xi
LIST OF TABLES .....	xiv
1. INTRODUCTION.....	1
1.1. Background and motivation .....	1
1.2. Problem statement .....	3
1.3. Research objectives .....	7
1.4. Research methodology .....	8
1.5. Expected contribution .....	11
1.6. Dissertation organization .....	13
1.7. References .....	13
2. EVALUATING CONTRACTORS' PRODUCTION PERFORMANCE IN HIGHWAY PROJECTS USING HISTORICAL DAILY WORK REPORT DATA .....	18
2.1. Overview .....	18
2.2. Introduction .....	19
2.3. Background .....	22
2.3.1. Qualification criteria.....	22
2.3.2. Contractor qualification models .....	24
2.4. DWR-based framework for evaluating contractors' production performance.....	26
2.4.1. Step 1—Estimation and classification of production rates .....	27
2.4.2. Step 2—Development of a contractor evaluation system .....	31
2.4.3. Step 3—Application of the evaluation system for a new project.....	34
2.5. Case study .....	35
2.5.1. Step 1—Estimation and classification of production rates .....	35

2.5.2. Step 2—Development of a contractor evaluation system .....	39
2.5.3. Step 3—Application of the evaluation system for a new project.....	47
2.6. Discussion and conclusions.....	47
2.7. References .....	50
3. A SEQUENTIAL PATTERN MINING DRIVEN FRAMEWORK FOR DEVELOPING CONSTRUCTION LOGIC KNOWLEDGE BASES .....	58
3.1. Overview .....	58
3.2. Introduction .....	59
3.3. Literature review .....	61
3.3.1. Highway scheduling practices.....	61
3.3.2. Prior studies of construction sequencing.....	63
3.4. Research objective and scope.....	65
3.5. Methods and concepts utilized for framework development .....	67
3.5.1. Daily work report data.....	67
3.5.2. Sequential pattern mining.....	68
3.5.3. Theoretical foundation of extracting construction sequential patterns from DWR data .....	70
3.6. Framework for developing knowledge bases of construction sequencing.....	71
3.6.1. Step 1: Select a level of the WBS as the basis of the knowledge base .....	72
3.6.2. Step 2: Create a sequence database suitable for applying SPM algorithms ...	74
3.6.3. Step 3: Apply an SPM algorithm to obtain a list of frequent subsequences or patterns and their support.....	75
3.6.4. Step 4: Propose and calculate domain-specific measures and build a broad knowledge base of construction sequencing .....	76
3.6.5. Step 5: Examine the effect of project types on the discovered patterns and build a project type-specific knowledge base if necessary.....	78
3.6.6. Step 6: Input work items of interest and extract relevant patterns from the knowledge base .....	78
3.7. Case study .....	79
3.7.1. Step 1: Select a level of the WBS as the basis of the knowledge base .....	79
3.7.2. Step 2: Create a sequence database suitable for applying SPM algorithms ...	80
3.7.3. Step 3: Apply an SPM algorithm to obtain a list of frequent subsequences/patterns and their support .....	81
3.7.4. Step 4: Propose and calculate domain-specific measures and build a broad knowledge base of construction sequencing .....	82
3.7.5. Step 5: Examine the effect of project types on the discovered patterns and build a project type-specific knowledge base if necessary.....	83
3.7.6. Step 6: Input work items of interest to schedulers and extract relevant patterns from the knowledge base .....	85
3.8. Comparison between the expected outputs of the template approach and the SPM-driven approach.....	86
3.9. Discussions and conclusions .....	88



3.10. References .....	91
<b>4. NETWORK THEORY DRIVEN CONSTRUCTION LOGIC KNOWLEDGE NETWORK: PROCESS MODELING AND APPLICATION IN HIGHWAY PROJECTS.....</b>	<b>97</b>
4.1. Overview .....	97
4.2. Introduction .....	98
4.3. Background .....	101
4.4. Research objective and scope.....	103
4.5. Process model to develop and apply a sequencing knowledge network.....	104
4.5.1. Step 1: Select a project work type for network development and determine its list of activities.....	105
4.5.2. Step 2: Develop a sequence database of the selected project work type from the DWR data .....	107
4.5.3. Step 3: Apply an SPM algorithm to the sequence database to extract all two-event subsequences and their occurrence frequencies .....	109
4.5.4. Step 4: Determine a representative sequential pattern for each pair of events from the two events' subsequences.....	109
4.5.5. Step 5: Interlink the identified sequential patterns and develop a network of as-built construction sequence patterns.....	112
4.5.6. Step 6: Implement algorithms to apply the developed network for sequencing a new project .....	114
4.6. Case study .....	118
4.6.1. Step 1: Select a project work type for network development and determine its list of activities.....	119
4.6.2. Step 2: Develop a sequence database of the selected project work type from the DWR data .....	120
4.6.3. Step 3: Apply an SPM algorithm to the sequence database to extract all two-event subsequences and their occurrence frequencies .....	121
4.6.4. Step 4: Determine a representative sequential pattern for each pair of events from the two events' subsequences.....	122
4.6.5. Step 5: Interlink the identified sequential patterns and develop a network of as-built construction sequence patterns.....	124
4.6.6. Step 6: Implement algorithms to apply the developed network for sequencing a new project .....	125
4.7. Discussion and practical implications .....	128
4.8. Summary and conclusions.....	130
4.9. References .....	132
<b>5. PARETO PRINCIPLE IN SCOPING-PHASE COST ESTIMATING: A MULTI-OBJECTIVE OPTIMIZATION APPROACH FOR SELECTING AND EVALUATING OPTIMAL MAJOR WORK ITEMS .....</b>	<b>137</b>

5.1. Overview .....	137
5.2. Introduction .....	138
5.3. Research scope and objective.....	142
5.4. Methodology .....	144
5.4.1. Input data.....	145
5.4.2. Model development.....	146
5.4.3. Model output .....	151
5.5. Data analysis and results .....	152
5.5.1. Model 1.....	152
5.5.2. Model 2.....	156
5.6. Discussion and practical implications .....	161
5.7. Summary and conclusions.....	163
5.8. References .....	165
6. CONCLUSIONS.....	169

## LIST OF FIGURES

	Page
Fig. 1.1. Continuous improvements in decision making by project owners .....	3
Fig. 1.2. Overall research methodology .....	9
Fig. 2.1. Overview of the proposed framework .....	27
Fig. 2.2. Classification of production rates .....	31
Fig. 2.3. Distribution fitting for excavation-unclassified (EU)—Tier 3 (m <sup>3</sup> /day).....	38
Fig. 2.4. Production rate distributions for Tiers 1 and 2, excavation-unclassified (m <sup>3</sup> /day).....	38
Fig. 2.5. Production rate distributions for Tiers 2 and 3, excavation-unclassified (m <sup>3</sup> /day).....	39
Fig. 2.6. Project locations and boundaries of urban areas.....	40
Fig. 2.7. Cluster analysis of project amounts .....	42
Fig. 2.8. Production rate distributions for Tiers 1 and 2, plant mix surfacing (PMS) (t/day).....	46
Fig. 2.9. Production rate distributions for Tiers 2 and 3, plant mix surfacing (PMS) (t/day).....	47
Fig. 3.1. DWR data attributes.....	68
Fig. 3.2. An example of a sequence database.....	69
Fig. 3.3. Extracting precedence relationships between two work items from DWR data.....	71
Fig. 3.4. An overall framework for developing a knowledge base of construction sequencing .....	72
Fig. 3.5. Transforming DWR data to sequence databases .....	74
Fig. 3.6. An output of applying an SPM algorithm to a sequence database .....	75
Fig. 3.7. Illustration of the average distance measure.....	77

Fig. 4.1. Process model to develop and apply a sequencing knowledge network .....	104
Fig. 4.2. DWR data attributes.....	106
Fig. 4.3. Transform the DWR data of a project into a sequence .....	108
Fig. 4.4. Visualization of a pattern .....	112
Fig. 4.5. Find the lag between two events of a pattern.....	112
Fig. 4.6. An example of a directed network .....	113
Fig. 4.7. The simplified version of subnetwork N1 .....	116
Fig. 4.8. An example of the lag time of a pattern.....	124
Fig. 4.9. A network of construction sequence patterns of overlay projects .....	125
Fig. 4.10. Successors of event 272 – the finish of the pavement mix surfacing activity	126
Fig. 4.11. Predecessors of event 171 – the start of the curbs and gutters activity.....	127
Fig. 4.12. Sequential relationships among a sample set of activities .....	128
Fig. 5.1. Timing of scoping-phase cost estimating in the project development phases .	142
Fig. 5.2. Data attributes of historical bid data .....	143
Fig. 5.3. Proposed multi-objective optimization models .....	144
Fig. 5.4. Calculations of evaluation metrics.....	148
Fig. 5.5. Optimal trade-offs between mean and <i>CV</i> of cost percentages of a 42-major-item set over total project cost in HMA resurfacing projects (Work type 1523).....	153
Fig. 5.6. Convergence of optimal trade-offs solutions .....	154
Fig. 5.7. Optimal trade-offs between mean and <i>CV</i> of cost percentages of a major item set over total project cost in different projects: A comparison between two work types (1523 — HMA resurfacing and 1014 — PCC pavement)....	155
Fig. 5.8. Optimal trade-offs between the mean of cost percentages of a 42-major-item set over total project cost in HMA resurfacing projects and MAPE using mean.....	157

Fig. 5.9. Optimal trade-offs between the median of cost percentages of a 42-major-item set over total project cost in HMA resurfacing projects and MAPE using median .....	158
Fig. 5.10. Comparison between the mean and median of cost percentages of a 42-major-item set over total project cost in HMA resurfacing projects .....	159
Fig. 5.11. Comparison between MAPE using mean and MAPE using median on the optimization dataset and the hold-out dataset.....	160
Fig. 5.12. Changes in average MAPE using median with four-fold cross-validation and increases in the number of major items .....	161

## LIST OF TABLES

	Page
Table 2.1. Statistical measures of production rates from DWR data .....	36
Table 2.2. Means of production rates (per day) for two area groups .....	41
Table 2.3. Means of production rates (per day) for two budget groups .....	43
Table 2.4. Means of production rates (per day) for two weather groups .....	44
Table 2.5. Comparison of production rates (per day) of three quantity groups .....	45
Table 3.1. An example of WBS .....	73
Table 3.2. List of sections and corresponding frequencies in the DWR data .....	79
Table 3.3. Examples of the discovered patterns, the output measure from SPM in Step 3, and proposed domain-specific measures in Step 4 .....	82
Table 3.4. Pattern “{Aggregate Surfacing (301)}, {Bituminous Materials (402)}” under different project types .....	84
Table 3.5. Examples of extracting the relevant patterns of interest from the reconstruction and grading knowledge base .....	85
Table 3.6. Construction sequence patterns of the reconstruction, grading project type under the schedule template approach and the SPM-driven approach .....	87
Table 4.1. Algorithm S1 – Find the successors of an event .....	116
Table 4.2. Algorithm S2 – Find the predecessors of an event .....	117
Table 4.3. Algorithm S3 – Find the sequencing relationships among a set of activities .....	118
Table 4.4. List of construction activities .....	120
Table 4.5. Examples of pairs of events and their subsequences .....	122
Table 4.6. Examples of pairs of events and their representative sequential patterns .....	123
Table 5.1. STAs' guidance on applying the Pareto principle to cost estimating .....	143
Table 5.2. Top five work items of four different work types .....	146

# 1. INTRODUCTION

## 1.1. Background and motivation

Construction project outcomes are not only affected by contractors' operations during construction but also by many decisions made by project owners in different project development phases before construction starts. Some representative examples include a) cost estimates in the planning and scoping phases necessary to estimate potential funds for projects and authorize project budgets (AASHTO 2013), b) contract time estimation and determination in the design phase to dictate the required completion date or duration of a project (TxDOT 2018), and c) contractor selection in the letting phase that directly affects the successful completion of a project.

Project maturity or definition increases as projects progress from planning to letting (AASHTO 2013; MnDOT 2008). Methods and tools used for even the same task vary with the amount and level of detail of input information, decision purposes, and required accuracy (AASHTO 2013). For example, conceptual project-level cost estimating is used in the planning phase, while detailed activity-level estimating is required in the final design phase (Anderson et al. 2007; Elmousalami 2020). Estimates in a later phase are expectedly more robust, more accurate, and closer to the actual construction cost at the end of construction as fewer assumptions are needed to deal with the lack of or uncertainty in the project information.

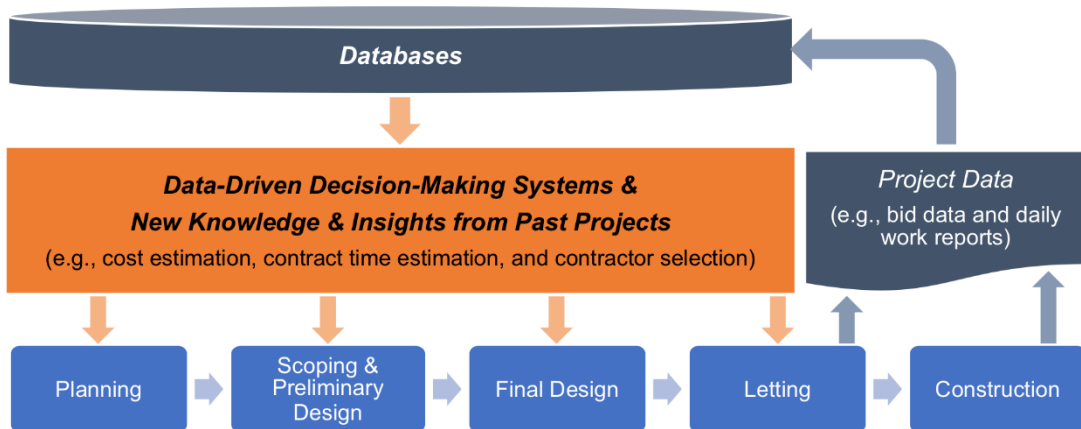
Post-construction evaluation can help validate assumptions made in pre-construction phases and therefore enhance future decision-making processes. Lessons

learned can also be captured to avoid repeating mistakes made in past projects and seek continuous improvements (Garsden 1995). However, the use and effectiveness of post-construction evaluations in state highway agencies (SHAs) are still limited (Taylor et al. 2017). One of the main reasons is the temporality of the construction industry (Carrillo 2005). Once a project is complete, the project's team is quickly disbanded, with team members transferring to a new project/position or leaving the agency (Kärnä and Junnonen 2005).

Since digital project data have been collected and retained by SHAs (Shrestha and Jeong 2017), data-driven approaches have become promising alternatives to continuous decision-making improvements. Historical project data, such as digital daily work reports (DWRs), are still primarily used for administrative purposes only or not fully leveraged for decision-making processes (Shrestha and Jeong 2017).

Fig. 1.1 shows the concept of continuous decision-making improvements. Historical project data, especially those in the letting and construction phases, are stored in databases and used for developing data-driven systems and gaining new knowledge and insights to support the decision-making of new projects in earlier phases. The new projects' data are, in turn, retained in the databases to improve or update the systems, creating a feedback loop for continuous improvements. Furthermore, such data-driven systems can help alleviate the heavy dependence of SHAs on decision-makers' judgments and experiences in the current decision-making processes by providing data-backed solutions.





**Fig. 1.1.** Continuous improvements in decision making by project owners

## 1.2. Problem statement

SHAs have spent significant efforts collecting historical project data (e.g., DWRs or bid tabulation) (Tang and McHale 2016). However, they have not fully leveraged the collected data to improve their current business practices (Shrestha and Jeong 2017). The data can help improve SHAs' crucial decisions before construction, ultimately enhancing project performances, particularly from time and cost perspectives. For example, evaluating contractors' past production performances in contractor prequalification or selection can provide project owners with greater assurance that the selected contractors will complete their projects on time. Reasonable contract time determination is also crucial to on-time and on-budget project completion (FHWA 2002; Le et al. 2021). Also, reasonable budget decisions are critical to project outcomes (Gardner et al. 2017).

One of the most critical tasks of a project's owner is to select the right contractor for the project due to its direct influences on project outcomes (Afshar et al. 2017; Chini et al. 2018). One dominant criterion for the selection is contractor bid prices. However,

selecting the lowest bidder does not guarantee the lowest cost at the end of construction due to additional costs associated with possible claims, delays, or low construction quality (Chini et al. 2018; Pesek et al. 2019). Therefore, project owners such as SHAs have also considered factors other than bid prices via pre-qualification or post-qualification processes for contractor selection (Dye Management Group 2014; Forcada et al. 2017). Some examples are experiences, financial capabilities, and technical abilities. However, SHAs have rarely considered a contractor's past production performance in the selection process despite its effect on project durations. The common notion that *"The best predictor of future behavior is past behavior"* might be true to contractors.

Additionally, the current body of knowledge has relied on decision-makers' subjective judgments to develop multiple-criteria evaluation models (Afshar et al. 2017; Lam et al. 2009; Nasab and Ghamsarian 2015). This reliance raises concerns about transparency (Tran et al. 2017); judgments of different evaluators on the same contractor may vary significantly (Lam et al. 2001). There is a need for more objective approaches to contractor evaluations to complement the existing systems.

Another critical task of a project's owner is establishing a reasonable contract time for the project (Echeverry et al. 1991; Son et al. 2019). SHAs use the bar chart method or the critical path method (CPM) for this task, including two main time-consuming subtasks: 1) the estimation of production rates of controlling activities and 2) the sequencing of construction activities (Taylor et al. 2017). This process can also be subjective because schedulers usually use their judgments and experiences in previous

projects in schedule development. In efforts to determine realistic and defensible contract time, SHAs have developed data-driven models for estimating production rates using historical project data, such as DWRs (Jang et al. 2019; Jeong et al. 2019). Construction sequencing, on the other hand, still relies on subjective judgments (Taylor et al. 2017). This dependence poses concerns regarding knowledge retention because experienced schedulers may leave with all knowledge gained over the years without transferring to their successors. Furthermore, experienced schedulers are not always available, especially in decentralized SHAs with novices or less experienced schedulers likely taking charge of contract time estimation.

Some SHAs developed logic templates for common project work types to support construction sequencing (Bruce et al. 2012; Jeong et al. 2009; Taylor et al. 2017), but these expert-based templates are static and quickly outdated. Analysis of as-built construction data can reveal construction sequence patterns adopted by contractors in previous projects, thereby supporting the sequencing of a new project with data-back evidence and bridging the gap between schedules developed by project owners and actual as-built schedules.

Compared to contract time estimation, SHAs are more mature in utilizing historical data for construction cost estimation. For conceptual estimating in the planning phase, SHAs typically use simple parametric methods such as applying cost per parameter (e.g., dollars per centerline mile or square foot of bridge deck) of past similar projects for their estimation due to minimally available project information (AASHTO 2013; MnDOT 2008). Numerous research studies have also developed advanced

artificial intelligence-based parametric models to predict total project construction cost to improve accuracy and overcome the lack of work item-level information in the planning phase (Elmousalami 2020; Gardner et al. 2017).

As projects progress, work item-level information becomes available, and SHAs commonly use the historical bid-based method to estimate work items' costs. Unlike the final design phase, estimators in the scoping phase (about 10% to 30% of project definition completed) do not have enough details to estimate all project work items (AASHTO 2013). They typically focus on high-cost impact work items, as suggested by the Pareto principle, that approximately 20% of the work items comprise 80% of a project's total cost (PennDOT 2018; TxDOT n.d.). A percentage or minor item allowance is used to consider the remaining work items (ConnDOT 2017; ITD 2020). Major work items and their contribution to total project cost vary with project work types and work-item breakdown structures (PennDOT 2018). Nevertheless, SHAs have limited guidelines and rely on estimators' judgments to select major work items and determine minor item allowances. Also, as numerous sets of major items can be used for estimation, selecting an optimal set is desirable to minimize the set size, maximize its cost percentage over total project cost, and minimize the uncertainty associated with using only a percentage to consider all minor work items.

In summary, there is a need for data-driven approaches to enhance cost and time-related decision-making by project owners such as SHAs in different project development phases. Specifically, this study aims to address the following research questions:

**Question 1:** How can production rates be calculated from historical DWR data and then used to evaluate contractors' past production performance considering the existence of numerous factors influencing production rates?

**Question 2:** How to extract common sequential patterns, such as pairwise logical relationships, between construction activities in past projects and determine the confidence level of applying a discovered pattern for new projects?

**Question 3:** How to automatically suggest immediate successors and predecessors of a construction activity or sequence any set of activities, including those whose activities did not frequently occur together in past projects?

**Question 4:** How to find optimal major work items for cost estimation in the scoping phase considering project types, work-item breakdown structures, and multiple objectives?

### **1.3. Research objectives**

This research's primary goal is to develop data-driven approaches and frameworks to enhance decision making by project owners before construction, alleviate the reliance on subjective judgments, and enable mechanisms for continuous improvements by utilizing historical data for the decision-making of new projects. The following are specific objectives:

**Objective 1:** Develop a data-driven framework that allows project owners to evaluate contractors more objectively and comprehensively, compared with the existing systems, by considering an additional but critical criterion, contractors' past production performance.

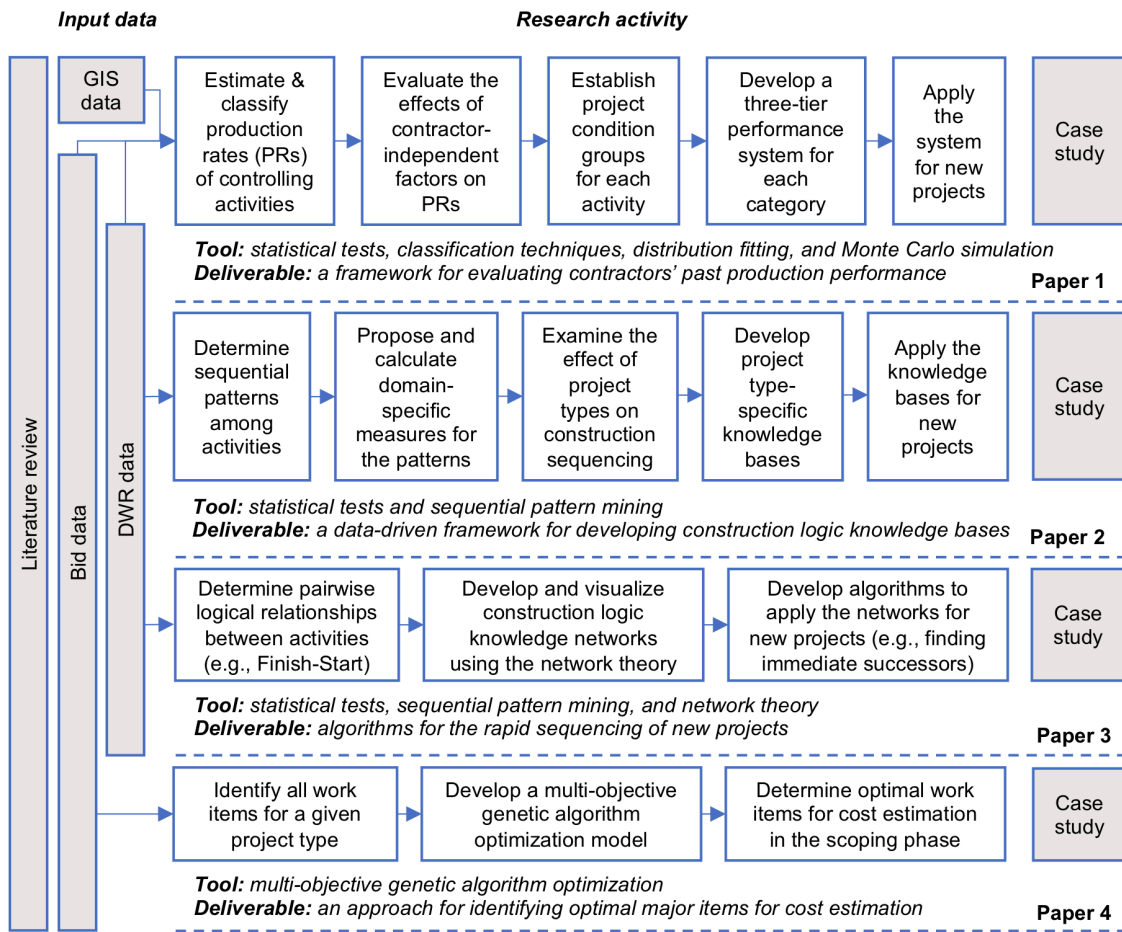
**Objective 2:** Develop a data-driven framework for the automated creation of construction logic knowledge bases under different project conditions from DWR data, including sequential patterns among activities and measures proposed to evaluate and apply the patterns for new projects.

**Objective 3:** Develop algorithms that can automatically suggest the immediate predecessors and successors of an activity and the sequence of a given set of activities while eliminating redundant logical relationships.

**Objective 4:** Develop a multi-objective optimization approach that can automatically find optimal major work items and definitive information for their application to scoping-phase cost estimation for different project work types and work-item breakdown structures.

#### **1.4. Research methodology**

Fig. 1.2 presents the overall research approach. The process is started with a literature review of the current body of knowledge to identify knowledge gaps, formulate research questions and objectives, and determine methods, techniques, and approaches to achieve the objectives.



**Fig. 1.2.** Overall research methodology

Regarding the first objective, contractors' production rates in past projects are calculated from DWR data and classified into three tiers, i.e., Tier 1—high performance, Tier 2—medium performance, and Tier 3—low performance, using the equal frequency interval method, distribution fitting, and Monte Carlo simulation. These production rates and classifications can be used for production rate estimation but are not adequate for contractor performance evaluation. Ideally, all contractor-independent factors influencing production rates need to be considered to allow for a fair evaluation of contractor performance. Due to data constraints and practical application purposes, the

effects of four common influential factors, i.e., project location, project budget, weather, and quantity of work, on production rates are evaluated using various statistical tests. These four factors constitute different project condition groups, and only production rates in the same group are classified into three performance tiers for contractor evaluation and comparison. Activity- and project-level scores are proposed to apply the evaluation system for new projects.

One of the main steps toward the second objective is to apply sequential pattern mining algorithms to historical DWR data to obtain sequential patterns among construction activities. Domain-specific measures, such as sequencing confidence, are proposed to evaluate the usability of the discovered patterns for future projects. Statistical tests are employed to evaluate the effect of project work types on construction sequencing. The discovered patterns for a project work type and their measures together form a construction logic knowledge base. If a set of activities involves more than one sequential pattern, statistical tests can be used to suggest the most probable one.

The sequential pattern mining-driven knowledge bases, however, have limitations. First, they do not support the determination of the immediate successors and predecessors of an activity. Second, they only include patterns whose activities frequently occurred together in past projects. Third, the patterns offer significant overlapping information. The third objective focuses on addressing these disadvantages. The network theory is applied to visualize and connect the separated discovered patterns, especially pairwise logical relationships between activities, to build a visual knowledge



network for each project work type. Algorithms are designed to support the rapid application of the developed network for the sequencing of new projects.

The final deliverable is a multi-criteria optimization model that can find optimal major work items for cost estimating in the scoping phase. An analysis of historical bid data from an SHA can provide a list of all work items associated with each project work type and specific to the agency's work breakdown structure. Optimization algorithms are then employed to select optimal major items that satisfy pre-defined criteria.

### **1.5. Expected contribution**

This study consists of four papers that aim to achieve each of the objectives mentioned above (see Fig. 1.2). The expected contributions of each paper are as follows.

**Paper #1**: Evaluating contractors' production performance in highway projects using historical daily work report data

- A novel data-driven approach that allows for the objective evaluation of contractors' past production performance using historical DWR data;
- Actual production rates and their statistical measures (i.e., mean and quartiles) for controlling activities;
- Validation of the effects of common contractor-independent influential factors on production rates; and
- A practical approach to enhance SHAs' current practices of evaluating contractors' qualifications.

**Paper #2**: A sequential pattern mining driven framework for developing construction logic knowledge bases

- A novel data-driven approach that allows for the automated creation of a knowledge base of construction sequence patterns (e.g., Start-Start and Finish-Start relationships) under different project conditions;
- Certainty level associated with each pattern;
- A formal method to evaluate the effect of an influential factor such as project work types on construction sequencing; and
- A practical approach to enhance SHAs' construction sequencing practices.

**Paper #3:** Network theory-driven construction logic knowledge network: process modeling and application in highway projects

- A novel application of the network theory to visualize, interlink, and store pairwise logical relationships extracted from DWR data;
- Algorithms for determining immediate successors and predecessors of an activity from DWR data; and
- An algorithm for sequencing activities even if they did not frequently occur together in past projects while eliminating redundant relationships.

**Paper #4:** Pareto principle in scoping-phase cost estimating: a multi-objective optimization approach for selecting and evaluating optimal major work items

- A novel application of multi-criteria optimization to find optimal major work items for cost estimation in the scoping phase.
- New knowledge about optimal major work items, their contribution to total project cost and variation, and the Pareto principle approach's error.

- A practical approach to enhance SHAs' scoping-phase cost estimation practices.

## **1.6. Dissertation organization**

This dissertation is organized into six chapters. Chapter 1 presents the research background and motivation, problem statement, objectives, methodologies, and expected contributions. Papers #1 to #4 are presented in Chapters 2 to 5, respectively. The final chapter, chapter 6, concludes the dissertation with major findings.

## **1.7. References**

AASHTO (2013). "Practical guide to cost estimating." AASHTO Washington, DC.

Afshar, M. R., Alipouri, Y., Sebt, M. H., and Chan, W. T. (2017). "A type-2 fuzzy set model for contractor prequalification." *Automation in Construction*, 84, 356-366.

Anderson, S. D., Molenaar, K. R., and Schexnayder, C. J. (2007). *Guidance for cost estimation and management for highway projects during planning, programming, and preconstruction*, Transportation Research Board.

Bruce, R. D., Slattery, D. K., Slattery, K. T., and McCandless, D. (2012). "An Expert Systems Approach to Highway Construction Scheduling." *Technology Interface International Journal*, 13(1), 21-28.

Carrillo, P. (2005). "Lessons learned practices in the engineering, procurement and construction sector." *Engineering, construction and architectural management*, 12(3), 236-250.

- Chini, A., Ptschelinzew, L., Minchin, R. E., Zhang, Y., and Shah, D. (2018). "Industry Attitudes toward Alternative Contracting for Highway Construction in Florida." *Journal of Management in Engineering*, 34(2), 04017055.
- ConnDOT (2017). "Cost Estimating Guidelines." Connecticut Department of Transportation.
- Dye Management Group (2014). "Performance-Based Contractor Prequalification as an Alternative to Performance Bonds." *Publication No. FHWA-HRT-14-034*, Federal Highway Administration, McLean, VA.
- Echeverry, D., Ibbs, C. W., and Kim, S. (1991). "Sequencing Knowledge for Construction Scheduling." *Journal of Construction Engineering and Management*, 117(1), 118-130.
- Elmousalami, H. H. (2020). "Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review." *Journal of Construction Engineering and Management*, 146(1).
- FHWA (2002). "FHWA Guide for Construction Contract Time Determination Procedures." Federal Highway Administration, Washington, D.C.
- Forcada, N., Serrat, C., Rodríguez, S., and Bortolini, R. (2017). "Communication Key Performance Indicators for Selecting Construction Project Bidders." *Journal of Management in Engineering*, 33(6).
- Gardner, B. J., Gransberg, D. D., and Rueda, J. A. (2017). "Stochastic Conceptual Cost Estimating of Highway Projects to Communicate Uncertainty Using Bootstrap

- Sampling." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3(3), 05016002.
- Garsden, B. R. (1995). "Postconstruction Evaluation." *Journal of Construction Engineering and Management*, 121(1), 37-42.
- ITD (2020). "Construction Cost Estimating Guide." Idaho Transportation Department.
- Jang, Y., Jeong, I.-B., Cho, Y. K., and Ahn, Y. (2019). "Predicting Business Failure of Construction Contractors Using Long Short-Term Memory Recurrent Neural Network." *Journal of Construction Engineering and Management*, 145(11).
- Jeong, H. D., Le, C., and Devaguptapu, V. (2019). "Effective Production Rate Estimation Using Construction Daily Work Report Data." Montana. Dept. of Transportation. Research Programs.
- Jeong, H. S., Atreya, S., Oberlender, G. D., and Chung, B. (2009). "Automated contract time determination system for highway projects." *Automation in Construction*, 18(7), 957-965.
- Kärnä, S., and Junnonen, J.-M. "Project feedback as a tool for learning." *Proc., ANNUAL CONFERENCE OF THE INTERNATIONAL GROUP FOR LEAN CONSTRUCTION*, 47-55.
- Lam, K. C., Hu, T., Ng, S. T., Skitmore, M., and Cheung, S. O. (2001). "A fuzzy neural network approach for contractor prequalification." *Construction Management and Economics*, 19(2), 175-188.
- Lam, K. C., Palaneeswaran, E., and Yu, C.-y. (2009). "A support vector machine model for contractor prequalification." *Automation in Construction*, 18(3), 321-329.

- Le, C., Yaw, M. W., Jeong, H. D., and Choi, K. (2021). "Comprehensive Evaluation of Influential Factors on Public Roadway Project Contract Time." *Journal of Management in Engineering*, 37(5), 04021044.
- MnDOT (2008). "Cost Estimation and Cost Management - Technical Reference Manual." Minnesota Department of Transportation.
- Nasab, H. H., and Ghamsarian, M. M. (2015). "A fuzzy multiple-criteria decision-making model for contractor prequalification." *Journal of Decision Systems*, 24(4), 433-448.
- PennDOT (2018). "Estimating Manual." Pennsylvania Department of Transportation.
- Pesek, A. E., Smithwick, J. B., Saseendran, A., and Sullivan, K. T. (2019). "Information Asymmetry on Heavy Civil Projects: Deficiency Identification by Contractors and Owners." *Journal of Management in Engineering*, 35(4).
- Shrestha, K. J., and Jeong, H. D. (2017). "Computational algorithm to automate as-built schedule development using digital daily work reports." *Automation in Construction*, 84, 315-322.
- Son, J., Khwaja, N., and Milligan Duane, S. (2019). "Planning-Phase Estimation of Construction Time for a Large Portfolio of Highway Projects." *Journal of Construction Engineering and Management*, 145(4), 04019018.
- Tang, T., and McHale, G. (2016). "Big Data."  
<<https://www.fhwa.dot.gov/publications/publicroads/16sepoct/06.cfm>>. (May 24th, 2019).

- Taylor, T. R. B., Sturgill, R. E., and Li, Y. (2017). *Practices for Establishing Contract Completion Dates for Highway Projects*.
- Tran, D. Q., Molenaar, K. R., and Kolli, B. (2017). "Implementation of best-value procurement for highway design and construction in the USA." *Engineering, Construction and Architectural Management*, 24(5), 774-787.
- TxDOT (2018). "Contract Time Determination Guidance." Texas Department of Transportation, Texas.
- TxDOT (n.d.). "Risk-Based Construction Cost Estimating - Reference Guide." Texas Department of Transportation.

## **2. EVALUATING CONTRACTORS' PRODUCTION PERFORMANCE IN HIGHWAY PROJECTS USING HISTORICAL DAILY WORK REPORT DATA\***

### **2.1. Overview**

One of the most crucial tasks that a project owner has to undertake is choosing the most competent contractor for the project. The current body of knowledge on contractor prequalification and selection through bidding focuses on the development of multiple-criteria models using decision makers' subjective judgments. Few studies have used existing data from past projects. This study presents a data-driven approach to evaluating contractors' production performance on past highway projects, a critical but rarely considered aspect of current evaluation systems. Using historical daily work report data, actual production rates were calculated for controlling activities in past projects. These rates were classified into three tiers of contractor production performance; four contractor-independent factors, which influence rates, location, project budget, weather, and quantity of work, were considered by applying classification techniques, distribution fitting, and Monte Carlo simulation. Performance indexes were proposed in order to enable comparisons. The proposed framework allows project owners to evaluate contractors more objectively and comprehensively by

---

\* Reprinted with permission (from ASCE) from "Evaluating Contractors' Production Performance in Highway Projects Using Historical Daily Work Report Data" by Le, C., Jeong, H. D., Le, T., and Kang, Y., 2020. *Journal of Management in Engineering*, 36(3), 04020015.



considering an additional but critical criterion in the existing evaluation process, that is, contractors' past production performance.

## **2.2. Introduction**

Selection of the right contractor for a construction project is one of the most critical and difficult decisions the project owner can make to ensure successful completion of the project (Afshar et al. 2017; Chini et al. 2018; Forcada et al. 2017). Failure to select a qualified contractor can lead to cost overruns, schedule delay, and poor quality (Afshar et al. 2017; Awwad and Ammouy 2019). The process of selecting contractors differs for different project delivery methods. Design-bid-build (DBB), design-build (DB), and construction manager/general contractor (CM/GC) are the most widely used delivery methods by state highway agencies (SHAs) in the U.S. (Sullivan et al. 2017). While the traditional method, DBB, is still the most common for highway projects, an increasing number of SHAs use DB and CM/GC as primary alternative contracting methods (Antoine et al. 2019; Molenaar et al. 2014; Shalwani et al. 2019). For DBB projects, the low-bid approach is typically applied to select contractors, in which bid price is the main criterion for the selection (Pesek et al. 2019; Shalwani et al. 2019). However, the choice of the lowest bidder does not ensure the lowest cost at the end of the project, due to possible claims and litigation during the construction phase and potential additional costs associated with project delays and poor product quality (Chini et al. 2018; Pesek et al. 2019; Tran et al. 2017). To increase the likelihood of the successful delivery of construction projects, project owners apply a prequalification process to rule out low-performance contractors based on the evaluation of multiple

aspects (e.g., experience, past performance, and technical ability), and only prequalified contractors are allowed to submit bid prices for final selection (Afshar et al. 2017; Forcada et al. 2017). For DB and CM/GC projects, SHAs typically use best-value (BV) procurement and qualification-based selection (QBS) to procure a qualified contractor (Alleman et al. 2017; Molenaar et al. 2014; Shalwani et al. 2019). The characteristic feature of BV procurement and QBS is the consideration of various factors in selecting the winning bidder (in contrast to the consideration of bid price alone in low-bid procurement). Previous studies have identified various criteria for contractor prequalification and selection processes, such as the contractors' experience, financial capability, technical ability, and past performance evaluations (Dye Management Group 2014; Forcada et al. 2017). However, contractors' past production performance is rarely taken into consideration in the prequalification stage (for DBB projects) or the final selection process (for DB and CM/GC projects), in spite of its direct influence on the schedule performance of construction projects. The inclusion of the production performance criterion in contractor qualification can help project owners avoid selecting contractors with poor production performance; this enhances project outcomes, particularly schedule performance.

A clear advantage of multi-criteria evaluation is to give project owners more confidence in the outcome of the contractor selection process. However, the consideration of multiple factors other than price raises concerns about transparency, especially when subjective judgments of contractor evaluators involve (Elyamany and Abdelrahman 2010; Tran et al. 2017). A considerable amount of literature has been

published on developing quantitative models for contractor qualification using decision-makers' subjective judgment as input. There are two types of multi-criteria models: (1) consolidation and (2) classification. Whereas the consolidation models focus on consolidating the ratings of contractors on multiple criteria into a single measure reflecting the contractors' overall qualification (Afshar et al. 2017; El-Abbasy et al. 2013; Nasab and Ghamsarian 2015), the classification models determine whether a contractor meets the required set of qualifications for the project (e.g., qualified or disqualified) (Lam et al. 2009; Wong 2004). However, the judgments by different evaluators of the same contractor may vary significantly (Lam et al. 2001), which raises a concern in terms of the effectiveness of the models. Therefore, there is a need for more objective and rational approaches to contractor evaluation in addition to the current assessment systems. Historical project data may be a more objective resource than human judgment in assessing contractors' performance and qualifications. One example of such data is the daily work reports (DWRs) of historical projects. DWRs collect and store the field activities of a project, including the various types of work items performed, quantities of work performed, equipment and labor usage, materials used, inspection results, and significant conversations with contractors (Shrestha and Jeong 2017). Leveraging readily available data can further benefit owners by avoiding the extra expenditure of collecting information for a qualification system. However, transforming such data into insightful information is not a trivial task, as it requires the use of a new data-driven approach.

The objective of this study is to develop a data-driven performance-based contractor evaluation approach for SHAs. In particular, this study investigates the use of digital DWR data for evaluating contractors' past production performance. Because the digital DWR data are widely used by highway agencies in the United States (Shrestha et al. 2015), this DWR-based approach can be quickly adopted by SHAs to enhance their contractor qualification evaluation practices. Based on the DWR data of past projects, the actual production rates and their statistical measures were calculated. Because of the existence of factors other than contractors' performance that influence production rates, it is not fair to evaluate contractors without considering project conditions or important contractor-independent factors such as project location, project budget, and weather. Various tools and techniques were applied to validate the effects of those factors on the production rates, and then only were the production rates of projects with the same project conditions analyzed to form a three-tier classification system of contractors' production performance. A method to apply the evaluation system to a new project is also proposed.

## **2.3. Background**

### **2.3.1. Qualification criteria**

The literature on contractor prequalification and contractor selection has revealed a significant difference in the criteria used for evaluating contractors or choosing the most suitable contractor for a project (Ng and Skitmore 1999; Nieto-Morote and Ruz-Vila 2012). Apart from the distinct characteristics of each project (e.g., project type, project size, and project location) (Hatush and Skitmore 1997), Ng and Skitmore (1999)

identified two other possible factors that affect the selection of qualification criteria: (1) the project owner's objectives and (2) decision-maker perceptions. The former factor indicates that different types of owners have different objectives. Whereas public owners are accountable to the public and the government, private owners need to ensure shareholders benefits. The latter factor relates to the background and perception of the decision-makers. For example, a scheduler may focus on contractors' time performance, while a cost estimator may be more interested in their financial soundness.

Nevertheless, considerable literature exists on the identification of common criteria for contractor evaluation. In one of the early studies on the topic, Russell (1990) proposed a model for contractor prequalification, including five main criteria: "references/reputation/past performance," "financial stability," "the status of current work," "technical expertise," and "project-specific criteria." Each criterion, in turn, consists of multiple sub-criteria. For example, financial stability can be decomposed into "credit rating," "banking arrangement," and "financial statement." Subsequent studies have investigated more potential criteria, such as the contractor's organization (Holt et al. 1994), management resources (Holt et al. 1994), past experience (Holt et al. 1994), project management capabilities (Bubshait and Al-Gobali 1996), health and safety policy (Anagnostopoulos and Vavatsikos 2006), quality performance (Elyamany and Abdelrahman 2010; Nasab and Ghamsarian 2015), communication performance (Forcada et al. 2017), risk management capability (Iyer et al. 2019; Perrenoud et al. 2017), safety performance (Khalafallah et al. 2019), and sustainability criteria (Montalbán-Domingo et al. 2019).

### **2.3.2. Contractor qualification models**

Several studies have developed models that consolidate evaluations of multiple criteria into a single measure to compare contractors' capabilities. The studies have focused on identifying the weights of qualifying criteria using different data analysis techniques. The following are selected examples. The analytic hierarchy process (AHP), which allows for pairwise comparisons among criteria by professionals, was utilized by Anagnostopoulos and Vavatsikos (2006) to estimate the weights of evaluation criteria for contractor prequalification. Plebankiewicz (2009) employed fuzzy theory to represent professionals' judgments of the criteria more naturally than with the use of ordinal variables (e.g., a Likert scale from 1—"not important" to 7—"very important") by assigning some levels of uncertainty to the professionals' responses. Jaskowski et al. (2010) applied both AHP and fuzzy numbers to determine criteria weights and claimed that their proposed method outperformed the traditional AHP method. Similarly, Nasab and Ghamsarian (2015) used the fuzzy AHP method to create a contractor prequalification model, including six criteria and 22 sub-criteria. The analytic network process (ANP), an extended version of the AHP method, was deployed by El-Abbasy et al. (2013) to develop a contractor assessment model for highway projects. Afshar et al. (2017) developed a fuzzy set model to deal with the decision-makers' differences in opinion in judging contractors for prequalification.

Another type of contractor qualification models is classification models such as a binary-classification model that determines whether a contractor is qualified or not. Khosrowshahi (1999) developed a neural network model to predict whether a contractor

would pass prequalification. Lam et al. (2001) incorporated fuzzy numbers into neural network models to minimize the subjectivity of the input values for contractor prequalification. Wong (2004) developed a logistic regression model for predicting the performance of a contractor (i.e., “good contractor” or “poor contractor”) in the United Kingdom with an accuracy of 75% using an input of 31 criteria via a survey. Lam et al. (2009) applied a support vector machine to classify contractors as prequalified or not with the accuracy of over 90%.

The majority of the studies above used responses from experienced practitioners as a starting point to estimate the weights of qualifying criteria or asked professionals to evaluate contractors using ordinal or interval variables (e.g., a Likert scale ranging from 1 to 5). The assessment of most qualifying criteria was subjective, except for several criteria such as financial soundness, which was evaluated via financial statements (Hatush and Skitmore 1998).

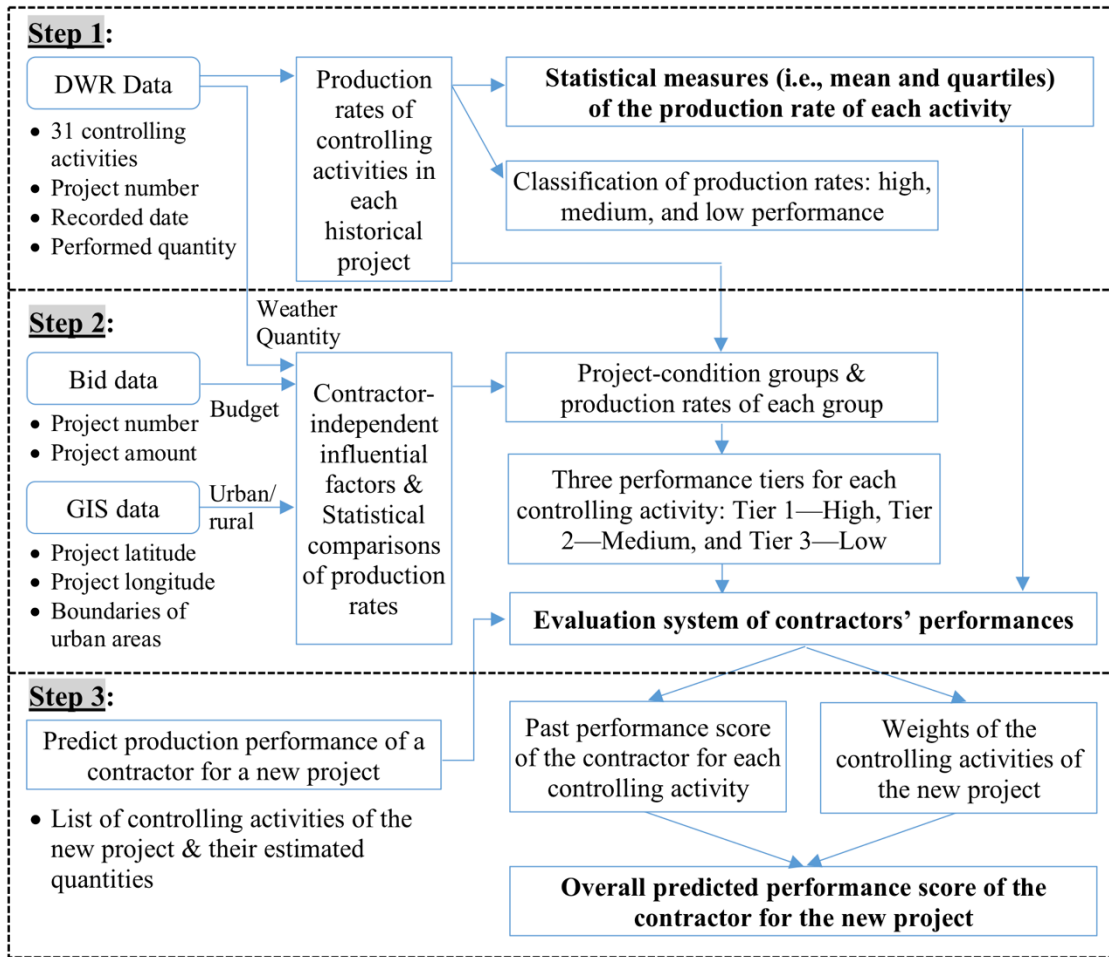
Most SHAs in the United States implement a questionnaire-based qualification system that entails a high level of subjectivity (Minchin and Smith 2001). In an attempt to obtain more objective evaluations, Hancher and Lambert (2002) provided a more detailed description for each level of rating of each qualification question to limit evaluators’ subjectivity. For example, a contractor would be evaluated at Level 3 if its work “met project requirements, but required moderate rework” or at Level 4 if the work “met project requirements, and required only minor rework” (Hancher and Lambert 2002). However, the terms “moderate” and “minor” themselves indicate some level of subjectivity, and questionnaire-based contractor qualification systems require significant

efforts from contractors and evaluators for handling qualification forms and collecting relevant information. Also, SHAs have concerns about the transparency of the contractor ratings by evaluators and desire to have more transparent procedures for the evaluation of contractors (Tran et al. 2017). One possible solution is to deploy the preexisting data of historical projects for evaluating contractors' past performance.

#### **2.4. DWR-based framework for evaluating contractors' production performance**

This study presents a DWR-based framework for evaluating contractors' production performance in highway projects. Fig. 2.1 describes an overview of the proposed framework, which includes three steps: estimation and classification of production rates, development of a contractor evaluation system, and application of the evaluation system for a new project.





**Fig. 2.1.** Overview of the proposed framework

### 2.4.1. Step 1—Estimation and classification of production rates

In the first step, the production rates of controlling activities in past projects are calculated, after which statistical measures are computed and production rates are classified into three levels.

#### 2.4.1.1. Step 1.1—Production rate estimation

Realistic and reliable production rates can be obtained from historical DWR data, which records the actual performance of contractors in the construction phase. The main

variables of DWR data are the project number, recorded date, work item code, item description, unit of measurement, performed quantity of a work item on a recorded date, contractor name, contractor identification number, weather conditions, and other variables regarding supervisors, workers, and equipment. DWRs are used by SHAs' resident construction engineers and field inspectors to document activities performed on the site by a contractor on a daily basis. For each working day, field inspectors record in their DWR system the work items that are performed on that day and their corresponding performed quantities and other relevant information (e.g., weather condition, number of workers, and number of equipment). At the end of a construction project, SHAs know, for a specific work item, the dates that the work item is performed and the corresponding quantity for each date, which can be used for determining the actual production rate of the work item for that specific project. Currently, SHAs use DWRs for payment and litigation purposes only (Shrestha and Jeong 2017).

SHAs usually identify a list of controlling activities, which are used by the agencies' schedulers to estimate projection duration. Controlling activities are the work items that are most likely to appear on the critical path of a construction project's schedule, and changes in the duration of the controlling activities influence the total project duration (Harmelink and Rowings 1998; Jeong et al. 2009). Each controlling activity is represented by one or more similar work items in the item list published by the agency. For example, topsoil salvaging and placing may be associated with two work items: (1) item code 203080100, topsoil salvaging and placing (unit: yd3), and (2) item code 203500000, topsoil salvaging and placing (unit: m3).

The production rate of a controlling activity in a past project is calculated by dividing the total performed quantity of the activity in the project by the total number of unique DWR dates recording the activity. Since an activity appears in multiple past projects, there is a sample of production rates for each activity. The sample size is the number of past projects that contain the activity. Statistical measures of the production rate of each activity are then estimated on the basis of the obtained sample. The measures include the mean, first quartile (Q1), median (Q2), and third quartile (Q3) values. The mean production rate of an activity is the average production rate of all past projects in the DWR data that included the activity. Q1, the median, and Q3 are the values for which the production rates of approximately 25%, 50%, and 75% of the projects are less than the values, respectively.

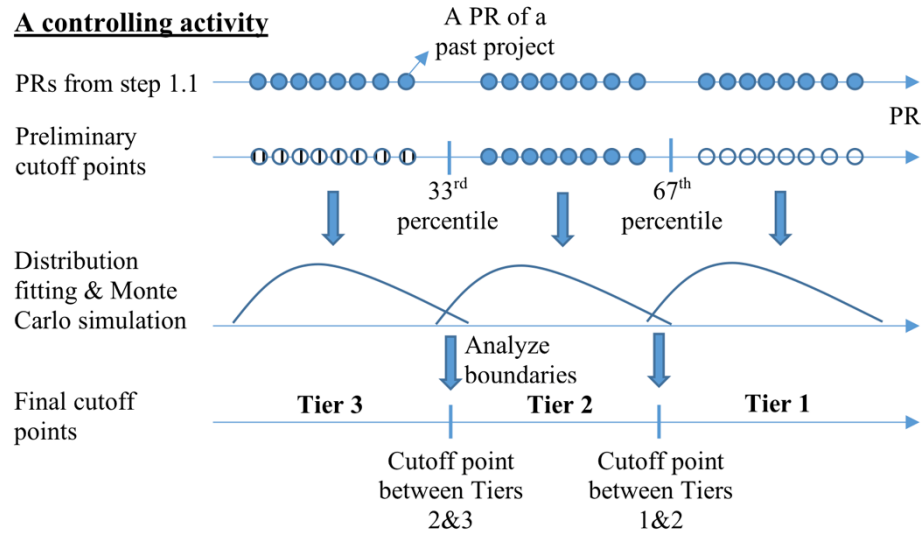
#### **2.4.1.2. Step 1.2—Classification of production rates**

The calculated production rates are reliable and realistic for estimating activity duration because they are the actual rates recorded in the DWR data. These production rates, however, may vary considerably. One of the reasons for the wide range of variation is that DWR data contain the production rates of all contractors regardless of their performance. Whereas some contractors might not have performed well in their previous projects, others might have performed very well. Production rates, therefore, can be classified into three levels: Tier 1—high performance; Tier 2—medium performance; and Tier 3—low medium. For a new project, SHAs can choose to use the medium level or the high level of production rates for their estimation, depending on the importance and the urgency of the project.

The two most popular methods for this type of classification (i.e., unsupervised classification with a known number of classes) are (1) the equal interval width method and (2) the equal frequency interval method (Dash et al. 2011; Dougherty et al. 1995). Whereas the former merely splits a range of values into smaller bins of equal width, the latter divides the value range into a number of intervals with equal numbers of values (Dash et al. 2011). Although both methods can provide a consistent rule to classify the production rates of controlling activities, the equal frequency interval method is preferred for classification because it is more robust against outliers than the other method (Dash et al. 2011). For the three-level classification of production rates, the cutoff points between classes are approximately the 33rd percentile and 67th percentile of each production rate sample. These percentiles, however, are not fixed but are expected to fluctuate when new projects are added to the sample. Due to the limited number of historical projects, additional steps need to be taken to acquire more static cutoff points.

Fig. 2.2 presents the whole process. First, the production rates of a controlling activity are divided into three groups by the 33rd and 67th percentiles, the two preliminary cutoff points. Second, each production rate group is analyzed separately to find the best-fitted distribution model for each group. Third, the Monte Carlo simulation is applied to generate a production rate distribution for each group on the basis of the fitted model identified in the previous step. Finally, the boundaries of the derived distributions are analyzed to identify the final cutoff points for classification: the cutoff point between Tier 1 and Tier 2 and the cutoff point between Tier 2 and Tier 3. In the

case of small sample sizes, which did not allow for distribution fitting, the preliminary cutoff points became the final ones.



**Fig. 2.2.** Classification of production rates

#### 2.4.2. Step 2—Development of a contractor evaluation system

The above classification of production rates can be applied to evaluate contractors' production performance. On the basis of the achieved production rates in the past projects, a contractor can be classified as Tier 1, Tier 2, or Tier 3 for each controlling activity. However, this approach is not comprehensive. The production rates of a project also depend on the characteristics of the project itself. Studies have revealed that a variety of project-specific factors other than contractors' performance contribute to the production rates, such as project location, project size, weather, the quantity of work, number of workers, and number of equipment (Jiang and Wu 2007; O'Connor et al. 2004; Woldesenbet et al. 2012). Therefore, comparing production rates alone is not an

equitable way to evaluate contractors' performance. Influential factors for production rates can be classified into two groups: contractor-independent factors (e.g., project location, project size, and weather) and contractor-dependent factors (e.g., the number of workers and equipment). Ideally, all of the contractor-independent influential factors should be considered when comparing the production rates of multiple contractors.

Contractor-independent factors constitute different project conditions that influence production rates. Only past production rates with the same project condition are analyzed to evaluate contractors. Therefore, there are multiple tier systems for a controlling activity, and each of them corresponds to a project condition. This study restricted itself to the following four major factors: project location, project budget, weather, and quantity of work. The selection of the four factors is due to the following reasons. First, the factors have to be identified at the prequalification or bidding stage as the proposed system is developed for contractor prequalification and selection. Second, extra effort to collect additional information for evaluation on the part of SHAs should be limited, because the study aims to develop a practical approach that SHAs can easily apply to their existing data in order to have another dimension of contractor performance (i.e., past production performance) for decision making. Third, the factor should be the top common ones that influence production rates, identified by SHAs (ADOT 2018; TxDOT 2020; WVDOT 2013). Last, the trade-off between the number of factors and the statistical significance of the evaluation system needs to be considered. An increase in the number of factors may help explain better the variation of production rates, but at the same time, it reduces the number of past projects per project condition as the number of

project condition increases, hence smaller sample sizes and possibly less statistical significance. Also, an extensive inclusion of influential factors in the evaluation system may not be necessary since the classification of production rates (i.e., Tier 1, Tier 2, or Tier 3), not production rates themselves, is used to compare among contractors. Not very different production rates are likely to be in the same tier, hence no difference in the performance scores. For the prequalification of low-bid procurement, the proposed system only helps to rule out contractors with low past production performance. The selection of the winning bidder from the prequalified contractors is based upon bid prices, not affected by the proposed system.

Statistical tests are used to verify the effects of the four factors (i.e., project location, project budget, weather, and quantity of work) on the production rate of each controlling activity. The factors that have a statistically significant effect on the production rate of an activity constitute different project conditions for that activity. The production rate sample of the activity obtained in Step 1 is then divided into smaller subsamples; each subsample corresponds to a project condition. For each project condition, the same procedure as Step 1.2 is applied to the corresponding subsample to find cutoff points and form three different tiers of production rates.

A past production performance score of activity  $c$ ,  $P(c)$ , is assigned to each tier. Because Tier 1 has the highest production rate, its performance score is the highest.

$$\textit{Tier 1: } P(c) = 3 \tag{2.1}$$

$$\textit{Tier 2: } P(c) = 2 \tag{2.2}$$

$$\textit{Tier 3: } P(c) = 1 \tag{2.3}$$

To quantify the past performance of a contractor in implementing a specific activity, the production rate of each project performed by the contractor is compared with the corresponding cutoff points of the same project condition to find the tier and the performance score of the contractor in each project. The performance score of a contractor in performing an activity is the average performance score of all the projects implemented by the contractor. SHAs can use the performance scores of the controlling activities to compare the past production performance among contractors or set a minimum required score that a contractor needs to pass to be qualified.

### 2.4.3. Step 3—Application of the evaluation system for a new project

To determine the overall expected performance of a contractor for a new project, the following steps are taken. First, the list of controlling activities and the corresponding quantities of the new project need to be acquired. Let assume that there are  $k$  controlling activities with indexes  $n_i$  ( $i$  varying from 1 to  $k$ ) and their estimated quantities are  $Q(n_i)$ . Second, the performance score of the contractor for each controlling activity  $n_i$ ,  $P(n_i)$ , was extracted from the evaluation system to prepare for contractor evaluation. Third, the weights of the controlling activities need to be determined. The weight  $w(n_i)$  of activity  $n_i$  is calculated on the basis of the following equation:

$$w(n_i) = \frac{Q(n_i)/APR(n_i)}{\sum_{i=1}^k Q(n_i)/APR(n_i)} \quad (2.4)$$

where  $APR(n_i)$  is the average production rate of the controlling activity  $n_i$ . The weight is proportional to the estimated duration of each activity. Lastly, the overall expected performance score of the contractor is calculated as follows:

$$P = \sum P(n_i) \times w(n_i) \quad (2.5)$$



The performance score  $P$  reflects the overall expected performance of the contractor for the new project. Because the weights range from 0 to 1 and the sum of the weights is equal to 1,  $P$  also receives values from 1 (low performance) to 3 (high performance). Similar to the individual performance scores of the controlling activities, the overall performance  $P$  can be used to compare among contractors for contractor selection or to test whether a contractor passes a predetermined threshold value to be prequalified, in the case of prequalification. Furthermore, the score  $P$  takes into consideration not only the performance of the contractor in past projects but also the specific characteristics of the new project, because every project has a different set of controlling activities and corresponding quantities.

## **2.5. Case study**

The proposed framework was applied to a case study to test whether it can generate useful information for the decision makers. Using DWR data from a SHA, this section presents how each step in the previous section was conducted.

### **2.5.1. Step 1—Estimation and classification of production rates**

#### **2.5.1.1. Step 1.1—Production rate estimation**

The authors obtained historical DWR data from a SHA: the Montana Department of Transportation. The DWR data consisted of 731 projects implemented from 2008 to 2017. On the basis of the DWR data, the authors calculated the production rates for the 31 controlling activities that were determined by the agency. Table 2.1 provides the mean, first quartile (Q1), median (Q2), and third quartile (Q3) values of the production rates of the 31 controlling activities.

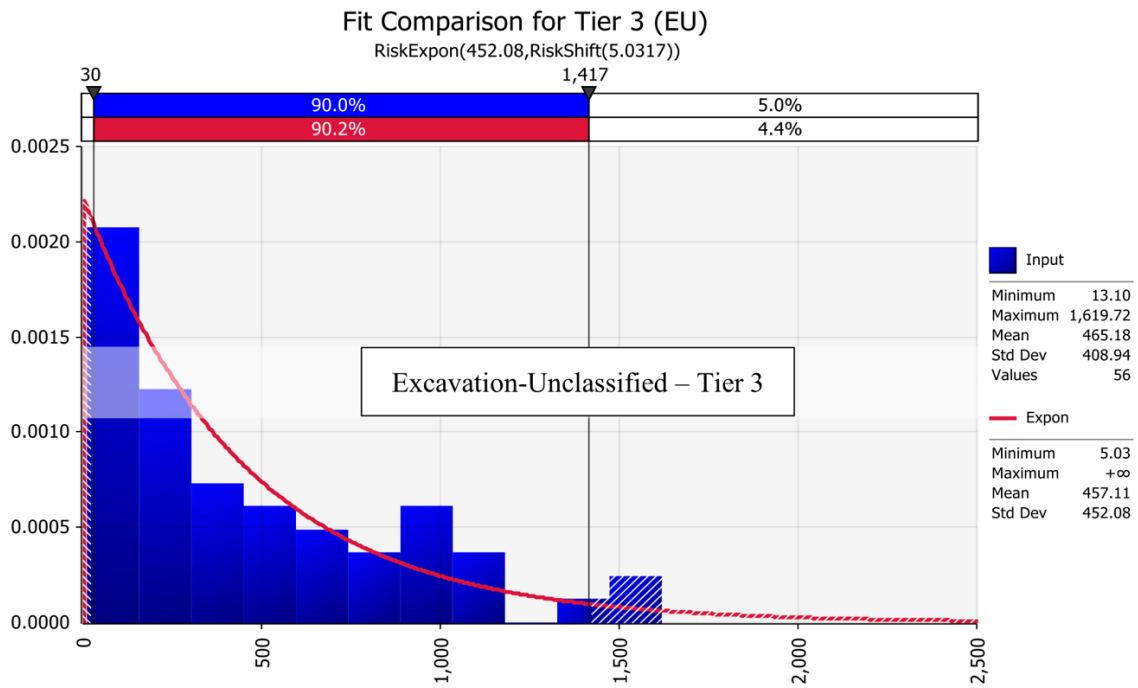
**Table 2.1.** Statistical measures of production rates from DWR data

No.	Controlling activity	Unit	Production rates (per day)			
			Mean	Q1	Median	Q3
1	Topsoil salvaging and placing	m <sup>3</sup>	1,769	234	969	2,348
2	Excavation-unclassified	m <sup>3</sup>	6,785	719	3,785	9,589
3	Special borrow	m <sup>3</sup>	2,783	493	1,382	2,836
4	Excavation-street	m <sup>3</sup>	1,161	396	748	1,855
5	Crushed aggregate course	m <sup>3</sup>	1,596	101	486	2,029
6	Base-cement treated	m <sup>3</sup>	2,640	1,185	2,726	3,857
7	Drainage pipe ( $D \leq 600$ mm)	m	29	18	26	34
8	Drainage pipe ( $D > 600$ mm)	m	28	17	22	36
9	Reinforced concrete box	m	29	15	20	43
10	Steel structural plate pipe	m	20	6	16	30
11	Riprap	m <sup>3</sup>	104	18	72	155
12	Cold milling	m <sup>2</sup>	12,60	1,598	6,873	17,117
13	Plant mix surfacing	t	1,369	343	932	1,975
14	Cover	m <sup>2</sup>	70,13	13,287	47,124	100,718
15	Micro-surfacing	t	421	370	402	477
16	Crack sealing	kg	2,878	1,362	2,327	3,630
17	Portland cement concrete pavement	m <sup>2</sup>	475	198	374	924
18	Curb and gutter	m	124	42	80	171
19	Sidewalk	m <sup>2</sup>	206	51	109	236
20	Farm fence	m	672	165	434	738
21	Guardrail steel	m	207	46	129	275
22	Concrete barrier rail	each	58	12	25	84
23	Seeding	ha	5	1	3	7
24	Reinforcing steel	kg	6,348	2,356	4,770	8,288
25	Drilled shaft	m	31	18	27	42
26	Concrete-class deck	m <sup>3</sup>	56	33	46	65
27	Class A bridge deck repair	m <sup>2</sup>	12	4	7	14
28	Concrete barrier rail bridge	m	68	27	43	77
29	Concrete-class overlay	m <sup>3</sup>	24	17	27	33
30	Bridge deck milling	m <sup>2</sup>	395	261	330	488
31	Revise bridge concrete barrier	m	61	21	48	75

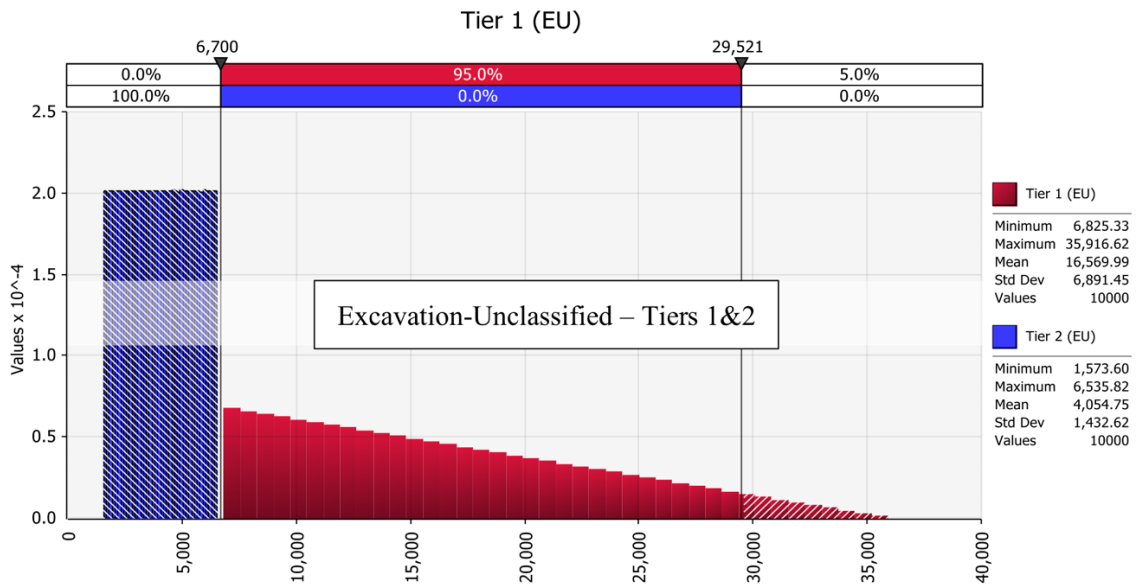
### 2.5.1.2. Step 1.2—Classification of production rates

The excavation-unclassified activity is taken as an illustrative example. The production rates of 167 projects in the DWR data including the activity were calculated.

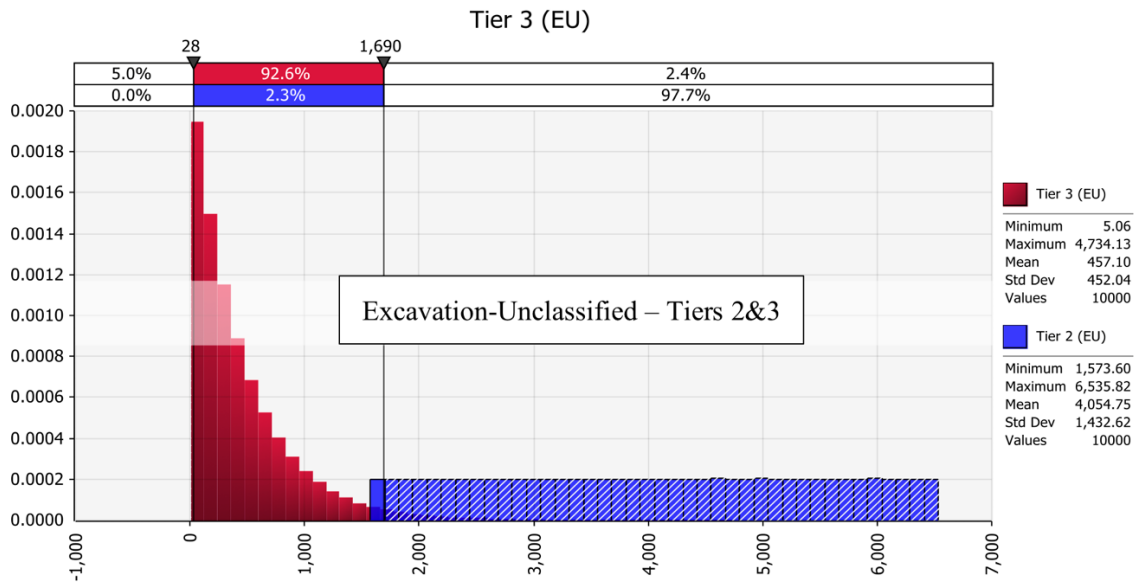
The 33rd percentile was 1,634 m<sup>3</sup>/day, and the 67th percentile was 6,699 m<sup>3</sup>/day. Production rates greater than or equal to 6,699 m<sup>3</sup>/day were preliminarily classified as Tier 1 (56 values). Similarly, production rates lower than 1,634 m<sup>3</sup>/day were assigned to Tier 3 (56 values), and the remaining values were assigned to Tier 2 (55 values). Each production rate tier was analyzed separately to identify the most fitted distribution using @Risk (version 7.6), a Microsoft Excel-based add-in tool that allows for distribution fitting and Monte Carlo simulation. Fig. 2.3 shows the result of the distribution fitting for Tier 3. The exponential distribution was identified as the most appropriate distribution for Tier 3. Similarly, uniform and triangular distributions were identified for Tier 2 and Tier 1, respectively. Once the most fitted distribution model was determined for each tier, a 10,000-iteration simulation was run to form its distribution. The distributions of two adjacent tiers were then placed next to each other to identify the final cutoff points (see Fig. 2.4 for Tier 1 and Tier 2, and Fig. 2.5 for Tier 2 and Tier 3). As shown in Fig. 2.4, a clear cutoff point between Tier 1 and Tier 2 was 6,700 m<sup>3</sup>/day, because the probability of a Tier-1 production rate higher than 6,700 m<sup>3</sup>/day was 100% and that of a Tier-2 production rate lower than 6,700 m<sup>3</sup>/day was 100%. Similarly, 1,690 m<sup>3</sup>/day was the final cutoff point between Tier 2 and Tier 3 (see Fig. 2.5).



**Fig. 2.3.** Distribution fitting for excavation-unclassified (EU)—Tier 3 (m<sup>3</sup>/day)



**Fig. 2.4.** Production rate distributions for Tiers 1 and 2, excavation-unclassified (m<sup>3</sup>/day)



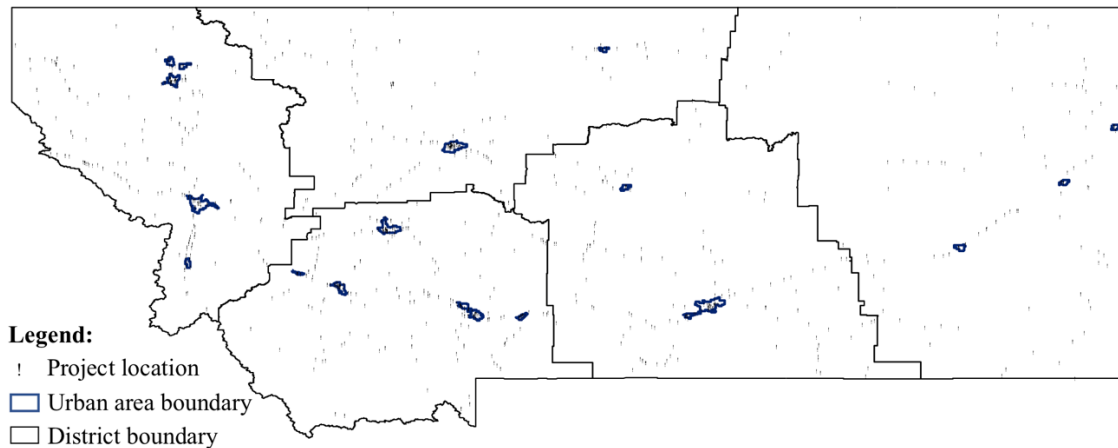
**Fig. 2.5.** Production rate distributions for Tiers 2 and 3, excavation-unclassified (m<sup>3</sup>/day)

## 2.5.2. Step 2—Development of a contractor evaluation system

Statistical tests were used to examine the effects of the four influential factors on the production rates of controlling activities.

### 2.5.2.1. Project location

The authors obtained project location data (i.e., project latitudes and longitudes) from the historical bid data via common project identification numbers with the projects in the DWR data. The projects were then mapped in ArcGIS Desktop (version 10.6.1), a geographic information system (GIS) platform dealing with spatial data. Montana currently has 19 qualifying urban and urbanized areas with a population of 5,000 or higher (MDT 2017). The boundaries of urban regions were superimposed on the project locations to identify whether a project was located in an urban or rural area. Fig. 2.6 shows the map of the project locations and urban area boundaries.



**Fig. 2.6.** Project locations and boundaries of urban areas

To validate the effect of the area type (i.e., urban or rural) on the production rates, statistical analysis was performed on all controlling activities to ensure statistical significance of the results. For each activity, the mean production rates of the two area groups (i.e., the urban group and the rural group) were calculated. Two types of statistical tests were used to compare the two groups: the two-sample t-test and the Wilcoxon rank-sum test. When two samples are from normally distributed populations or have large sample sizes (i.e.,  $n \geq 30$ ), the t-test can be applied to compare the means of the two populations (Ott and Longnecker 2015). When the conditions are not valid, the Wilcoxon rank-sum test, a nonparametric test, should be used to compare the two distributions (Ott and Longnecker 2015). Table 2.2 shows the results of six activities that have a significant difference in the production rates of the two area groups. For example, of those projects that contained crushed aggregate course, there were 225 projects in rural areas with the mean production rate of 1,940 m<sup>3</sup>/day, and there were 57 projects in urban areas with the mean production rate of 762 m<sup>3</sup>/day. Since the sample sizes of two

group were larger than 30, t-test was used to compare the difference between two groups with a p-value of smaller than 0.0001. The hypothesis that the two groups had equal mean values was rejected at the significance level of 0.05. Moreover, the mean production rates of the six activities of the rural group were significantly larger than those of the urban group.

**Table 2.2.** Means of production rates (per day) for two area groups

Activity Description	Unit	Rural		Urban		Comparison ( $\alpha = 0.05$ )		
		Mean	<i>n</i>	Mean	<i>n</i>	Test	<i>p</i> -value	Difference
Crushed aggregate course	m <sup>3</sup>	1,940	225	762	57	t-test	< 0.0001	Significant
Cold milling	m <sup>2</sup>	14,619	193	5,873	53	t-test	< 0.0001	Significant
Plant mix surfacing	t	1,584	331	700	78	t-test	< 0.0001	Significant
Cover	m <sup>2</sup>	75,974	308	32,199	61	t-test	< 0.0001	Significant
Farm fence	m	721	142	300	11	t-test <sup>a</sup>	< 0.0001	Significant
Revise bridge concrete	m	62	39	3	2	WRS	0.0492	Significant

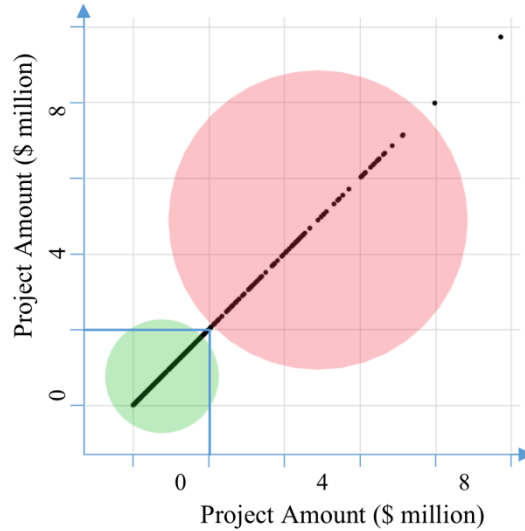
Note: *n* = number of projects in the dataset; and WRS = Wilcoxon rank-sum test.

<sup>a</sup> Although the sample size of the urban group is smaller than 30, the sample is normally distributed via normal quantile plots and normality tests.

### 2.5.2.2. Project budget

Total project cost was also extracted from the historical bid data via common project identification numbers with the projects in the DWR data. The Montana Department of Transportation (DOT), however, does not have an official rule to classify project amounts into different groups. Cluster analysis was applied to find the optimal number of categories because the two aforementioned classification methods (i.e., the equal interval width method and the equal frequency interval method) did not provide that function. Fig. 2.7 shows the results of a cluster analysis with two budget groups:

project amounts below \$4 million and project amounts equal to or higher than \$4 million.



**Fig. 2.7.** Cluster analysis of project amounts

To verify the effect of the budget type on the production rates, statistical analysis was performed on all controlling activities. As shown in Table 2.3, there were significant differences in the mean production rates for the two groups for eleven activities, such as topsoil salvaging and placing, excavation unclassified, crushed aggregate course, drainage pipe  $D \leq 600$  mm, plant mix surfacing, cover, farm fence, and seeding. Moreover, the mean production rates of the eleven activities of the larger-budget group were significantly higher than those of the smaller-budget group.



**Table 2.3.** Means of production rates (per day) for two budget groups

Activity Description	Unit	< \$4 million		≥ \$4 million		Comparison ( $\alpha = 0.05$ )		
		Mean	<i>n</i>	Mean	<i>n</i>	Test	<i>p</i> -value	Difference
Topsoil salvaging and placing	m <sup>3</sup>	971	120	3,311	73	t-test	< 0.0001	Significant
Excavation-unclassified	m <sup>3</sup>	4,391	89	10,823	67	t-test	< 0.0001	Significant
Special borrow	m <sup>3</sup>	1,674	70	4,251	57	t-test	0.0015	Significant
Crushed aggregate course	m <sup>3</sup>	899	183	3,456	90	t-test	< 0.0001	Significant
Drainage pipe (D ≤ 600 mm)	m	27	116	34	79	t-test	0.0061	Significant
Plant mix surfacing	t	1,157	290	2,169	107	t-test	< 0.0001	Significant
Cover	m <sup>2</sup>	62,660	259	79,888	97	t-test	0.0432	Significant
Curb and gutter	m	111	75	190	24	WRS	0.0024	Significant
Farm fence	m	477	87	992	64	t-test	0.0001	Significant
Seeding	ha	3	90	8	72	t-test	< 0.0001	Significant
Reinforcing steel	kg	4,203	30	8,218	42	t-test	0.0037	Significant

Note: *n* = number of projects in the dataset; and WRS = Wilcoxon rank-sum test.

### 2.5.2.3. Weather

Weather conditions can reduce production rates significantly. Production rates in the winter season, therefore, may be lower than those in the construction season (i.e., during the remaining time of the year). For each activity, the mean production rates of the two weather groups (i.e., the construction season versus the winter season) were calculated. The two-sample t-test and Wilcoxon rank-sum test were used to compare the production rates of the two weather groups. As shown in Table 2.4, there were significant differences in the production rates for the two groups for eight activities, such as excavation unclassified, drainage pipe  $D \leq 600$  mm, cold milling, cover, farm fence, and guardrail steel. Moreover, the mean production rates of the construction season group were significantly higher than those of the winter season group for the eight activities.

**Table 2.4.** Means of production rates (per day) for two weather groups

Activity Description	Unit	Construction		Winter		Comparison ( $\alpha = 0.05$ )		
		Mean	<i>n</i>	Mean	<i>n</i>	Test	<i>p</i> -value	Difference
Excavation-unclassified	m <sup>3</sup>	7,357	145	3,011	22	WRS	0.0006	Significant
Drainage pipe (D ≤ 600 mm)	m	30	187	23	24	WRS	0.0166	Significant
Cold milling	m <sup>2</sup>	13,377	251	3,812	22	WRS	0.0006	Significant
Cover	m <sup>2</sup>	74,130	373	27,593	35	t-test	< 0.0001	Significant
Farm fence	m	745	120	464	42	t-test	0.0121	Significant
Guardrail steel	m	222	192	148	50	t-test	0.0066	Significant
Reinforcing steel	kg	7,126	59	3,480	16	WRS	0.0240	Significant
Class A bridge deck repair	m <sup>2</sup>	13	49	1	2	WRS	0.0494	Significant

Note: *n* = number of projects in the data; and WRS = Wilcoxon rank-sum test.

#### 2.5.2.4. Quantity of work

The quantity of a work item can lead to an increase or a decrease in the production rate of that activity due to various reasons such as better utilization of resources and more optimized construction methods. SHAs such as Texas DOT and Virginia DOT classify the quantity of a work item into three levels (i.e., large, medium, and small) when estimating production rates. However, the details of how to distinguish the three levels are not provided. To determine a consistent rule to classify quantities of the 31 controlling activities, the equal frequency interval method was employed. For the three-level classification of quantities, the cutoff points were approximately the 33rd percentile and 67th percentile of each quantity sample. Quantities equal to or larger than the corresponding 67th percentile were considered large, whereas quantities lower than the corresponding 33rd percentile were considered small.

For each of the controlling activities, the number of historical projects in each quantity group and the mean production rates of the three quantity groups (i.e., low, medium, and high) were calculated. For three-sample comparison, the ANOVA method

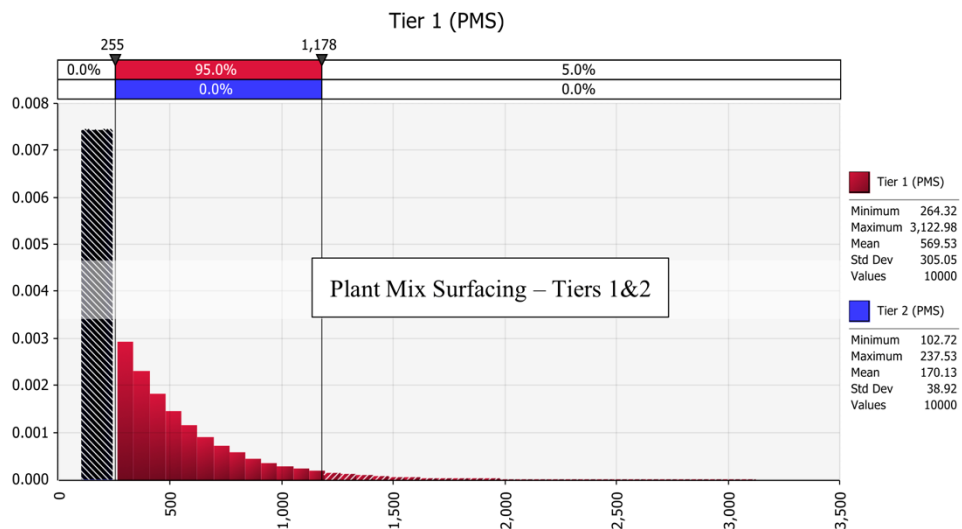
was used instead of the two-sample t-test, and the Kruskal-Wallis test replaced the Wilcoxon rank-sum test. As shown in Table 2.5, there were significant differences in the mean production rates of the three groups of 26 activities. Moreover, the high-quantity group had the highest production rates among the three groups, whereas the low-quantity group had the lowest mean.

**Table 2.5.** Comparison of production rates (per day) of three quantity groups

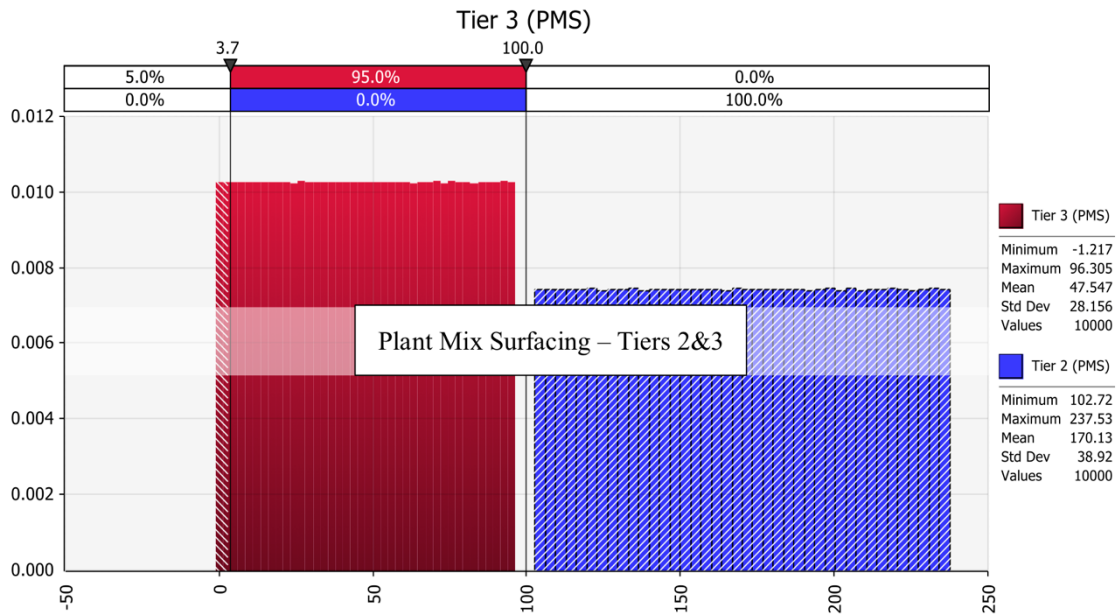
Activity Description	Unit	Low		Medium		High		Comparison ( $\alpha = 0.05$ )	
		Mean	<i>n</i>	Mean	<i>n</i>	Mean	<i>n</i>	Test	<i>p</i> -value
Topsoil salvaging and placing	m <sup>3</sup>	205	70	1,166	69	3,958	69	Anova	< 0.0001
Excavation-unclassified	m <sup>3</sup>	623	56	4,272	55	15,415	56	Anova	< 0.0001
Special borrow	m <sup>3</sup>	551	45	1,681	45	6,193	44	Anova	< 0.0001
Excavation-street	m <sup>3</sup>	253	8	1,061	8	2,168	8	KW	0.0004
Crushed aggregate course	m <sup>3</sup>	76	104	790	103	3,937	103	Anova	< 0.0001
Drainage pipe ( <i>D</i> ≤ 600 mm)	m	18	71	30	70	40	70	Anova	< 0.0001
Drainage pipe ( <i>D</i> > 600 mm)	m	19	41	30	39	34	40	Anova	0.0002
Reinforced concrete box	m	16	11	33	9	40	10	KW	0.0094
Riprap	m <sup>3</sup>	16	51	102	50	196	50	Anova	< 0.0001
Cold milling	m <sup>2</sup>	1,373	91	9,476	91	26,970	91	Anova	< 0.0001
Plant mix surfacing	t	269	151	1,512	150	2,326	151	Anova	< 0.0001
Cover	m <sup>2</sup>	10,645	136	61,075	136	138,694	136	Anova	< 0.0001
Micro-surfacing	t	331	4	410	4	524	4	KW	0.0488
Crack sealing	kg	1,465	18	3,157	17	4,097	17	KW	0.0001
Portland cement concrete pavement	m <sup>2</sup>	128	4	491	4	806	4	KW	0.0308
Curb and gutter	m	54	38	96	37	226	37	Anova	< 0.0001
Sidewalk	m <sup>2</sup>	54	40	183	39	385	39	Anova	< 0.0001
Farm fence	m	169	54	520	54	1,328	54	Anova	< 0.0001
Guardrail steel	m	38	81	167	80	415	81	Anova	< 0.0001
Concrete barrier rail	each	13	13	25	11	136	12	KW	< 0.0001
Seeding	ha	1	60	4	57	10	58	Anova	< 0.0001
Reinforcing steel	kg	1,796	25	5,029	25	12,220	25	KW	< 0.0001
Concrete-class deck	m <sup>3</sup>	34	20	48	19	84	20	KW	< 0.0001
Class A bridge deck repair	m <sup>2</sup>	3	17	9	17	25	17	KW	< 0.0001
Concrete barrier rail bridge	m	31	10	51	9	126	9	KW	0.0026
Revise bridge concrete barrier	m	17	16	75	15	93	15	KW	< 0.0001

Note: *n* = number of projects in the data; and KW = Kruskal-Wallis test.

The plant mix surfacing controlling activity was taken as an illustrative example. According to the statistical results above (Tables 2.2–2.5), project location, project budget, and quantity of work have significant effects on the production rate of plant mix surfacing. The three factors constitute different project conditions. The following gives an example of the project location that corresponds to a rural area, a project amount lower than \$4 million, and a low quantity of plant mix surfacing. A corresponding subsample of 83 values was formed to estimate cutoff points for the project condition, using the same procedure mentioned in Step 1.2. The results are shown in Fig. 2.8 and Fig. 2.9. The final values of the cutoff points were 100 t/day for Tier 2 & Tier 3 and 255 t/day for Tier 1 & Tier 2. A contractor that had a production rate of plant mix surfacing equal to or greater than 255 t/day was classified as Tier 1 for plant mix surfacing and received a corresponding performance score of 3. Similarly, a contractor that had a production rate lower than 100 t/day received a performance score of 1.



**Fig. 2.8.** Production rate distributions for Tiers 1 and 2, plant mix surfacing (PMS) (t/day)



**Fig. 2.9.** Production rate distributions for Tiers 2 and 3, plant mix surfacing (PMS) (t/day)

### 2.5.3. Step 3—Application of the evaluation system for a new project

An SHA can apply the classifications of the 31 controlling activities to determine the overall expected production performance of a contractor for a new project with the procedure outlined in the proposed framework. This expected performance can be used as an additional criterion for the agency’s current prequalification or selection system.

### 2.6. Discussion and conclusions

A significant amount of literature has focused on developing multiple-criteria models for contractor prequalification and contractor selection using decision makers’ judgments as input variables. The judgments are subjective and depend on the evaluators’ intuition and experience. These characteristics lead to variability in the ratings of different evaluators for the same contractor, thereby affecting the reliability of the developed models. In the highway sector, SHAs mostly employ questionnaire-based

systems for contractor qualification. These systems require considerable efforts from the respondents in terms of collecting supporting information, rating contractors and validating the ratings, as well as maintaining and updating the systems periodically. Nevertheless, these systems are still not free from subjectivity bias due to the nature of questionnaire methods, and hence the need for more objective approaches to contractor evaluation.

This study proposes an approach that employs DWR data for evaluating contractors' past production performance for SHAs. The actual production rates and their statistical measures (i.e., mean and quartiles) of the 31 controlling activities used by an SHA were estimated. On the basis of these production rates and the application of various tools and techniques (e.g., GIS, cluster analysis, and statistical tests), the effects of the four main contractor-independent influential factors (i.e., location, project budget, weather, and quantity of work) on production rates were validated. These four factors were used to classify projects into different project-condition groups. For each pair of controlling activity and project condition, a three-tier classification of contractors' performance was established, including Tier 1—high performance, Tier 2—medium performance, and Tier 3—low performance. Cutoff points between two adjacent tiers were determined by applying classification techniques, distribution fitting, and Monte Carlo simulation to past production rates. In addition, performance indexes for contractors, that is, performance scores for individual controlling activities and overall expected performance scores, were proposed for comparisons among contractors or against thresholds predetermined by SHAs.

The primary contribution to the body of knowledge of this study is a data-driven approach that allows for an objective evaluation of contractors' past production performance using historical DWR data. As data collection and storage technologies have today advanced substantially, using existing data for better decision making is a promising technique for achieving better project management. Indeed, the Federal Highway Administration (FHWA) is investigating various initiatives to adopt and use data-driven decision making and management (Tang and McHale 2016). Due to the increasing use of digital DWRs, SHAs can easily apply the proposed framework to enhance their current evaluation practices of contractor qualification. This DWR-based approach not only reduces human involvement in the qualification process but is also time and cost efficient. The results of contractor qualification in terms of past production performance can be achieved immediately after the required input (e.g., the contractor identification number) is provided, instead of waiting for contractors to fill out qualification forms and for qualifiers to validate the forms. Moreover, SHAs can save the expenditure of collecting data because the data are readily available. Considering the fact that large project owners tend to record DWRs and maintain their own databases (Barlow et al. 2017; Jones and Laquidara-Carr 2016), this approach should be useful for and applicable to building and industrial projects as well.

The study is not free of limitations. The first limitation is related to the data availability for each contractor. In case there is not enough information for a contractor in the DWR data for assessments, the DWR-based system cannot be applied to evaluate the contractor. In addition, when sample sizes for each contractor vary substantially,

comparing their production rates in terms of statistical measures such as means and standard deviations may be problematic. However, such situations indicate that contractors with no or few project records may not have much experience for the job in question; therefore, careful consideration is needed before awarding them contracts. The second limitation is the fact that this study did not consider all of the factors influencing the production rates. Concurrent works can be one example. If there are multiple activities that a contractor should execute concurrently, the contractor must allocate labor and equipment resources. Because the production rate is defined by quantity per day, the allocation may impact the production rates. However, to take such factors into account, SHAs need to collect additional data, which will be a burden for them. Overall, the key is to be able to use the existing database to extract meaningful information so that users can make better decisions. From this standpoint, this study makes a valuable contribution, because users can access information useful for their decision making without putting in much additional effort.

## **2.7. References**

ADOT (2018). "Production Rates Guidelines for Arizona Highway Construction."

Arizona Department of Transportation (ADOT).

Afshar, M. R., Alipouri, Y., Sebt, M. H., and Chan, W. T. (2017). "A type-2 fuzzy set model for contractor prequalification." *Automation in Construction*, 84, 356-366.

Alleman, D., Antoine, A., Gransberg, D. D., and Molenaar, K. R. (2017). "Comparison of Qualifications-Based Selection and Best-Value Procurement for Construction



- Manager–General Contractor Highway Construction." *Transportation Research Record: Journal of the Transportation Research Board*, 2630(1), 59-67.
- Anagnostopoulos, K. P., and Vavatsikos, A. P. (2006). "An AHP model for construction contractor prequalification." *Operational Research*, 6(3), 333-346.
- Antoine, A. L. C., Alleman, D., and Molenaar, K. R. (2019). "Examination of Project Duration, Project Intensity, and Timing of Cost Certainty in Highway Project Delivery Methods." *Journal of Management in Engineering*, 35(1), 04018049.
- Awwad, R., and Ammouy, M. (2019). "Owner's Perspective on Evolution of Bid Prices under Various Price-Driven Bid Selection Methods." *Journal of Computing in Civil Engineering*, 33(2).
- Barlow, G., Tubb, A., and Riley, G. (2017). "Driving business performance: Project Management Survey 2017." KPMG, NZ.
- Bubshait, A. A., and Al-Gobali, K. H. (1996). "Contractor Prequalification in Saudi Arabia." *Journal of Management in Engineering*, 12(2), 50-54.
- Chini, A., Ptschelinzew, L., Minchin, R. E., Zhang, Y., and Shah, D. (2018). "Industry Attitudes toward Alternative Contracting for Highway Construction in Florida." *Journal of Management in Engineering*, 34(2), 04017055.
- Dash, R., Paramguru, R. L., Dash, R. J. I. J. o. A. i. S., and Technology (2011). "Comparative analysis of supervised and unsupervised discretization techniques." 2(3), 29-37.

- Dougherty, J., Kohavi, R., and Sahami, M. (1995). "Supervised and unsupervised discretization of continuous features." *Machine Learning Proceedings 1995*, Elsevier, 194-202.
- Dye Management Group (2014). "Performance-Based Contractor Prequalification as an Alternative to Performance Bonds." *Publication No. FHWA-HRT-14-034*, Federal Highway Administration, McLean, VA.
- El-Abbasy, M. S., Zayed, T., Ahmed, M., Alzraiee, H., and Abouhamad, M. (2013). "Contractor Selection Model for Highway Projects Using Integrated Simulation and Analytic Network Process." *Journal of Construction Engineering and Management*, 139(7), 755-767.
- Elyamany, A., and Abdelrahman, M. (2010). "Contractor Performance Evaluation for the Best Value of Superpave Projects." *Journal of Construction Engineering and Management*, 136(5), 606-614.
- Forcada, N., Serrat, C., Rodríguez, S., and Bortolini, R. (2017). "Communication Key Performance Indicators for Selecting Construction Project Bidders." *Journal of Management in Engineering*, 33(6).
- Hancher, D. E., and Lambert, S. E. (2002). "Quality-Based Prequalification of Contractors." *Transportation Research Record: Journal of the Transportation Research Board*, 1813, 260-274.
- Harmelink, D. J., and Rowings, J. E. (1998). "Linear Scheduling Model: Development of Controlling Activity Path." *Journal of Construction Engineering and Management*, 124(4), 263-268.

- Hatush, Z., and Skitmore, M. (1997). "Criteria for contractor selection." *Construction Management and Economics*, 15(1), 19-38.
- Hatush, Z., and Skitmore, M. (1998). "Contractor selection using multicriteria utility theory: An additive model." *Building and Environment*, 33(2), 105-115.
- Holt, G. D., Olomolaiye, P. O., and Harris, F. C. (1994). "Evaluating prequalification criteria in contractor selection." *Building and Environment*, 29(4), 437-448.
- Iyer, K. C., Kumar, R., and Singh, S. P. (2019). "Understanding the role of contractor capability in risk management: a comparative case study of two similar projects." *Construction Management and Economics*, 1-16.
- Jaskowski, P., Biruk, S., and Bucon, R. (2010). "Assessing contractor selection criteria weights with fuzzy AHP method application in group decision environment." *Automation in Construction*, 19(2), 120-126.
- Jeong, H. S., Atreya, S., Oberlender, G. D., and Chung, B. (2009). "Automated contract time determination system for highway projects." *Automation in Construction*, 18(7), 957-965.
- Jiang, Y., and Wu, H. (2007). "Production Rates of Highway Construction Activities." *International Journal of Construction Education and Research*, 3(2), 81-98.
- Jones, S. A., and Laquidara-Carr, D. (2016). "SmartMarket Brief: Optimizing the Owner Organization." Dodge Data & Analytics, New York.
- Khalafallah, A., Kartam, N., and Razeq, R. A. (2019). "Bilevel Standards-Compliant Platform for Evaluating Building Contractor Safety." *Journal of Construction Engineering and Management*, 145(10).

- Khosrowshahi, F. (1999). "Neural network model for contractors' prequalification for local authority projects." *Engineering construction and architectural management*, 6(3), 315-328.
- Lam, K. C., Hu, T., Ng, S. T., Skitmore, M., and Cheung, S. O. (2001). "A fuzzy neural network approach for contractor prequalification." *Construction Management and Economics*, 19(2), 175-188.
- Lam, K. C., Palaneeswaran, E., and Yu, C.-y. (2009). "A support vector machine model for contractor prequalification." *Automation in Construction*, 18(3), 321-329.
- MDT (2017). "MDT Urban Boundaries." Montana Department of Transportation, [http://gis-mdt.opendata.arcgis.com/datasets/ded82829e789400f9b7bcb8e35b880c6\\_1](http://gis-mdt.opendata.arcgis.com/datasets/ded82829e789400f9b7bcb8e35b880c6_1)
- Minchin, R. E., and Smith, G. R. (2001). "Quality-based performance rating of contractors for prequalification and bidding purposes." *NCHRP Web Document 38*, Transportation Research Board, Washington, D.C.
- Molenaar, K., Harper, C., and Yugar-Arias, I. (2014). "Guidebook for Selecting Alternative Contracting Methods for Roadway Projects: Project Delivery Methods, Procurement Procedures and Payment Provisions." *Transportation Pooled Fund Program Study TPF-5 (260)*, 410.
- Montalbán-Domingo, L., García-Segura, T., Amalia Sanz, M., and Pellicer, E. (2019). "Social Sustainability in Delivery and Procurement of Public Construction Contracts." *Journal of Management in Engineering*, 35(2).

- Nasab, H. H., and Ghamsarian, M. M. (2015). "A fuzzy multiple-criteria decision-making model for contractor prequalification." *Journal of Decision Systems*, 24(4), 433-448.
- Ng, S. T., and Skitmore, R. M. (1999). "Client and consultant perspectives of prequalification criteria." *Building and Environment*, 34(5), 607-621.
- Nieto-Morote, A., and Ruz-Vila, F. (2012). "A fuzzy multi-criteria decision-making model for construction contractor prequalification." *Automation in Construction*, 25, 8-19.
- O'Connor, J. T., Chong, W. K., Huh, Y., and Kuo, Y.-c. (2004). "Development of improved information for estimating construction time." Center for Transportation Research, The University of Texas at Austin, Texas.
- Ott, R. L., and Longnecker, M. (2015). *An Introduction to Statistical Methods and Data Analysis*, Cengage Learning.
- Perrenoud, A., Lines, B. C., Savicky, J., and Sullivan, K. T. (2017). "Using Best-Value Procurement to Measure the Impact of Initial Risk-Management Capability on Qualitative Construction Performance." *Journal of Management in Engineering*, 33(5).
- Pesek, A. E., Smithwick, J. B., Saseendran, A., and Sullivan, K. T. (2019). "Information Asymmetry on Heavy Civil Projects: Deficiency Identification by Contractors and Owners." *Journal of Management in Engineering*, 35(4).
- Plebankiewicz, E. (2009). "Contractor prequalification model using fuzzy sets." *Journal of Civil Engineering and Management*, 15(4), 377-385.

- Russell, J. S. (1990). "Model for Owner Prequalification of Contractors." *Journal of Management in Engineering*, 6(1), 59-75.
- Shalwani, A., Lines, B. C., and Smithwick, J. B. (2019). "Differentiation of Evaluation Criteria in Design-Build and Construction Manager at Risk Procurements." *Journal of Management in Engineering*, 35(5).
- Shrestha, K. J., and Jeong, H. D. (2017). "Computational algorithm to automate as-built schedule development using digital daily work reports." *Automation in Construction*, 84, 315-322.
- Shrestha, K. J., Jeong, H. D., and Gransberg, D. D. (2015). "Current Practices of Collecting and Utilizing Daily Work Report Data and Areas for Improvements." *Proc., 6th International Conference on Construction Engineering and Project Management*.
- Sullivan, J., Asmar Mounir, E., Chalhoub, J., and Obeid, H. (2017). "Two Decades of Performance Comparisons for Design-Build, Construction Manager at Risk, and Design-Bid-Build: Quantitative Analysis of the State of Knowledge on Project Cost, Schedule, and Quality." *Journal of Construction Engineering and Management*, 143(6), 04017009.
- Tang, T., and McHale, G. (2016). "Big Data." <https://www.fhwa.dot.gov/publications/publicroads/16sepoct/06.cfm>. (May 24th, 2019).

- Tran, D. Q., Molenaar, K. R., and Kolli, B. (2017). "Implementation of best-value procurement for highway design and construction in the USA." *Engineering, Construction and Architectural Management*, 24(5), 774-787.
- TxDOT (2020). "Construction Production Rates." Texas Department of Transportation.
- Woldesenbet, A., Jeong, H. D., and Oberlender, G. D. (2012). "Daily Work Reports–Based Production Rate Estimation for Highway Projects." *Journal of Construction Engineering and Management*, 138(4), 481-490.
- Wong, C. H. (2004). "Contractor Performance Prediction Model for the United Kingdom Construction Contractor: Study of Logistic Regression Approach." *Journal of Construction Engineering and Management*, 130(5), 691-698.
- WVDOT (2013). "Design Directive 803: Determination of Contract Completion Date." West Virginia Department of Transportation (WVDOT).

### **3. A SEQUENTIAL PATTERN MINING DRIVEN FRAMEWORK FOR DEVELOPING CONSTRUCTION LOGIC KNOWLEDGE BASES\***

#### **3.1. Overview**

One vital task of a project's owner is to determine a reliable and reasonable construction time for the project. A U.S. highway agency typically uses the bar chart or critical path method for estimating project duration, which requires the determination of construction logic. The current practice of activity sequencing is challenging, time-consuming, and heavily dependent upon the agency schedulers' knowledge and experience. Several agencies have developed templates of repetitive projects based on expert inputs to save time and support schedulers in sequencing a new project. However, these templates are deterministic, dependent on expert judgments, and get outdated quickly. This study aims to enhance the current practice by developing a data-driven approach that leverages the readily available daily work report data of past projects to develop a knowledge base of construction sequence patterns. With a novel application of sequential pattern mining, the proposed framework allows for the determination of common sequential patterns among work items and proposed domain measures such as the confidence level of applying a pattern for future projects under different project

---

\* Reprinted with permission (from Elsevier) from "A sequential pattern mining driven framework for developing construction logic knowledge bases" by Le, C., Shrestha, K. J., Jeong, H. D., and Damnjanovic, I., 2021. *Automation in Construction*, 121, 103439.



conditions. The framework also allows for the extraction of only relevant sequential patterns for future construction time estimation.

### **3.2. Introduction**

Establishing a reliable and reasonable project duration is one of the most vital tasks of a project owner before construction since it has a critical impact on the successful completion of a project (Echeverry et al. 1991; Son et al. 2019). State Departments of Transportation (DOTs) in the U.S. currently use the bar chart and the critical path method (CPM) to develop schedules of simple and complex projects, respectively (Taylor et al. 2017). However, developing a realistic schedule is time-consuming and challenging for both new and experienced schedulers (Fischer and Aalami 1996; Shrestha et al. 2019). The process of developing an as-planned schedule before construction usually involves three main tasks: 1) the identification of construction activities, 2) the estimation of activity production rates and durations, and 3) the determination of construction logic among activities (Mubarak 2015). Of the three tasks, the first one is the simplest because design plans and a list of work items and quantities are available to schedulers at the end of the design phase. DOTs have guidance and tools to support the second task (Leandro et al. 2018), and several studies have been conducted to improve production rate estimation processes (Le and Jeong 2020; Le et al. 2020; Woldesenbet et al. 2012). However, there are few data-driven studies dedicated to the third task. Prior studies were heavily dependent on experienced schedulers' knowledge and experience for construction sequencing (Bruce et al. 2012; Jeong et al. 2009). Also, DOTs have limited guidance about sequencing construction

activities and rely on schedulers' expertise or schedule templates derived from expert opinions (Taylor et al. 2017). This dependence on individuals is not a sustainable solution since active knowledge retention programs are not practiced in many agencies (Taylor et al. 2017). The retirement of experienced schedulers or their move to the private sector basically translates into the retirement of the knowledge and experience in the current business environment. Schedule templates are a form of knowledge retention, but they are static and get outdated quickly when construction means, methods, and conditions change (Shrestha et al. 2019). Thus, there is a need for a data-driven approach to construction sequencing.

DOTs have invested a significant amount of time, money, and effort in collecting the digital data of highway projects (Tang and McHale 2016). However, the collected data, such as digital Daily Work Reports (DWRs), have not been fully leveraged to enhance current business practices and increase the return on investment (Shrestha and Jeong 2017). A national survey (Jeong et al. 2015) revealed that 37 DOTs used an electronic DWR system (e.g., AASHTOWare SiteManager) for their projects. However, these systems are mainly used for monitoring progress, making payments to contractors, and resolving possible claims (Shrestha and Jeong 2017). Some studies have applied DWR systems for as-built schedule development, production rate estimation, and contractors' past production performance evaluation (Jeong et al. 2019; Le et al. 2020; Shrestha and Jeong 2017). However, most DOTs do not capture the lessons learned during construction to enhance future project planning and scheduling (Taylor et al. 2017).

The main goal of this research is to innovate the current scheduling practices of the owner agencies by developing a knowledge base of various types and patterns of construction activity sequencing. An easily accessible knowledge base driven from past construction projects would work as a robust and reliable resource, which may offer and suggest the most probable relationships among key work activities for a new project based on historical as-built data of similar types of projects. It may also help reduce the dependency on expert-based resources such as logic templates. A set of proposed measures associated with each pattern also provides schedulers with a multi-perspective evaluation of the pattern. Although schedulers may modify the proposed patterns in some cases due to the unique characteristics of a new project, the proposed patterns may provide significant support and confidence to schedulers, especially inexperienced ones, than DOTs' current limited sequencing resources, thereby saving time and improving productivity. Additionally, the knowledge base is not static and can be updated as newer DWR data becomes available.

### **3.3. Literature review**

#### **3.3.1. Highway scheduling practices**

Two major approaches for estimating project duration are a) a bottom-up approach using production rates/durations of construction activities and sequential logic among the activities to formulate construction schedules and b) a top-down approach using prediction models built upon past project data (Stephenson et al. 2010).

A national survey of U.S. highway agencies reveals that bottom-up methods such as the bar chart and CPM are dominant for project duration estimation (Taylor et al.

2017). A limited number of DOTs have applied top-down methods for their projects (Taylor et al. 2017). Ohio DOT developed a multiple linear regression (MLR) model to estimate preliminary construction duration using construction cost, work type, location, working season, and others as predictors (Ohio DOT 2013). Similarly, the Kentucky Transportation Cabinet (KYTC) developed an MLR based estimation model for small projects (i.e., less than \$1 million) and another MLR model for large projects (Zhai et al. 2016). Also, Colorado DOT is developing an artificial neural network model for establishing construction time using project size, project type, estimated construction cost, and bid item quantities as input variables.

The top-down methods have a clear advantage of fast estimation time compared to the bottom-up approach (Zhai et al. 2016). However, the use of the top-down methods is limited to early project development phases when design and other detailed project information are not available for applying the bottom-up approach (Son et al. 2019; Stephenson et al. 2010). In later phases (e.g., final design and procurement), the bar chart method and CPM are the primary scheduling tools in establishing construction time due to their ability to leverage detailed project information and consider unique project characteristics (Stephenson et al. 2010).

Existing top-down models have only considered the impact of general project characteristics (e.g., project location and project type) and significant activity quantities on construction time (Ohio DOT 2013; Zhai et al. 2016). These explanatory variables may not be enough to produce a reliable estimate due to various other influential factors, such as project phasing, maintenance of traffic, environmental restrictions, adverse

weather, and working time restrictions (Taylor et al. 2017). Adding more predictors into prediction models is theoretically possible but practically constrained by the availability of relevant data.

Due to the above reasons, this study focuses on the bottom-up approach. Of the two main components of the approach, i.e., activity duration estimation and construction sequencing determination, various research findings and guidance are available for the former (Jeong et al. 2019; Le and Jeong 2020; Leandro et al. 2018). At the same time, few studies have investigated the utilization of historical data for sequencing construction activities.

### **3.3.2. Prior studies of construction sequencing**

A large and growing body of literature has investigated construction sequencing as a primary component of trending network-based scheduling topics (e.g., automated schedule development and schedule optimization) and as a separate research topic itself (Bruce et al. 2012; Fan et al. 2012; Jeong et al. 2009; Kim et al. 2013). This review was not limited to construction sequencing-focused studies but also included other relevant scheduling research to synthesize how previous studies determined constructions sequences or developed precedence networks.

Various studies have attempted to transform the knowledge and experience of scheduling experts into written forms such as sequencing rules and templates to be used by other schedulers. For example, Echeverry et al. (1991) described four essential factors considered by skilled schedulers for sequencing. They were physical relationships among building components (e.g., paint-covered walls), trade interactions (e.g., resource

limitations), optimized movement of equipment and materials, and safety considerations. In recognition of some level of repetition of construction schedules of a project type, Chevallier and Russell (2001) proposed a partially automated approach to schedule development by capturing experience from the past projects in knowledge-based templates. Each template contains the typical activities of a recurring project type and sequencing relationships between the activities. Similar studies were conducted to develop scheduling templates for DOTs based on interviews with DOT scheduling experts and reviews of past project records (Bruce et al. 2012; Jeong et al. 2009). Some studies also attempted to automate sequencing of activities from expert inputs such as functional dependencies among walls, columns, and beams (Chua et al. 2013) and predetermined relationships among cutting, fitting, and welding operations of pipe spool fabrication (Hu and Mohamed 2014). Apart from traditional logical relationships (e.g., Start-Start or Finish-Start), new logical relationships (e.g., maximal, point-to-point, and continuous) have been proposed to describe better the interdependencies between construction activities (Hajdu 2015; Hajdu 2018).

Two major scheduling research topics involving construction sequencing are automated schedule development and schedule optimization. Regarding the first topic, previous studies leveraged activity information embedded in computer-aided-design (CAD) drawings (Cherneff et al. 1991; Fischer and Aalami 1996) and, more recently, in building information modeling (BIM) models (Kim et al. 2013; Liu et al. 2015; Wang et al. 2014) to automate the development of construction schedules. However, CAD drawings and BIM models do not typically provide information about construction logic.

These studies have mostly relied on a predetermined precedence network (Wang et al. 2014), expert-based sequencing rules (Cherneff et al. 1991; Fischer and Aalami 1996; Kim et al. 2013), or a predetermined process pattern (Liu et al. 2015) for sequencing construction activities. An example of sequencing rules is that if A supports B, install A before B and remove B before A (Kim et al. 2013). A sample process pattern for cast-in-place building elements is “Erect Inner Formwork -> Install Rebar -> Erect External Formwork -> Pour Concrete -> Cure Concrete -> Remove Formwork” (Liu et al. 2015). These expert-based resources have also supported schedule optimization research (Fan et al. 2012; Florez 2017; Jaśkowski and Sobotka 2006; Lim et al. 2014). Based on a basic sequence among activities, researchers have considered changes in other influential components and factors to optimize project outputs. Some examples considered are different duration and resources to complete an activity (Jaśkowski and Sobotka 2006), the maximum number of concurrent units for an activity (Fan et al. 2012), overlap levels between two activities (Lim et al. 2014), and crew allocation (Florez 2017).

One common feature of the above studies is the heavy dependence on senior schedulers’ knowledge and experience. Thus, an effective methodology to extract sequencing knowledge from historical construction data, such as DWR data, is highly desirable. This innovative approach may be able to support schedulers, especially junior schedulers, in determining realistic sequences of construction activities for new projects.

### **3.4. Research objective and scope**

This study’s objective is to better support DOT schedulers in sequencing construction activities for a new project by using an advanced data driven approach to

the historical project performance data. A contractor conducts sequencing in the construction phase with detailed information on how to build the project (e.g., phasing, crew allocation, construction methods, and logistics) and input from key project team members such as project managers, superintendents, and subcontractors (Hajdu 1996; Newitt 2009; Pierce 2014). With that detailed information and inputs, sequencing logic development is still an iterative process evolving during construction due to unique project characteristics and other constraints on the sequence such as safety, quality, and resources (Hinze 2012; Newitt 2009; Pierce 2014). It is also more art than science because there is “no one right way to build any project” (Mubarak 2015; Newitt 2009). Those issues create a variety of possible sequencing solutions for a given project. Unlike schedule development by a contractor, DOT schedulers do not have that detailed information and inputs from the contractor. They need to make assumptions about construction techniques and phasing options to develop a reasonable sequencing logic and a reliable project duration estimate that will be included in the request for bid (CDOT 2019). The logic is not likely to be the same as the later developed sequence by the winning bidder, let alone the final actual sequence at the end of construction. An analysis of as-built schedules can help bridge the difference between owner-developed and as-built sequences.

This study proposes a framework to extract common sequence patterns adopted by contractors under different project conditions from the historical DWR data. DOT schedulers can directly apply or adjust the patterns depending on unique project characteristics to quickly develop logic networks for new projects. The framework also



provides the certainty level of each pattern as well as irregular sequential relationships between activities to illustrate the variation of activity sequencing or to point out unique scenarios that need extra attention from DOT schedulers. This data-driven approach also helps alleviate the heavy and sole dependency on experienced schedulers' experience and judgment in logic sequencing.

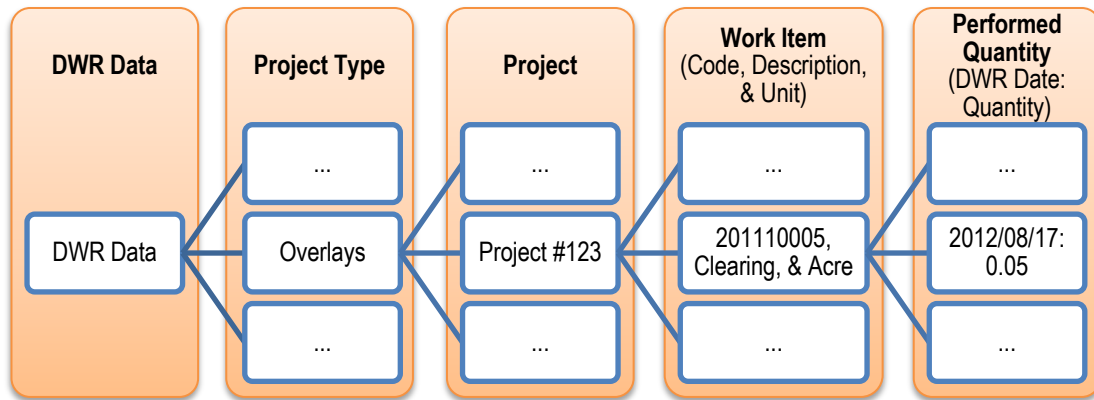
### **3.5. Methods and concepts utilized for framework development**

This section provides background on the primary concept and method utilized for framework development: DWR data and SPM algorithms.

#### **3.5.1. Daily work report data**

As the primary owners of highway projects in the U.S., DOTs assign their site inspectors to monitor construction activities performed by contractors to ensure successful project completion on a daily basis. Site inspectors collect daily construction-related information (e.g., the amount of work performed by contractors, labor, equipment, and weather conditions) in DWRs for different purposes such as monitoring progress, making payments to contractors, and resolving possible claims and disputes in the future (Jeong et al. 2015). A digital DWR system allows easy data extraction for other potential applications. One of them is to identify sequential relationships between construction activities in past projects.

Fig. 3.1 illustrates the data attributes that are typically available in a DWR system. DOTs categorize projects into different project work types. A project contains a list of work items to be performed. For a specific working day, DOT inspectors record the performed quantities of the work items.



**Fig. 3.1.** DWR data attributes

### 3.5.2. Sequential pattern mining

Pattern mining is a fundamental task of data mining that aims to discover patterns of interest from a database (Aggarwal 2015). Different types of pattern mining have been proposed and applied depending on a particular application (Fournier-Viger et al. 2017). For example, frequent itemset and association rule mining algorithms aim to detect event co-occurrences in a database without considering event orders (Fournier-Viger et al. 2017). Conversely, sequential pattern mining (SPM) is a natural choice to deal with a time-associated database. SPM analyzes a database of sequences, such as DWR data, to extract meaningful sequential patterns or subsequences of interest (Fournier-Viger et al. 2017). The applications of SPM started with customer transaction databases (e.g., retail customer transactions in a grocery store) (Agrawal and Srikant 1995) but have expanded to other domains, such as telecommunication, web access analysis, e-learning, scientific experiments, text analysis, natural disasters, DNA research, and protein formations (Chand et al. 2012; Fournier-Viger et al. 2017). Fig. 3.3 gives an example of a five-sequence database. Each sequence corresponded to a

customer's transactions in the first week of opening at a grocery store. For example, Customer #1 bought Item *a* on Day 1, Items *b* and *c* on Day 2, Item *d* on Day 4, and Item *f* on Day 6. These transactions corresponded with Sequence #1, indicating that the customer bought Item *a*, then purchased Items *b* and *c* together, then bought Item *d*, and then purchased Item *f*.

Customer	Transactions of six items in the 1 <sup>st</sup> week							ID	Sequence
	D1	D2	D3	D4	D5	D6	D7		
1	<i>a</i>	<i>b, c</i>		<i>d</i>		<i>f</i>		1	$\langle \{a\}, \{b, c\}, \{d\}, \{f\} \rangle$
2		<i>a, c</i>		<i>f</i>	<i>d</i>		<i>e, c</i>	2	$\langle \{a, c\}, \{f\}, \{d\}, \{e, c\} \rangle$
3	<i>a, b</i>		<i>d</i>			<i>e</i>		3	$\langle \{a, b\}, \{d\}, \{e\} \rangle$
4		<i>b</i>	<i>c</i>	<i>d, e</i>				4	$\langle \{b\}, \{c\}, \{d, e\} \rangle$
5		<i>a</i>	<i>b</i>		<i>f</i>			5	$\langle \{a\}, \{b\}, \{f\} \rangle$

Note: D = Day

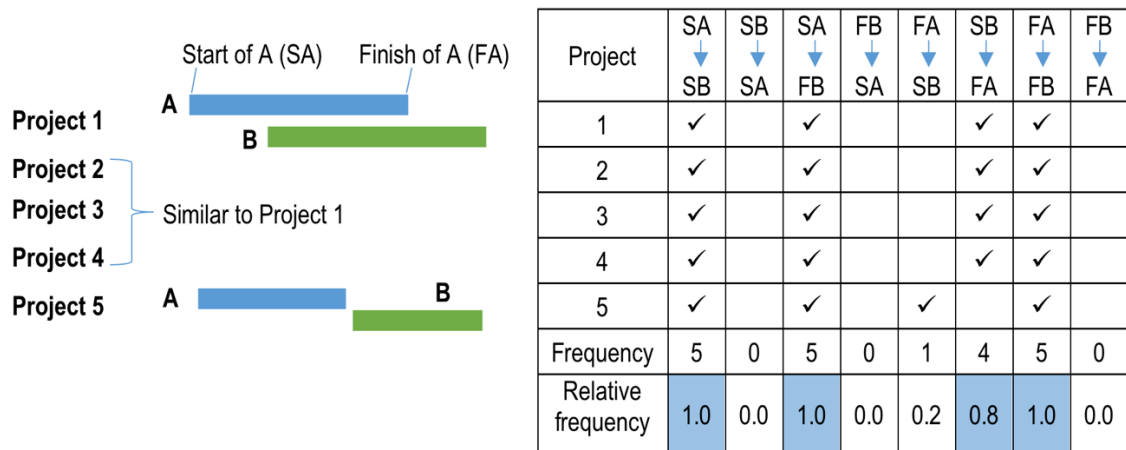
**Fig. 3.2.** An example of a sequence database

The support of a subsequence (e.g.,  $\langle \{a\}, \{b\} \rangle$ ) is the number of sequences containing the subsequence. For example, subsequence  $\langle \{a\}, \{b\} \rangle$  (i.e., customers bought *a* then *b*) has a support of 2 (IDs #1 & 5). One primary focus of SPM is to find frequent subsequences in a sequence database. A frequent subsequence is the one with support not smaller than the minimum support threshold defined by users. The most straightforward way of searching all frequent subsequences is to determine all possible subsequences and their support and then choose ones that meet the predefined criterion. However, that approach is not practical and efficient for most real-life applications since the number of possible subsequences can be huge (Fournier-Viger et al. 2017). Therefore, various SPM algorithms, such as GSP, ClaSP, PrefixSpan, and Spam, have

been developed to improve the search for desired patterns (Chand et al. 2012; Fournier-Viger et al. 2017).

### **3.5.3. Theoretical foundation of extracting construction sequential patterns from DWR data**

As-built construction data contain information about the sequential order of work items in past projects. An analysis of those sequential orders can help reveal the precedence relationships adopted in past projects. Fig. 3.3 illustrates the principles of extracting sequential patterns between two work items, namely item A and item B, from DWR data. Two events represent item A: Start of A (SA) and Finish of A (FA). Similarly, item B is represented by SB and FB. The two events of item A, two events of item B, and two sequential directions may result in at most eight pairwise relationships between A and B, such as “SA -> SB,” which is a Start-Start relationship between A and B (see Fig. 3.3). Assume that only five past projects contain both items, and their as-built orders are available on the left of Fig. 3.3. Those as-built orders can be easily transformed into a frequency table on the right of Fig. 3.3. The frequency of a relationship then demonstrates how frequently the relationship appeared in past projects. For example, item A started before the start of item B in 100% of the past projects. The analysis can provide not only one sequential relationship but multiple relationships between two items. The measures associated with each relationship will help a scheduler decide to use the relationship for a future project.

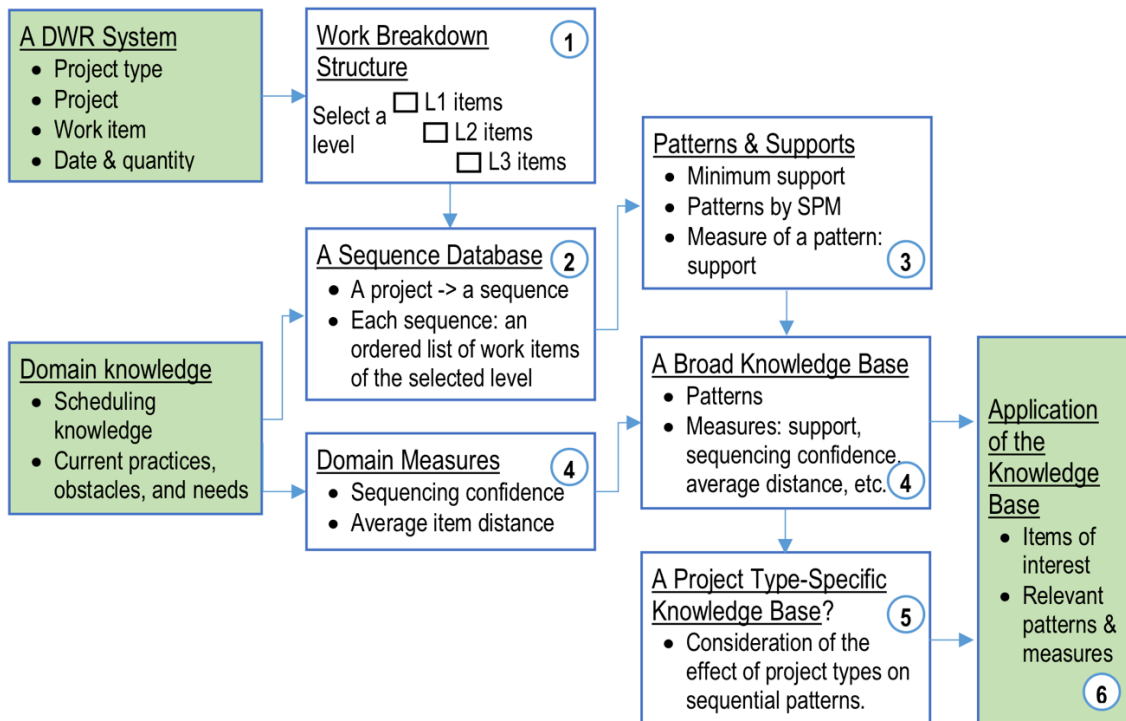


**Fig. 3.3.** Extracting precedence relationships between two work items from DWR data

### 3.6. Framework for developing knowledge bases of construction sequencing

This section presents a framework for developing a knowledge base of construction sequencing from an electronic DWR system (see Fig. 3.4). A DWR system is typically associated with a work breakdown structure (WBS) with three detail levels of work items. Depending on a schedule's required level of detail, one of the three levels can be used as scheduling items. For detecting logic patterns among the chosen work items, the DWR system needs to be transformed into a sequential database. The application of an SPM algorithm helps extract sequential patterns from the database and calculate the level of support (i.e., the number of projects in the database containing the pattern). However, the support alone is not helpful for sequencing purposes. Some domain measures (i.e., sequencing confidence and average item distance) were proposed to evaluate the extracted patterns. The patterns and measures constitute a broad knowledge base of construction logic patterns. A project context-specific knowledge base can be developed by examining the effects of project-specific influential factors on

construction sequencing. In this study, only project types are considered due to data availability issues, but the framework applies to other factors if data are available. For a given set of work items, relevant patterns and measures can be extracted from the knowledge base for schedulers' reference in sequencing a new project, with a suggestion of the most probable patterns among the items.



**Fig. 3.4.** An overall framework for developing a knowledge base of construction sequencing

### 3.6.1. Step 1: Select a level of the WBS as the basis of the knowledge base

A DWR system is typically associated with the agency's WBS. For example, each agency has a standard specification for road and bridge construction, specifying the agency's WBS. Table 3.1 shows a sample of a DOT's WBS with three levels of work: divisions (e.g., earthwork), sections (e.g., excavation and embankment), and specific

work items (e.g., excavation-unclassified). Depending on the required level of details of a construction schedule and project complexity, each of the three levels can serve as a basis for scheduling, hence three levels of work items, as the following:

- Level 1 – Division: L1 work item,
- Level 2 – Section: L2 work item, and
- Level 3 – Specific work item: L3 work item.

An agency’s WBS can contain thousands of specific work items, dozens of sections, and less than a dozen divisions. The selection of work-item levels, therefore, affects the number of relationships of work items in the final knowledge base.

**Table 3.1.** An example of WBS

DIVISION		SECTION		SPECIFIC WORK ITEM	
200	Earthwork	201	Clearing and Grubbing	201110005	Clearing
				201130000	Clearing and Grubbing
				...	...
		202	Removal of Structures and Obstructions	202020040	Remove Structure
				202020055	Remove Obstructions
				...	...
		203	Excavation and Embankment	203020100	Excavation-Unclassified
				203020175	Excavation-Unclass. Channel
				...	...
		...	...	...	...
300	Aggregate Surfacing and Base Courses	301	Aggregate Surfacing	301020340	Crushed Aggregate Course
				301020348	Drain Aggregate
				...	...
...	...	...	...	...	...
...	...	...	...	...	...

### 3.6.2. Step 2: Create a sequence database suitable for applying SPM algorithms

After selecting the work-item level, the next step is to extract relevant data from the DWR system and transfer them to a sequence database format suitable for applying SPM algorithms. The DWR system contains information regarding the start and end dates of L3 work items, the lowest WBS level. Information on L2 or L1 work items can be obtained by concatenating lower-level work items. Fig. 3.5 shows the proposed process of transforming the start and end dates of work items in a specific project into a corresponding sequence with three types of sequence databases: Start-Start, Finish-Finish, and Start/Finish-Start/Finish. The Start-Start database includes sequences of ordered start dates of work items, used for identifying Start-Start relationships between work items, with each sequence representing a project in the DWR system. A similar interpretation applies to the other two databases.

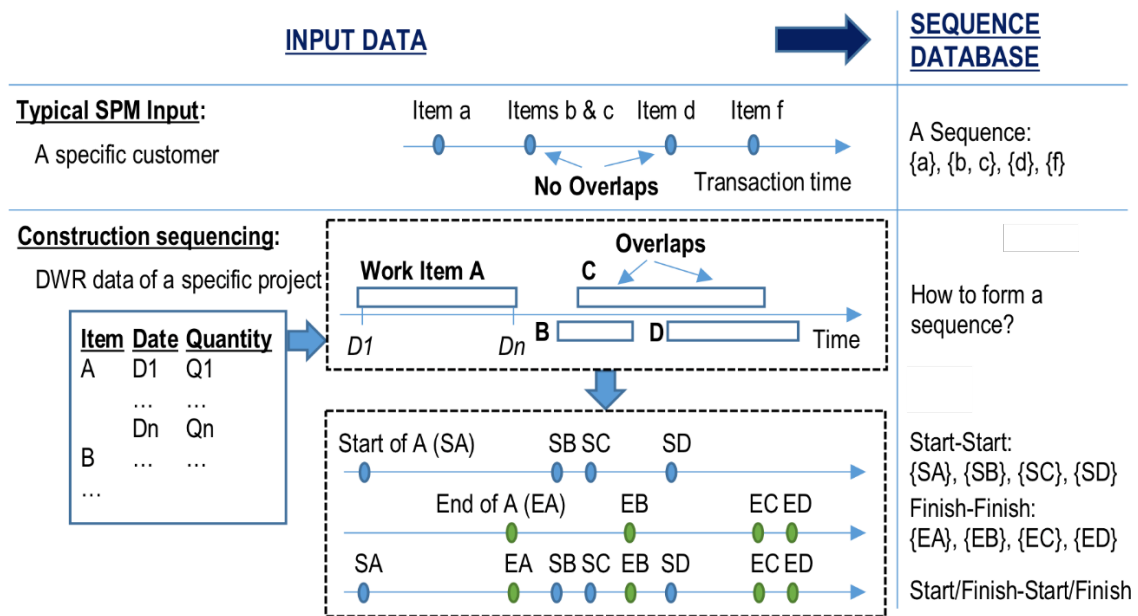


Fig. 3.5. Transforming DWR data to sequence databases



### 3.6.3. Step 3: Apply an SPM algorithm to obtain a list of frequent subsequences or patterns and their support

Given a sequence database and user-defined minimum threshold support, an SPM algorithm is used to find the subsequences having the support not smaller than the minimum threshold. Since SPM is an iteration problem with definite and unique solutions, various SPM algorithms should provide the same results.

Fig. 3.6 shows a simplified illustrative example. The input sequence database includes five sequences, and each sequence is the ordered start date of work items from a past project. For instance, in sequence ID #0, the start of Item 1 is followed by the start of Items 2 & 3, followed by Item 4 and then Item 6. Applying an SPM algorithm such as PrefixSpan in the Sequential Pattern Mining Framework (SPMF) library with the minimum support of three gives the output of eleven frequent subsequences (or patterns) and their support. For example, the pattern “{1}, {4}” (Item 4 starts after the start of Item 1) occurs in three sequences (IDs 0, 1, and 2) of the input.

Input		An SPM algorithm			Output		
ID	Sequence	Subsequence	Support	IDs			
0	<{1}, {2, 3}, {4}, {6}>	{1}	4	0,1,2,&4			
1	<{1, 3}, {4}, {5, 6}>	{1}, {4}	3	0,1,&2			
2	<{1, 2}, {4}, {5}>	{1}, {6}	3	0,1,&4			
3	<{2}, {3}, {4, 6}>	{2}	4	0,2,3,&4			
4	<{1}, {2}, {6}>	{2}, {4}	3	0,2,&3			
		{2}, {6}	3	0,3,&4			
		{3}	3	0,1,&3			
		{3}, {4}	3	0,1,&3			
		{3}, {6}	3	0,1,&3			
		{4}	4	0,1,2,&3			
		{6}	4	0,1,3,&4			

**Fig. 3.6.** An output of applying an SPM algorithm to a sequence database

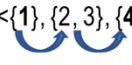


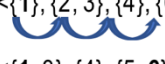
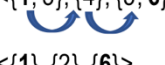

#### **3.6.4. Step 4: Propose and calculate domain-specific measures and build a broad knowledge base of construction sequencing**

The results from an SPM algorithm need to be processed further to develop sequences for future projects. The patterns discovered in the preceding step do not indicate the level of certainty associated with them, a critical feature needed for scheduling future projects. This study proposed two new measures specific to construction sequencing applications:

- Sequencing confidence and
- Average distance.

Sequencing confidence is the support divided by the number of sequences containing all items in the pattern. For example, the support of the pattern “Item 1 starts before the start of Item 4” is 3, and a total of three sequences of the input database contain both Item 1 and Item 4 (see Fig. 3.6). According to the definition, the sequencing confidence of the pattern is 100% ( $= 3/3$ ). In other words, the pattern occurred in 100% of the projects containing both Item 1 and Item 4. For a future project that contains both Items 1 & 4, schedulers are likely to arrange Item 1 before Item 4.

Also, schedulers are more interested in the relationships among activities that are typically closer to than far away from each other. For example, the pattern “{1},{6}” has the same support and sequencing confidence with “{1},{4}”, but the average distances of the two patterns are different (see Fig. 3.7). If a pattern has more than two items, the average distance is calculated based on the distance between the first and the last items.

Pattern	Sequence containing the pattern			
	ID	Detail	Distance	Average Distance
{1}, {4}	0	<{1}, {2, 3}, {4}, {6}> 	2	1.3
	1	<{1, 3}, {4}, {5, 6}> 	1	
	2	<{1, 2}, {4}, {5}> 	1	
{1}, {6}	0	<{1}, {2, 3}, {4}, {6}> 	3	2.3
	1	<{1, 3}, {4}, {5, 6}> 	2	
	4	<{1}, {2}, {6}> 	2	

**Fig. 3.7.** Illustration of the average distance measure

Another useful measure is the number of items in a pattern or the size of a pattern in items. One-item patterns (e.g., “{1}”) only show popular work items in past projects and are not useful for sequencing construction activities. Two-item patterns (e.g., “{1}, {4}”) show pairwise relationships between two work items and usually used by schedulers in sequencing a project. When the number of items in a pattern increases, the pattern provides more information to schedulers, but its sequencing confidence is more likely to decrease.

By the end of this step, the discovered patterns and their measures (e.g., support, sequencing confidence, and average distance) create a broad knowledge base of construction sequencing from a DWR system. The knowledge base can be further detailed by considering project characteristics that may affect work sequences. Thus, in this proposed framework, project types are a factor to be considered since this characteristic has been the primary factor influencing activity logic in previous research. Information about the project type is also available in a typical DWR system.

### **3.6.5. Step 5: Examine the effect of project types on the discovered patterns and build a project type-specific knowledge base if necessary**

The availability of project type information in the DWR system allows for the calculation of the support and confidence of a specific pattern for each project type. The Homogeneity Test can be used to compare the sequencing confidence of a pattern among project types. More details about the test can be found in (Ott and Longnecker 2015). If some project types have significantly different sequencing confidence of the same pattern, the patterns should be analyzed separately for each project type. The sequence database established at the end of Step 2 should be divided into smaller separate databases by project types before applying the same procedures as Steps 3 & 4, thereby creating a project-type specific knowledge base.

### **3.6.6. Step 6: Input work items of interest and extract relevant patterns from the knowledge base**

The knowledge base generally consists of a large number of patterns of construction work items, and many of the patterns may not be relevant to a new project for which a schedule is to be developed. For a given set of work items of interest, the knowledge base can return the patterns that only contain the items of interest. For example, if a scheduler is interested in the relationship between Items 1 & 2, the knowledge base can return at most three possible patterns: “{1},{2}” (Item 1 starts before Item 2), “{2},{1}” (Item 2 starts before Item 1), and “{1,2}” (Items 1 & 2 start on the same day). A McNemar exact binomial test can be used to compare the sequencing confidence among the patterns, which is similar to comparing proportions among

categories in statistics. More details about the test can be found in (Ott and Longnecker 2015). If a pattern has sequencing confidence significantly higher than the others, schedulers can be more confident in using the pattern for sequencing future projects.

### 3.7. Case study

Data from a DOT’s DWR system were obtained to illustrate the proposed framework. The data contained the DWRs of more than 700 projects from 2008 to 2017. The application of the framework on the dataset is described in this section.

#### 3.7.1. Step 1: Select a level of the WBS as the basis of the knowledge base

The DOT’s WBS consisted of six divisions (L1 work items), 46 sections (L2 work items as shown in Table 3.2), and more than 2,000 specific work items (L3 work items). The middle level was chosen to illustrate the framework since the number of sections was manageable by schedulers for scheduling purposes.

**Table 3.2.** List of sections and corresponding frequencies in the DWR data

Section	Description	Frequency (Project)
201	Clearing and Grubbing	30
202	Removal of Structures and Obstructions	188
203	Excavation and Embankment	350
207	Culvert Excavation and Trench Excavation	138
208	Water Pollution Control and Stream Preservation	264
209	Structure Excavation	27
212	Obliterate Roadway	23
301	Aggregate Surfacing	556
304	Portland Cement-Treated Base	16
401	Plant Mix Pavement	482
402	Bituminous Materials	503
403	Crack Sealing	25
409	Seal Coat	153
411	Cold Milling	324
501	Portland Cement Concrete Pavement	83
552	Concrete Structures	152

**Table 3.2. Continued**

Section	Description	Frequency (Project)
553	Prestressed Concrete Members	59
555	Reinforcing Steel	76
556	Steel Structures	25
557	Steel Bridge Railing	138
559	Piling	80
561	Bridge Deck Milling	63
562	Bridge Deck Repair	13
563	Modified Concrete Overlay	10
601	Water Service Lines	49
602	Remove and Relay Pipe Culvert	180
603	Culverts, Storm Drains, Sanitary Sewers, Stockpasses, and Underpasses	262
604	Manholes, Combination Manholes and Inlets, and Inlets	102
606	Guardrail and Concrete Barrier Rail	371
607	Fences	217
608	Concrete Sidewalks	139
609	Curbs and Gutters	196
610	Roadside Re-Vegetation	341
611	Cattle Guards	79
613	Riprap and Slope and Bank Protection	178
614	Retaining Walls	40
615	Irrigation Facilities and Headwalls	25
616	Conduits and Pull Boxes	155
617	Traffic Signals and Lighting	167
618	Traffic Control	722
619	Signs, Delineators, and Guideposts	507
620	Pavement Marking Application	562
621	Remove, Reset, and Adjust Facilities	94
622	Geotextiles	247
623	Mailboxes	97
624	Welding	64

### 3.7.2. Step 2: Create a sequence database suitable for applying SPM algorithms

For each project in the DWR data, the start dates of L2 work items were calculated and compared to form a sequence or an ordered list of work items in chronological order, with the project number as its identification number. For example, the sequence of the project #1420053000 had 18 items with “Traffic Control” (Code:

618) as the first and “Pavement Marking Application” (Code: 620) as the last. Such sequences of all projects in the DWR data constituted a sequence database for discovering Start-Start relationships among the work items, with 726 sequences representing 726 projects. As all 46 chosen L2 work items did not necessarily happen in the same project, the maximum number of work items in a sequence in the database was 35. If the L3 level was chosen, the number of items in a project would be much higher.

### **3.7.3. Step 3: Apply an SPM algorithm to obtain a list of frequent subsequences/patterns and their support**

Some items had a frequency as low as ten (see Table 3.2). Thus, a minimum support of five occurrences was selected to extract frequent subsequences/patterns. Furthermore, the selection of smaller minimum support was preferred to extract a larger number of patterns and reduce the elimination of potentially useful patterns. In the end, the support of each pattern was available for users in decision making. The SPM algorithm identified and extracted 127,325 subsequences or patterns using the criterion.

The first three columns of Table 3.3 show examples of the discovered patterns and their support, i.e., SPM output measure. The pattern {618} (or Traffic Control) had the largest support among the patterns (i.e., 716). However, it had only one item, which is not useful for construction sequencing. Pattern #2 indicated that “Traffic Control” started before “Pavement Marking Applications” in 506 past projects. However, the support alone was not a good indicator of the pattern’s level of confidence due to the lack of information about the number of projects containing the two items. Also, the two items were usually one of the first and one of the last items in a project, such as the

previous example of the 18-item sequence (Project #1420053000), so their relationship was pretty obvious and not useful to schedulers. Compared to Pattern #2, Pattern #3 had a much smaller support value, but it might provide more interesting and useful information to schedulers. Similar comparisons applied to Patterns #4 & 5. These observations emphasized the need for the calculation of domain-specific measures in the next step.

**Table 3.3.** Examples of the discovered patterns, the output measure from SPM in Step 3, and proposed domain-specific measures in Step 4

No.	Pattern	Support (SPM Output)	Proposed measures	
			Sequencing Confidence	Average Distance
<b>One-Item Pattern</b>				
1	{Traffic control (618)}	716	N/A	N/A
<b>Two-Item Patterns</b>				
2	{Traffic Control (618)}, {Pavement Marking Application (620)}	506	0.91	5.7
3	{Concrete Sidewalks (608)}, {Pavement Marking Application (620)}	99	0.77	2.9
<b>Three-Item Patterns</b>				
4	{Traffic Control (618)}, {Aggregate Surfacing (301)}, {Pavement Marking Application (620)}	219	0.65	9.3
5	{Traffic Control (618)}, {Culverts, Storm Drains, Sanitary Sewers, Stockpasses, and Underpasses (603)}, {Aggregate Surfacing (301)}	143	0.67	6.6

#### 3.7.4. Step 4: Propose and calculate domain-specific measures and build a broad knowledge base of construction sequencing

The proposed measures (i.e., sequencing confidence, average distance, and size) of each of the 127,325 patterns were calculated. The patterns, along with the measures,



constituted a broad knowledge base of construction sequencing without considering project characteristics yet. Users of the knowledge base can rely on the measures of a pattern to consider whether to apply it for scheduling a future project. From a scheduler's perspective, high sequencing confidence, support, and size are preferable while a small average distance is preferred. However, these references are not likely to coincide.

Table 3.3 also shows the proposed measures of the example patterns in Step 3. The sequencing confidence and average distance measures do not apply to one-item patterns as they are not useful for sequencing purposes. The sequencing confidence of a pattern provides users with the level of certainty associated with the pattern. For example, the sequencing confidence of 0.91 of Pattern #2 indicates that the pattern held for 91% of the past projects containing Items #618 & #620. However, its average distance (i.e., 5.7) suggests that the pattern may not be of interest due to the existence of other items between Items #618 & #620. Pattern #3, however, is more useful in that perspective. The default and proposed measures together provide multi-perspective evaluations about a pattern.

#### **3.7.5. Step 5: Examine the effect of project types on the discovered patterns and build a project type-specific knowledge base if necessary**

The DOT had 24 project types, such as drainage, overlays, Portland cement concrete pavement, and reconstruction & grading. The effect of project types on construction sequences was tested using the Homogeneity Test. Table 3.4 provides details for testing the effect of project types on a specific pattern.

**Table 3.4.** Pattern “{Aggregate Surfacing (301)}, {Bituminous Materials (402)}” under different project types

No.	Project type	Number of projects	Support of the pattern	Sequencing confidence
1	Reconstruction, grading	74	71	0.96
2	Overlays	65	34	0.52
3	Bridge construction, rehab and removal	17	16	0.94
4	Seal & cover	16	9	0.56
5	Safety	11	9	0.82
6	Rahab (Minor grade & overlay)	6	4	0.67
7	Others	10	5	0.50
	All types	199	148	0.74

The pattern “{Aggregate Surfacing (301)}, {Bituminous Materials (402)}” had overall sequencing confidence of 0.74 since the pattern held for 148 projects out of a total of 199 projects containing the two items. The sequencing confidence of the pattern for each project type was calculated to test whether there was a significant difference among project types. Those project types that had the support of smaller than five were combined to ensure the robustness of the test (see Table 3.4). This test involved seven populations/project types with two response categories (i.e., whether a project contained the pattern or not). The sequencing confidence was the proportion of the response category that the pattern held. The null hypothesis was that the sequencing confidence was the same among the seven project types, and it was rejected with a p-value of smaller than 0.0001. Therefore, project types affected the pattern’s sequencing confidence, hence a need for a project type-specific knowledge base.

A knowledge base specific for a project type can be developed by applying Step 3 and Step 4 to only those projects of that specific project type. For example, the work type “reconstruction, grading” had 115 projects in the database. By applying Step 3 and

Step 4 with the same minimum support of five, a knowledge base specific to the work type “reconstruction, grading” was developed, with 75,672 patterns.

**3.7.6. Step 6: Input work items of interest to schedulers and extract relevant patterns from the knowledge base**

This step illustrates the use of the knowledge base of construction sequencing for schedule development. For example, the reconstruction and grading knowledge base developed in this case study consisted of 75,672 patterns. Many of them are not useful for the schedule development of a specific new project. The suggested solution is for the user to input the work items of interest, and the knowledge base filters relevant patterns along with their measures. Then, the algorithm or users can select the pattern with the highest sequencing confidence for the new project if no additional project condition information is available. Table 3.5 provides two examples of applying the knowledge base.

**Table 3.5.** Examples of extracting the relevant patterns of interest from the reconstruction and grading knowledge base

No.	Pattern	Support	Sequencing Confidence
<b>Example 1: 301 &amp; 402</b>			
1	{Aggregate Surfacing (301)}, {Bituminous Materials (402)}	71	0.96
2	Other subsequences not satisfying the minimum support requirement	3	
<b>Example 2: 203, 301, &amp; 402</b>			
3	{Excavation and Embankment (203)}, {Aggregate Surfacing (301)}, {Bituminous Materials (402)}	47	0.64
4	{Aggregate Surfacing (301)}, {Excavation and Embankment (203)}, {Bituminous Materials (402)}	17	0.23
5	Other subsequences not satisfying the minimum support requirement	9	

In the first example, the most probable sequence for Aggregate Surfacing and Bituminous Materials is Pattern #1 (Aggregate Surfacing starts before Bituminous Materials) since it is the only reported pattern and its sequencing confidence is close to 100%. Some other possible subsequences are not reported as they are not frequent and do not meet the minimum support required to be considered as a pattern.

In the second example, the most probable sequential pattern among three items is Pattern #3 (Excavation and Embankment -> Aggregate Surfacing -> Bituminous Materials) as its confidence level is significantly higher than that of the other reported pattern (i.e., Pattern #4). A McNemar exact binomial test can also be applied to reinforce that comparison. Pattern #4 is irregular and unexpected, but it is possible due to one limitation of this case study, i.e., only considering the effect of project types on sequencing. Other possible influential factors, such as project phasing, were not considered due to data availability issues. Nevertheless, Pattern #4 is not suggested by the developed knowledge base for future projects as it is dominated by Pattern #3.

### **3.8. Comparison between the expected outputs of the template approach and the SPM-driven approach**

In the template approach, construction sequences of a project are assumed to be deterministic when the project type is known. However, the assumption is rejected based on the results of the SPM-driven approach. Although the project type is considered, most sequencing confidence values of the patterns in Table 3.6 are still significantly smaller than 1. The null hypothesis that the patterns are deterministic is rejected by the Binomial Test.

**Table 3.6.** Construction sequence patterns of the reconstruction, grading project type under the schedule template approach and the SPM-driven approach

Pattern	Schedule-template approach	SPM-driven approach				
		Support	Sequencing Confidence ( $\rho$ )	Binomial test, $\alpha = 0.05$		Deterministic?
				$H_0: p \geq p_0, H_1: p < p_0,$		
				$p_0 = 1.00$ p-value	$p_0 = 0.95$ p-value	
{Traffic Control (618)}, {Cold Milling (411)}		29	0.97	0.000	0.785	
{Traffic Control (618)}, {Water Pollution Control and Stream Preservation (208)}	No information about the support of each pattern.	77	0.68	0.000	0.000	No
{Water Pollution Control and Stream Preservation (208)}, {Geotextiles (622)}		75	0.75	0.000	0.000	No
{Water Pollution Control and Stream Preservation (208)}, {Aggregate Surfacing (301)}	The pattern is deterministic, or the sequencing confidence is 1.00 (100%).	94	0.86	0.000	0.000	No
{Aggregate Surfacing (301)}, {Plant Mix Pavement (401)}		76	0.87	0.000	0.004	No
{Geotextiles (622)}, {Plant Mix Pavement (401)}, {Pavement Marking Application (620)}		51	0.65	0.000	0.000	No

Apart from detecting (not predicting) the typical construction sequence patterns of a project type as the template approach, the SPM-driven approach also provides the certainty level associated with each pattern as the discovered patterns are common and frequent but not always correct. There are factors other than project types affecting the construction sequencing of a project, such as construction methods and project phasing. Also, there is no unique right way to sequence a project. Therefore, construction sequence patterns are not likely to be deterministic but associated with some level of uncertainty. Even though this study only considers the effect of project types on construction sequencing due to the data availability issue, other influential factors can be easily accounted for by the framework if data are available.

### **3.9. Discussions and conclusions**

Unlike construction sequencing by a contractor in the construction phase with detailed construction plans and inputs from key project members, U.S. highway owners' schedulers need to make various assumptions to develop a reasonable sequencing solution in the earlier project phases. However, DOT guidance on sequencing is limited. Some more advanced DOTs have developed expert-based logic templates to support schedulers in sequencing a new project to save time and improve productivity. While these templates take a significant amount of time to develop, they are dependent upon subjective expert judgment, deterministic, and possibly quickly outdated due to the continuous changes in the construction industry. Furthermore, there is a lack of feedback loop from construction to earlier phases to enhance the current scheduling practices. Thus, there is a need for a data-driven resource of construction sequencing to better support DOT schedulers with this challenging task. This study proposed a six-step framework for developing a knowledge base of construction sequencing from a DOT DWR system, which may significantly help schedulers, particularly inexperienced ones, in determining a defensible construction logic for a new project with data-backed evidence.

This study's primary contribution to the body of knowledge is a data-driven approach that allows for the automated creation of a knowledge base of construction sequences under different project conditions (e.g., different project types). The developed knowledge base can provide DOT schedulers with common sequence patterns or different types of sequential relationships between work items (i.e., Start-Start, Start-

Finish, Finish-Start, and Finish-Finish) adopted by contractors in past projects. These patterns are especially beneficial to young schedulers with limited construction experience. The discovered patterns among common work items should not be new to experienced schedulers, as the proposed framework offers an alternative and complementary way to sequence patterns to alleviate the heavily and solely dependence of DOTs on individual experts. However, the patterns still help provide feedback from the construction stage or reinforce assumptions used in past projects. Furthermore, patterns involving the work items new to schedulers can provide them with new knowledge as well.

The proposed measure “sequencing confidence” of a pattern provides the certainty level associated with the pattern, which is not available with logic templates. It also allows for a formal way of examining and evaluating the effect of an influential factor (e.g., project types) on construction sequencing. However, as shown in the case study, the confidence of the most probable pattern, in some cases, can be significantly smaller than 1 (e.g., Pattern #3 in Table 3.5). The measure is associated with unavoided noises as not all influential factors on sequencing can be considered due to data availability issues. Nevertheless, the relative difference in sequencing confidence among alternative patterns of a set of work items can be used to compare and propose the most probable pattern for future projects.

A DOT can apply the proposed framework to its available DWR data to enhance the current scheduling practice without collecting any additional data. However, the agency probably needs external technical support to implement the framework and

develop its first knowledge base probably because the agency may not have an adequate level of data mining expertise. Once a new system is established, the agency's employees can periodically update the knowledge base by themselves by providing a more extensive or newer set of DWR data to the system to stay up to date with the changing conditions of highway construction. To ensure the accuracy of the knowledge base, the agency needs to ensure the reliability of the framework's input data. Although the DWR data are not initially collected to support construction sequencing, the proposed framework does not require the agency's inspectors and resident engineers to perform beyond what they currently do for contract administration, progress monitoring, payment, and litigation purposes. The agency's effort to control and verify the DWR data will help avoid possible suspicion about the reliability of the data and the accuracy of the resulting knowledge base and enhance the performances of the initially intended tasks of the DWR data such as making payments to contractors.

Furthermore, the proposed framework can be extended to other construction sectors or other entities that maintain a DWR system and are interested in discovering and documenting their construction activities' sequential patterns. Once the various types of sequencing patterns are identified, the patterns can be effectively used as an excellent material for training and education sessions to enhance their organization's scheduling competence.

The study is limited by the lack of information on possible project phasing in the current DWR system. Without that information, the study could not consider the potential effect of project staging on construction sequencing. If a project has multiple



phases, each stage's sequencing should be analyzed, which can be quickly done, provided that detailed phasing information is available. A project phase will correspond to a sequence in the sequence database, and the proposed framework will still be applicable to develop a corresponding knowledge base. To account for project phasing, DOTs need to collect additional data for their systems, causing additional burden on the agencies. The key contribution is to provide DOTs with a sequencing knowledge base to support schedule development from a readily available data source without additional data collection efforts.

### **3.10. References**

Aggarwal, C. C. (2015). *Data Mining: The Textbook*, Springer.

Agrawal, R., and Srikant, R. "Mining sequential patterns." *Proc., The 11th International Conference on Data Engineering (ICDE'95)*, 3-14.

Bruce, R. D., Slattery, D. K., Slattery, K. T., and McCandless, D. (2012). "An Expert Systems Approach to Highway Construction Scheduling." *Technology Interface International Journal*, 13(1), 21-28.

CDOT (2019). "Construction Manual – Section 100: General Provisions." Colorado Department of Transportation, Colorado.

Chand, C., Thakkar, A., and Ganatra, A. (2012). "Sequential pattern mining: Survey and current research challenges." *International Journal of Soft Computing and Engineering*, 2(1), 185-193.

Cherneff, J., Logcher, R., and Sriram, D. (1991). "Integrating CAD with Construction-Schedule Generation." *Journal of Computing in Civil Engineering*, 5(1), 64-84.

- Chevallier, N. J., and Russell, A. D. (2001). "Developing a Draft Schedule Using Templates and Rules." *Journal of Construction Engineering and Management*, 127(5), 391-398.
- Chua, D. K. H., Nguyen, T. Q., and Yeoh, K. W. (2013). "Automated construction sequencing and scheduling from functional requirements." *Automation in Construction*, 35, 79-88.
- Echeverry, D., Ibbs, C. W., and Kim, S. (1991). "Sequencing Knowledge for Construction Scheduling." *Journal of Construction Engineering and Management*, 117(1), 118-130.
- Fan, S.-L., Sun, K.-S., and Wang, Y.-R. (2012). "GA optimization model for repetitive projects with soft logic." *Automation in Construction*, 21, 253-261.
- Fischer, M. A., and Aalami, F. (1996). "Scheduling with Computer-Interpretable Construction Method Models." *Journal of Construction Engineering and Management*, 122(4), 337-347.
- Florez, L. (2017). "Crew Allocation System for the Masonry Industry." *Computer-Aided Civil and Infrastructure Engineering*, 32(10), 874-889.
- Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., and Thomas, R. (2017). "A survey of sequential pattern mining." *Data Science and Pattern Recognition*, 1(1), 54-77.
- Hajdu, M. (1996). *Network scheduling techniques for construction project management*, Springer Science & Business Media.

- Hajdu, M. (2015). "Point-to-point versus traditional precedence relations for modeling activity overlapping." *Procedia Engineering*, 123, 208-215.
- Hajdu, M. (2018). "Survey of precedence relationships: Classification and algorithms." *Automation in Construction*, 95, 245-259.
- Hinze, J. (2012). *Construction planning and scheduling*, Pearson/Prentice Hall.
- Hu, D., and Mohamed, Y. (2014). "A dynamic programming solution to automate fabrication sequencing of industrial construction components." *Automation in Construction*, 40, 9-20.
- Jaśkowski, P., and Sobotka, A. (2006). "Scheduling Construction Projects Using Evolutionary Algorithm." *Journal of Construction Engineering and Management*, 132(8), 861-870.
- Jeong, H. D., Gransberg, D. D., and Shrestha, K. J. (2015). "Framework for Advanced Daily Work Report System." Mid-America Transportation Center.
- Jeong, H. D., Le, C., and Devaguptapu, V. (2019). "Effective Production Rate Estimation Using Construction Daily Work Report Data." Montana. Dept. of Transportation. Research Programs.
- Jeong, H. S., Atreya, S., Oberlender, G. D., and Chung, B. (2009). "Automated contract time determination system for highway projects." *Automation in Construction*, 18(7), 957-965.
- Kim, H., Anderson, K., Lee, S., and Hildreth, J. (2013). "Generating construction schedules through automatic data extraction using open BIM (building information modeling) technology." *Automation in Construction*, 35, 285-295.

- Le, C., and Jeong, H. D. "A Daily Work Report Based Approach for Schedule Risk Analysis." Springer Singapore, 1131-1136.
- Le, C., Jeong, H. D., Le, T., and Kang, Y. (2020). "Evaluating Contractors' Production Performance in Highway Projects Using Historical Daily Work Report Data." *Journal of Management in Engineering*, 36(3), 04020015.
- Leandro, R., O'Connor, J. T., and Khwaja, N. (2018). "Development and Application of a Production-Rate Resource for Contract Time Determination." *Journal of Construction Engineering and Management*, 144(12), 06018005.
- Lim, T.-K., Yi, C.-Y., Lee, D.-E., and Arditi, D. (2014). "Concurrent Construction Scheduling Simulation Algorithm." *Computer-Aided Civil and Infrastructure Engineering*, 29(6), 449-463.
- Liu, H., Al-Hussein, M., and Lu, M. (2015). "BIM-based integrated approach for detailed construction scheduling under resource constraints." *Automation in Construction*, 53, 29-43.
- Mubarak, S. A. (2015). *Construction project scheduling and control*, John Wiley & Sons.
- Newitt, J. S. (2009). *Construction Scheduling: Principles and Practices*, Pearson/Prentice Hall, New Jersey.
- Ohio DOT (2013). "Construction Duration Estimation Tool." Ohio Department of Transportation.
- Ott, R. L., and Longnecker, M. (2015). *An Introduction to Statistical Methods and Data Analysis*, Cengage Learning.

- Pierce, D. R. (2014). *Project scheduling and management for construction*, John Wiley & Sons.
- Shrestha, K. J., and Jeong, H. D. (2017). "Computational algorithm to automate as-built schedule development using digital daily work reports." *Automation in Construction*, 84, 315-322.
- Shrestha, K. J., Le, C., Jeong, H. D., and Le, T. "Mining Daily Work Report Data for Detecting Patterns of Construction Sequences." *Proc., Creative Construction Conference 2019*, 578-583.
- Son, J., Khwaja, N., and Milligan Duane, S. (2019). "Planning-Phase Estimation of Construction Time for a Large Portfolio of Highway Projects." *Journal of Construction Engineering and Management*, 145(4), 04019018.
- Stephenson, L., Douglas, E., Hanks, D., Hollmann, J., Jagathnarayanan, A., Nosbich, M., Uppal, K., White, P., and Wolfson, D. (2010). "Schedule classification system." *Morgantown, West Virginia: AACE International*.
- Tang, T., and McHale, G. (2016). "Big Data."   
<<https://www.fhwa.dot.gov/publications/publicroads/16sepoct/06.cfm>>. (May 24th, 2019).
- Taylor, T. R. B., Sturgill, R. E., and Li, Y. (2017). *Practices for Establishing Contract Completion Dates for Highway Projects*.
- Wang, W.-C., Weng, S.-W., Wang, S.-H., and Chen, C.-Y. (2014). "Integrating building information models with construction process simulations for project scheduling support." *Automation in Construction*, 37, 68-80.

Woldesenbet, A., Jeong, H. D., and Oberlender, G. D. (2012). "Daily Work Reports–  
Based Production Rate Estimation for Highway Projects." *Journal of  
Construction Engineering and Management*, 138(4), 481-490.

Zhai, D., Shan, Y., Sturgill, R. E., Taylor, T. R. B., and Goodrum, P. M. (2016). "Using  
Parametric Modeling to Estimate Highway Construction Contract Time."  
*Transportation Research Record*, 2573(1), 1-9.

## **4. NETWORK THEORY DRIVEN CONSTRUCTION LOGIC KNOWLEDGE NETWORK: PROCESS MODELING AND APPLICATION IN HIGHWAY PROJECTS**

### **4.1. Overview**

Determining a reasonable project duration is one of the most critical activities required by project owner agencies for successful project letting and delivery. Most owner agencies, specifically in the highway sector, mainly rely on schedulers' judgment and experience in determining the sequence of construction activities to estimate the required amount of time of a project. A vast amount of historical project performance data available in owner agencies' databases provide highly rich and reliable resources that can significantly improve the current process in order to produce a consistent and repeatable quality of construction logic determination. This study proposes a novel data-driven process model utilizing pattern mining, statistical analysis, and network analysis techniques that can detect pairwise logical relationships among construction activities (e.g., Start-Start and Finish-Start) and develop knowledge networks of as-built construction sequence patterns to improve the scheduling process. Three algorithms are proposed to apply the knowledge networks to sequencing a new project: finding immediate predecessors and successors of an activity or ordering a given set of activities. Ten years of historical project data obtained from a state department of transportation were used in this research. A case study reveals that the process model developed in this study can successfully build the most reasonable construction

sequences of a highway project, which can significantly improve the scheduling and contract time determination process.

#### **4.2. Introduction**

Network scheduling methods, such as the Critical Path Method (CPM), have been widely used to develop construction schedules due to their ability to show a logical sequence of construction activities such as start-start, finish-start, and lead and lag times. A scheduler needs to have a thorough understanding of the project and related site experience to reasonably estimate the project's schedule using a network scheduling method (Carson et al. 2014; IDOT 2017). Scheduling still heavily relies on a scheduler's experience and judgment, leaving a significant part of it as art, which is difficult to make it repeatable with an acceptable quality of the results when different people perform the scheduling job.

An alternative or an augmented approach to experience and judgment-based schedule development for improved consistency of the scheduling results is to systematically analyze as-built schedules from previous projects of similar characteristics and have the information available for schedulers. For example, analysis of as-built construction data can provide a sound basis for estimating realistic production rates of major work items and reveal the typical logical relationships of activities adopted by contractors in past projects (Alikhani et al. 2020; Le and Jeong 2020b). In the highway industry sector, such data are available in Daily Work Report (DWR) systems in most highway owner agencies - State Departments of Transportation (DOTs), which



are used to support contract administration, project monitoring, and contractor payment activities (Shrestha and Jeong 2017).

DOTs, much like any other owner organizations, need a reasonable contract time estimate to finalize the bid and contract documents for letting (TxDOT 2018). In a typical project delivered through a design-bid-build process, this is not an easy task as only design documents are complete, and the actual sequencing logic of the project is not known since a contractor is not selected yet. Hence, the owner's schedule is only an approximation of the actual future schedule, put in place to obtain a reasonable estimate of contract duration. In fact, the owner's schedule or the contract time is not expected to be the same or at the same level of detail as those of the contractors. However, the owner's schedule should be reasonable enough not to discourage qualified contractors from bidding on the project and affect project outcomes negatively (TxDOT 2018).

Most DOTs rely on bar charts or network scheduling methods to estimate contract time (Le and Jeong 2020a; Taylor et al. 2017). To support this estimation, two activities are considered to be most critical: a) production rate estimation of activities and b) construction logic development (Abdel-Raheem et al. 2020; FHWA 2002; Taylor et al. 2017). Some approaches have been developed to derive reliable production rates using historical project performance data, such as production rate tables (Hancher et al. 1992) and regression models (Jang et al. 2019; Jeong et al. 2019). Some recent studies have also utilized DWR data for realistic production rate estimation and evaluation (Jeong et al. 2019; Le et al. 2020).

However, to the best of our knowledge, there is no approach or method available in the highway construction sector that taps into the historical project performance data to extract meaningful patterns that can help support the project scheduling process, specifically, the construction logic development. In fact, most DOTs solely rely on schedulers' experience and knowledge in sequencing construction activities as there is very limited DOT guidance on this work task (ADOT 2015; CDOT 2019). With an increasing number of retirements of experienced schedulers, they are not always available, and in many cases, novices and less experienced schedulers are tasked to develop sequencing logic and estimate contract time (Bruce et al. 2012; Jeong et al. 2009).

To address this issue, several DOTs have developed logic templates for common project work types (Bruce et al. 2012; Jeong et al. 2009; Taylor et al. 2017). Each template consists of controlling activities (i.e., the ones that are likely to appear in the project's critical path) and their pairwise logical relationships (e.g., Start-Start and Finish-Start). However, this approach also depends on subjective expert input and can be outdated quickly due to the continuous changes in the construction industry (Shrestha et al. 2019). Furthermore, the pairwise relationships in the templates are fixed for a given project work type (Bruce et al. 2012; Jeong et al. 2009) and are not flexible in considering other influential factors such as project phasing, construction methods, and other constraints (ADOT 2015; Hinze 2012; MassDOT 2014).

In this research, years of historical DWR as-built construction data from a DOT are investigated to develop a model that can structurally improve the development of

sequencing logic and reduce the reliance on schedulers' subjective judgments. By applying a) sequential pattern mining (SPM), b) statistical analysis, and c) network analysis on the DWR data, pairwise relationships between activities and their probabilities of occurrence can be obtained and interlinked together in creating a knowledge network of construction logic. Three algorithms are proposed to support rapid sequencing activities in a new project: finding predecessors and successors as well as suggesting the most probable sequences for a given set of activities with the removal of possible redundant sequential patterns.

### **4.3. Background**

Expert knowledge-based logic templates have been developed for common project work types to provide a good starting point for determining scheduling logic (Taylor et al. 2017). The template of a project work type consists of the controlling activities of that project work type and their typical logical interrelationships (Bruce et al. 2012; Jeong et al. 2009; McCrary et al. 1995). Template-based systems vary significantly among the highway agencies due to state-specific practices and system developers' judgments and preferences. For example, the Kentucky contract time determination system considers only six templates (Werkmeister et al. 2000), while Oklahoma DOT considers sixteen (Jeong et al. 2009). There is also a difference in the types of logical relationships used in each system. For example, Hancher et al. (1992) and Werkmeister et al. (2000) only applied Start-Start relationships with lags (e.g., concrete paving starts after the 75% completion of milling) and Finish-Start relationships

without lags (e.g., final clean up starts after permanent pavement marking is complete) for their systems.

Apart from U.S. DOT-related research, there are a significant amount of studies related to construction sequencing, which can be divided into two main research topics: automated schedule development and schedule optimization. For automated schedule development, some studies have leveraged expert inputs (e.g., functional dependencies among structural components or predetermined dependencies among pipe fabrication operations) (Chua et al. 2013; Hu and Mohamed 2014), while others have utilized information contained in computer-aided-design drawings (Cherneff et al. 1991; Fischer and Aalami 1996) and building information modeling models (Liu et al. 2015; Wang and Yuan 2017). In schedule optimization, some studies have also relied on expert knowledge but to capture the impact of changes in other schedule components (e.g., overlap between activities, crew allocation, or workspace interference) (Florez 2017; Lim et al. 2014; Tao et al. 2020).

The studies mentioned earlier, especially those dedicated to DOT contract time systems, heavily rely on the knowledge and experience of schedulers for logic development. For instance, DOTs applied different logical relationship types with and without lags for their logic templates as they deemed appropriate. This application is inherently subjective. Hence a more rational and objective process to select relationship types and decide on the use of lags is needed. An analysis of historical project data could be used as excellent resources to detect as-built logical relationships in past projects and

improve the owner's confidence in developing a project's construction logic with high reliability and consistency.

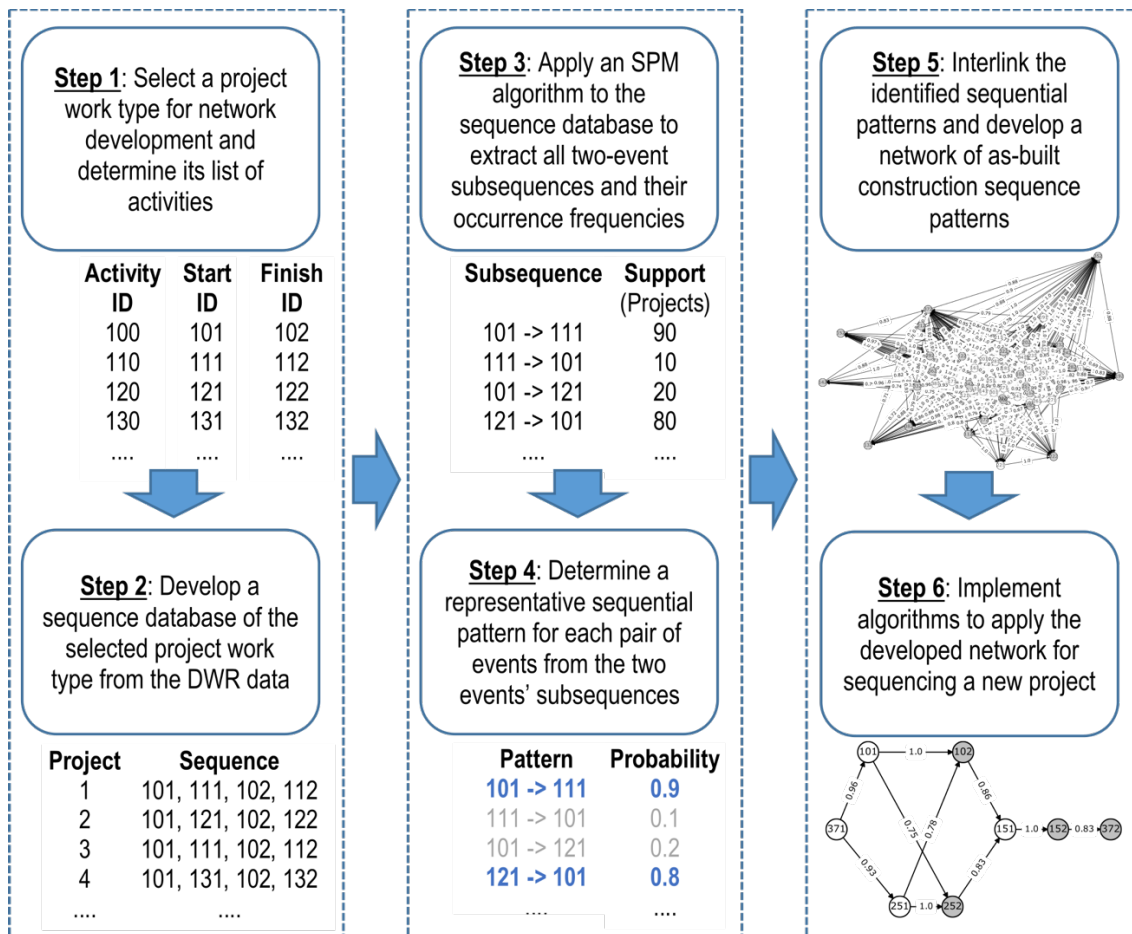
#### **4.4. Research objective and scope**

Developing sequencing logic is important for both project owners and contractors. For DOTs, this task is typically done by a DOT scheduler who determines construction sequencing at the end of the final design phase for estimating construction contract time before construction starts; note that this is often different from the sequencing logic the contractor will develop, considering its available resources, means and methods (Idaho DOT 2011; TxDOT 2018). Due to the lack of information regarding how the contractor will construct the project and input from construction team members (Mubarak 2015), and the existence of more than one right way to construct any project (Newitt 2009), the construction sequence determined by the DOT is not expected to be the same as that in a detailed construction schedule later developed by the contractor. However, it should be reasonable enough to avoid unreasonably short or long contract time (Idaho DOT 2011; TxDOT 2018) that can severely affect the bidding process and the selection of a winning contractor.

The purpose of this study is to develop an innovative DWR data-driven approach to construction sequencing to enhance DOTs' contract time estimation processes for common project work types. While large and complex projects are highly unique and may not benefit much from historical project data, as-built schedules of past projects of common project types have similarities and can provide insights for future project schedule development.

#### 4.5. Process model to develop and apply a sequencing knowledge network

This section presents a novel process model for developing knowledge networks of as-built construction sequence patterns in past project data that leverages sequential pattern mining (SPM), statistical analysis, and network analysis techniques (see Fig. 4.1).



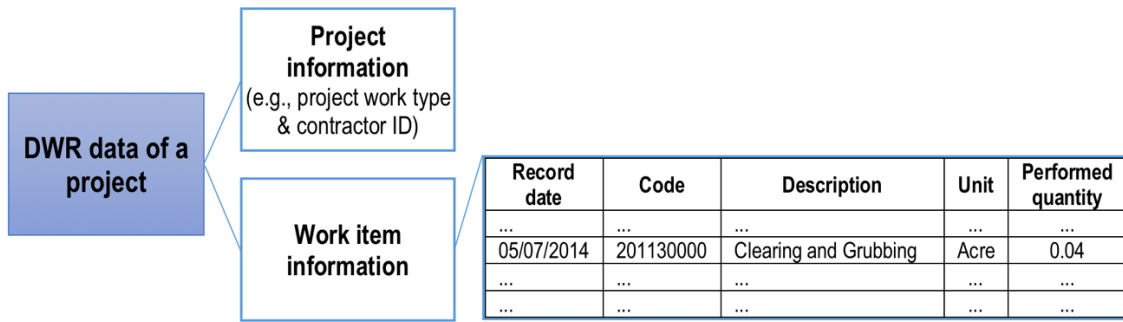
**Fig. 4.1.** Process model to develop and apply a sequencing knowledge network

The first step in this process is to establish a list of activities used for scheduling. Next, relevant DWR data of a project work type and related activities are then extracted from a DWR system and transformed into a sequence database. Each sequence is an

ordered list of events occurring in a specific project, while each event is the start or the finish of an activity in the project. The next step involves applying an SPM algorithm to the sequence database to extract two-event subsequences or pairwise sequential relationships between two activities. In Step 4, possible subsequences of each pair of events are compared to determine whether there is a pattern between the two events. The discovered patterns are then connected and visualized in a single network in Step 5. Finally, three network analysis algorithms are designed and implemented in Step 6 to support the application of the developed network in sequencing construction activities in a new project. In the next section, we provide more detailed descriptions of these sequential steps.

#### **4.5.1. Step 1: Select a project work type for network development and determine its list of activities**

DWRs are created and maintained by DOTs' site inspectors to record contractors' activities and performances on a daily basis in the construction phase for contract administration purposes (e.g., payment) (Shrestha and Jeong 2017). DOTs have a systematic pre-defined list of work items with item codes, descriptions, and units (e.g., 201130000, Clearing and Grubbing, and Acre) to be used across their projects, and DWR data attributes center around the work items included in a specific project. Main attributes include project identification numbers, project work types, record dates, work items' codes, descriptions, and units, accomplished quantities of the work items performed on a specific date, and contractor name and identification number (Le et al. 2020) (see Fig. 4.2).



**Fig. 4.2.** DWR data attributes

Project work type (e.g., overlay, reconstruction, or seal & cover) is one of the most important drivers of activity sequencing (Bruce et al. 2012; Chevallier and Russell 2001; Jeong et al. 2009; Shrestha et al. 2019). Construction sequences of projects of the same project work type tend to be more similar to each other than those of different types as required activities vary significantly by project work type. Therefore, the first action is to select a type of interest. Only projects of the selected type are extracted from the DWR data for further analysis.

Each project work type may involve hundreds of work items. For example, some work items of an overlay project could be “Clearing and Grubbing (Code = 201130000),” “Cold Milling (411010000),” “Cover – Type 1 (409000010),” “Excavation – Unclassified (203020100),” and “Sidewalk – Concrete 4 IN (608010020).” Schedulers typically do not use all of these relevant work items for scheduling purposes. They only use major items (e.g., Illinois DOT), critical items (e.g., Idaho DOT), or controlling items of work – the items are likely to be on the critical path (e.g., Colorado DOT) (CDOT 2019; Idaho DOT 2011; IDOT 2017). Other agencies use other terms, such as controlling work activities (Arizona DOT) or controlling operations



(North Carolina DOT) (ADOT 2015; NCDOT n.d.). In this paper, we use the general term “controlling activity.”

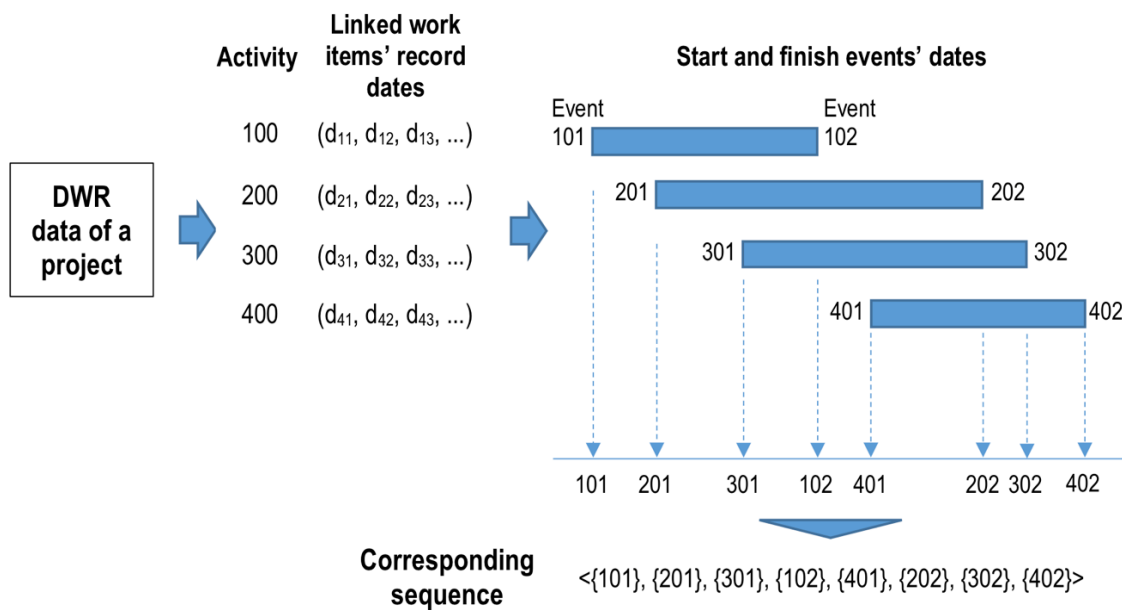
There are two main methods used by DOTs to determine controlling activities from a work item list. The first method is to link multiple similar work items into a single activity. For example, items “Reinforcing Steel,” “Reinforcing Steel-Epoxy Coated,” and “Reinforcing Steel-Stainless” can be represented by the activity “Reinforcing Steel.” The second method is to group miscellaneous work items into a composite activity, based on a DOT Work Breakdown Structure (WBS) specified in the DOT’s standard specification. For example, work items such as “Adjust Drop Inlet” and “Adjust Fire Hydrant” can be grouped into a single activity, “Remove, Reset, and Adjust Facilities.” After the grouping, a list of activities is formed. The number of activities can further be reduced by eliminating the activities that have rarely appeared in past projects.

#### **4.5.2. Step 2: Develop a sequence database of the selected project work type from the DWR data**

For a past project, the DWR data provide information about all work items performed in the project, including the dates each work item was performed. Therefore, the start and the finish dates of an activity identified in Step 1 can be determined by sorting the times recorded in the DWR dataset of the linked work items. Each project of the selected project work type corresponds to an ordered sequence of events, while each event is the start or finish of an activity included in the project

There is a need for coding the activities and events for later analysis. An activity can be coded as  $X0$  ( $X$  is a positive integer), and  $X1$  and  $X2$  are its start and finish events.

Fig. 4.4 shows the process of transforming the DWR data of a project into a corresponding sequence. For simplicity, assume that a project has four activities 100, 200, 300, and 400. Each activity has a start event and a finish event (e.g., events 101 and 102 of activity 100), resulting in eight events in total. Those events' occurring dates can be calculated from the project's DWR data and then compared to form an ordered sequence of events, i.e.,  $\langle \{101\}, \{201\}, \{301\}, \{102\}, \{401\}, \{202\}, \{302\}, \{402\} \rangle$ . A list of sequences or a sequence database is established by applying a similar process to every project, with each sequence given an identification number (ID) for later retrieval. As each project corresponds to a sequence, the number of sequences in the sequence database is equal to the number of projects of the selected project work type.



**Fig. 4.3.** Transform the DWR data of a project into a sequence

### **4.5.3. Step 3: Apply an SPM algorithm to the sequence database to extract all two-event subsequences and their occurrence frequencies**

An SPM algorithm can extract all two-event sub-sequences from the sequence database along with their support values. For example,  $\langle\{101\},\{201\}\rangle$ ,  $\langle\{201\},\{301\}\rangle$ , and  $\langle\{101\},\{301\}\rangle$  are three of many two-event subsequences of the sample sequence in Fig. 4.3. The support of a sub-sequence is the number of projects in the sequence database containing the sub-sequence. As each event is the start or finish of an activity, a two-event sub-sequence represents a pairwise logical relationship between activities, such as a Start-Start, Start-Finish, Finish-Start, or Finish-Finish relationship.

Various SPM algorithms are available to find pairwise logical relationships between activities and their support values. Some popular algorithms are Generalized Sequential Patterns (GSP), Prefix-projected Sequential Pattern Mining (PrefixSpan), and Sequential PAttern Mining (SPAM) (Fournier-Viger et al. 2017; Mooney and Roddick 2013).

### **4.5.4. Step 4: Determine a representative sequential pattern for each pair of events from the two events' subsequences**

The output of Step 3 is the list of two-event sub-sequences (each event is the start or finish of an activity) and the support/occurrence frequency of each sub-sequence. However, not every sub-sequence is a sequential pattern. For example, assume that 100 projects contained events  $A$  and  $B$ . The sub-sequence  $\langle\{A\},\{B\}\rangle$  (event  $B$  happened later than event  $A$ ) that occurred in only three out of 100 projects (the support = 3) may not be considered as a sequential pattern when the opposite sub-sequence (i.e.,  $\langle\{B\},$

$\{A\}$ ) that occurred in 97 out of 100 projects (the support = 97) could be declared as a reliable sequential pattern. As a pair of events (e.g.,  $A$  &  $B$ ) can have multiple subsequences (e.g.,  $\langle\{A\}, \{B\}\rangle$  and  $\langle\{B\}, \{A\}\rangle$ ), comparing their support is necessary to determine a representative sequential pattern of the pair of events.

Given two events  $A$  and  $B$ , there are three possible two-event subsequences:

1.  $\langle\{A, B\}\rangle$  (events  $A$  &  $B$  happened on the same day) with the support (i.e., the number of projects in the sequence database containing the subsequence) =  $n1$ ,
2.  $\langle\{A\}, \{B\}\rangle$  (event  $B$  happened at a later day than  $A$ ) with the support =  $n2$ , and
3.  $\langle\{B\}, \{A\}\rangle$  (event  $A$  happened at a later day than  $B$ ) with the support =  $n3$ .

While subsequence #1 does not inform about the order of  $A$  and  $B$ , subsequences #2 and #3 demonstrate two opposite orders. A comparison between  $n2$  and  $n3$  is necessary to determine which event ( $A$  or  $B$ ) is more likely to start first.

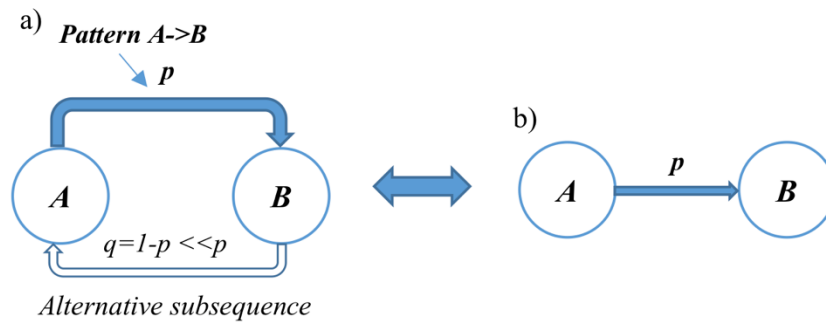
In the context of statistical analysis, the comparison between  $n2$  and  $n3$  can be framed as the following. A sample of  $n$  projects contain both events  $A$  and  $B$  ( $n = n1 + n2 + n3$ ). Of those projects,  $n2$  projects include  $\langle\{A\}, \{B\}\rangle$ , and  $n3$  projects contain  $\langle\{B\}, \{A\}\rangle$ . Statistical analysis can be applied to test whether there is a significant difference between the proportion of the projects containing  $\langle\{A\}, \{B\}\rangle$  and the percentage of the projects containing  $\langle\{B\}, \{A\}\rangle$ . As the two ratios are dependent and the sample size can be small, the McNemar exact binomial test is applied instead of Z-test. Ott and Longnecker (2015) provide detailed information about the test.

The test result is then used to determine a representative sequential pattern for the two events. There are three scenarios:

1. If  $n_2$  (i.e., the number of past projects containing  $\langle \{A\}, \{B\} \rangle$ ) is not significantly different from  $n_3$  (i.e., the number of past projects containing  $\langle \{B\}, \{A\} \rangle$ ), there is no sequential pattern between event  $A$  and event  $B$  as either event can start first.
2. If  $n_2$  is significantly larger than  $n_3$ , event  $A$  is more likely to be followed by event  $B$  ( $A \rightarrow B$ ), with the support of  $m = n_1 + n_2$  ( $n_1$ : the number of past projects that  $A$  and  $B$  happened on the same day, or  $A$  was followed by  $B$  with lag = 0).
3. If  $n_3$  is significantly larger than  $n_2$ , event  $B$  is more likely to be followed by event  $A$  ( $B \rightarrow A$ ), with the support of  $m = n_1 + n_3$ .

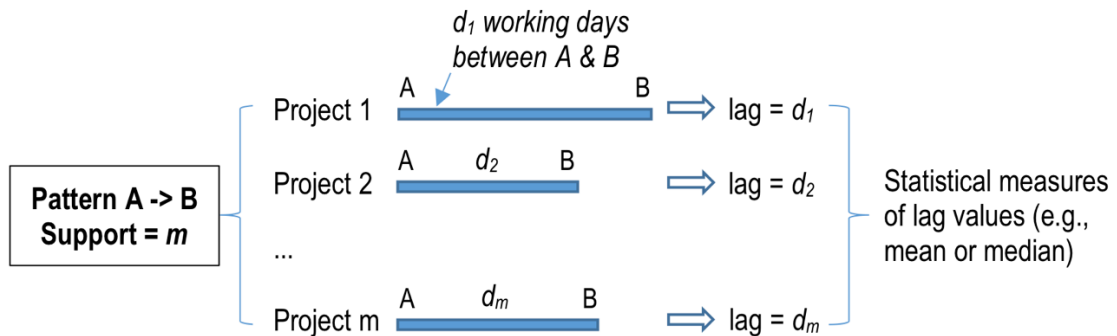
If  $m$  is larger than or equal to a minimum required support in scenarios #2 or #3, the corresponding order can be considered a pattern. The probability of the pattern given the two items is  $p$ , with  $p = m/n$ .

Pattern  $A \rightarrow B$ , with a probability of  $p$ , can be visualized in two different ways. As  $p$  can be smaller than 1, the likelihood of the opposite direction is  $q (=1-p)$ , although  $q$  is significantly lower than  $p$ , as shown in Fig. 4.4a. A more straightforward but equivalent representation of the pattern is illustrated in Fig. 4.4b. The value of  $p$  indicates the possibility of the opposite direction with the probability of  $(1-p)$ . The latter is adopted in this study.



**Fig. 4.4.** Visualization of a pattern

The lag time between two events of a pattern can also be calculated, as shown in Fig. 4.5. A pattern  $A \rightarrow B$  with support of  $m$  means event  $A$  was followed by event  $B$  in  $m$  projects in the database. The lag value in each project is the number of working days between event  $A$  and event  $B$ . Statistical measures of the lag time of the pattern can then be easily obtained.

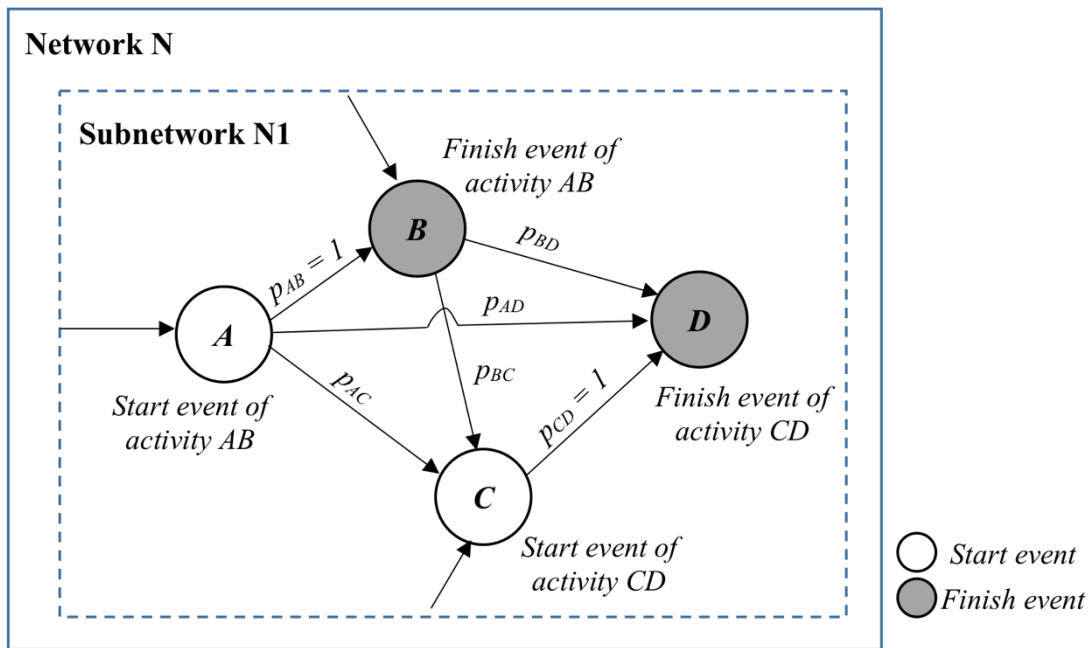


**Fig. 4.5.** Find the lag between two events of a pattern

#### 4.5.5. Step 5: Interlink the identified sequential patterns and develop a network of as-built construction sequence patterns

The output of Step 4 is a list of patterns (i.e., pairwise logical relationships between activities) and their probabilities. Network analysis can be applied to interlink

and analyze events from different patterns and visualize all discovered patterns in a directed network with nodes and directed edges. Fig. 4.6 shows a portion of network  $N$ , subnetwork  $N1$ , with four nodes for illustration, in which nodes represent events, and edges represent sequential relationships. A directed edge/arrow from event  $A$  to event  $B$  represents the pattern that event  $A$ , the head node of the edge, is followed by event  $B$ , the edge's tail node. The probability of each pattern is also shown on the corresponding edge to demonstrate the pattern's strength.



**Fig. 4.6.** An example of a directed network

Two colors are used to differentiate two types of events: white for start events and grey for finish events. One can quickly recognize whether an edge represents a Start-Start, Start-Finish, Finish-Start, or Finish-Finish relationship based on the two connected nodes' colors. For example, if events  $A$  and  $B$  are the start and finish of activity  $AB$ , and

events  $C$  and  $D$  are those of activity  $CD$ , then

- Edge  $B \rightarrow C$  (grey to white) represents the Finish-Start relationship between activity  $AB$  and activity  $CD$ , with a probability of  $p_{BC}$ ,
- Edge  $A \rightarrow C$  (white to white) represents the Start-Start relationship between activity  $AB$  and activity  $CD$ , with a probability of  $p_{AC}$ ,
- Edge  $B \rightarrow D$  (grey to grey) represents the Finish-Finish relationship between activity  $AB$  and activity  $CD$ , with a probability of  $p_{BD}$ , and
- Edge  $A \rightarrow D$  (white to grey) represents the Start-Finish relationship between activity  $AB$  and activity  $CD$ , with a probability of  $p_{AD}$ . However,
- Edge  $A \rightarrow B$  (white to grey) represents a trivial Start-Finish relationship: event  $A$  – the start of activity  $AB$  is followed by event  $B$  – the finish of activity  $AB$ . The probability in the case is equal to 1. Similarly,
- Edge  $C \rightarrow D$  also represents a trivial Start-Finish relationship: event  $C$  – the start of activity  $CD$  is followed by event  $D$  – the finish of activity  $AD$ .

In this study, the development of networks from the discovered patterns and network analysis were implemented in Python. A Python package named NetworkX was used to facilitate network modeling and analysis. Detailed information about this package can be found in Hagberg et al. (2008).

#### **4.5.6. Step 6: Implement algorithms to apply the developed network for sequencing a new project**

The developed network in Step 5 contains all discovered patterns of all activities identified in Step 1 based on the DWR data of past projects. However, only a subset of



the activities involves in a new project. Algorithms are necessary to extract only relevant patterns from the network, depending on the required information for scheduling. Three primary algorithms are needed for scheduling new projects:

1. Find the successors of an event (i.e., the start or the finish of an activity),
2. Find the predecessors of an event, and
3. Find the sequencing relationships among a set of activities of interest.

***Algorithm S1: Find the successors of an event***

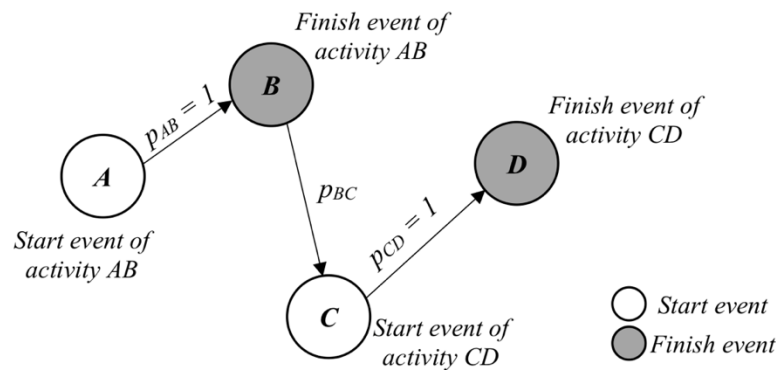
The developed network in Step 5 can be used to find not only the successors of an event of interest but also the relationships among the successors to maximize retrieval information. For example, an output as  $A \rightarrow B \rightarrow C \rightarrow D$  provides more information than three patterns:  $A \rightarrow B$ ,  $A \rightarrow C$ , and  $A \rightarrow D$ . Both cases indicate that  $B$ ,  $C$ , and  $D$  are successors of  $A$ , but the former may suggest that  $B$  is a direct successor of  $A$  and that  $C$  and  $D$  only follow  $A$  via  $B$ . A redundant relationship that can be inferred from other relationships should be eliminated to avoid unnecessary confusion.

Table 4.1 presents algorithm S1 and an example for illustration. The algorithm takes the developed network in Step 5 and a query event provided by a user as input. First, it searches the network to find all succeeding events/nodes of the query event/node, which are the tail nodes of edges starting from the query node. Second, a subnetwork is extracted from the network to include only the query node and its successors. Third, each relationship/edge in the subnetwork is checked whether it is redundant or, in other words, can be deducted from other edges. For each edge (e.g.,  $A \rightarrow D$ ), its head and tail nodes (e.g.,  $A$  and  $D$ ) and all paths from the head to the tail (e.g.,

A->D or A->B->D) are returned. If at least one path has more than two nodes (e.g., A->B->D), the edge under checking (e.g., A->D) is redundant. In the end, the algorithm returns a simplified subnetwork that contains only the event of interest and its successors and their sequential relationships without redundancy (see Fig. 4.7).

**Table 4.1.** Algorithm S1 – Find the successors of an event

Algorithm	Example
<p><u>Input:</u></p> <ul style="list-style-type: none"> <li>○ <u>From previous steps:</u> the developed network in Step 5</li> <li>○ <u>From users:</u> the event of interest</li> </ul> <p>1. Return all successor nodes of the query event, which are the tail nodes of edges starting from the query event.</p> <p>2. Return a subnetwork that contains the query event and its successors.</p> <p>3. Eliminate redundant edges that can be deduced from other edges to obtain a simplified subnetwork. Details are as follows.</p> <p style="padding-left: 20px;">For each edge in the subnetwork, return its head node and tail node.</p> <p style="padding-left: 20px;">Return all paths from the head to the tail.</p> <p style="padding-left: 40px;">If there is a path that contains more than two nodes, remove the directed edge from the head to the tail.</p> <p><u>Output:</u> A subnetwork that contains only the event of interest and its successors and their sequential relationships</p>	<p><u>Input:</u></p> <ul style="list-style-type: none"> <li>○ Network <math>N</math> in Fig. 4.6</li> <li>○ Event <math>A</math></li> </ul> <p>1. Events <math>B</math>, <math>C</math>, and <math>D</math></p> <p>2. Subnetwork <math>N1</math> in Fig. 4.6</p> <p>3. Pattern <math>A \rightarrow D</math> can be inferred from the two patterns <math>A \rightarrow B</math> and <math>B \rightarrow D</math>, so it can be eliminated from the subnetwork. Similarly, <math>A \rightarrow C</math> and <math>B \rightarrow D</math> can be eliminated.</p> <p>Fig. 4.7 shows the simplified subnetwork.</p> <p><u>Output:</u> The simplified subnetwork in Fig. 4.7.</p>



**Fig. 4.7.** The simplified version of subnetwork  $N1$

***Algorithm S2: Find the predecessors of an event***

Similarly, Table 4.2 presents algorithm S2 that takes the developed network in Step 5 and a query event provided by a user as input and returns a simplified subnetwork that contains only the event of interest and its predecessors and their sequential relationships.

**Table 4.2.** Algorithm S2 – Find the predecessors of an event

---

**Input:**

- From previous steps: the developed network in Step 5
  - From users: the event of interest
1. Return all predecessor nodes of the query event, which are the head nodes of edges ending at the query event.
  2. Return a subnetwork that contains the query event and its predecessors.
  3. Eliminate redundant edges that can be deduced from other edges to obtain a simplified subnetwork. Details are as follows.

For each edge in the subnetwork, return its head node and tail node.

Return all paths from the head to the tail.

If there is a path that contains more than two nodes, remove the directed edge from the head to the tail.

**Output:** A subnetwork that contains only the event of interest and its predecessors and their sequential relationships

---

***Algorithm S3: Find sequential relationships among a set of activities of interest***

The developed network in Step 5 can also be used to extract only relevant patterns to a set of interest activities and identify the first event(s), the last event(s), and paths from the first to the last event(s). Table 4.3 presents algorithm S3 and an example for illustration. The algorithm takes the developed network in Step 5 and query activities provided by a user as input. It returns a simplified subnetwork that contains only the events of the activities of interest and paths from the first event(s) to the last event(s).

**Table 4.3.** Algorithm S3 – Find the sequencing relationships among a set of activities

<b>Algorithm</b>	<b>Example</b>
<p><u>Input:</u></p> <ul style="list-style-type: none"> <li>○ <u>From previous steps:</u> the developed network in Step 5</li> <li>○ <u>From users:</u> activities and events of interest</li> </ul> <ol style="list-style-type: none"> <li>1. Return a subnetwork that contains the events of interest.</li> <li>2. Eliminate redundant edges that can be deduced from other edges to obtain a simplified subnetwork. Details are as follows.               <ul style="list-style-type: none"> <li>For each edge, return its head node and tail node.</li> <li>Return all paths from the head to the tail.</li> <li>If there is a path that contains more than two nodes, remove the directed edge from the head to the tail.</li> </ul> </li> <li>3. Return the first event(s) of the simplified subnetwork (i.e., the events that have no predecessors. Details are as follows.               <ul style="list-style-type: none"> <li>For each node in the simplified subnetwork, return its in-degree (e.g., the number of edges that come to the node).</li> <li>If a node has an in-degree of 0, it is the first event.</li> </ul> </li> <li>4. Return the last event(s) of the simplified subnetwork (i.e., the events that have no successors. Details are as follows.               <ul style="list-style-type: none"> <li>For each node in the simplified subnetwork, return its out-degree (e.g., the number of edges that leave the node).</li> <li>If a node has an out-degree of 0, it is the last event.</li> </ul> </li> <li>5. Return all the paths from the first event(s) to the last event(s)</li> </ol> <p><u>Output:</u> The simplified subnetwork of the events of interest and paths from the first events to the last events</p>	<p><u>Input:</u></p> <ul style="list-style-type: none"> <li>○ Network <i>N</i> in Fig. 4.6</li> <li>○ Events <i>A</i>, <i>B</i>, <i>C</i>, and <i>D</i></li> </ul> <ol style="list-style-type: none"> <li>1. Subnetwork <i>NI</i> in Fig. 4.6</li> <li>2. The simplified subnetwork in Fig. 4.7</li> <li>3. The in-degrees of events <i>A</i>, <i>B</i>, <i>C</i>, and <i>D</i> in the simplified network are 0, 1, 1, 1, respectively. <i>A</i> is the first event.</li> <li>4. The out-degrees of events <i>A</i>, <i>B</i>, <i>C</i>, and <i>D</i> in the simplified network are 1, 1, 1, 0, respectively. <i>D</i> is the last event.</li> <li>5. Paths from <i>A</i> to <i>D</i> in the simplified subnetwork include <i>A</i>-&gt;<i>B</i>-&gt;<i>C</i>-&gt;<i>D</i>.</li> </ol> <p><u>Output:</u> The simplified subnetwork in Fig. 4.7 and the path from the first event to the last event (i.e., <i>A</i>-&gt;<i>B</i>-&gt;<i>C</i>-&gt;<i>D</i>).</p>

#### 4.6. Case study

The DWR data of 190 overlay projects conducted from 2008 to 2017 were obtained from a DOT to illustrate the proposed approach. This section describes the application of the approach to the dataset.

#### **4.6.1. Step 1: Select a project work type for network development and determine its list of activities**

The 190 projects of work type overlay involved 639 work items, which were too detailed and unnecessary for scheduling purposes. A list of 135 activities was established by 1) linking similar work items to a single activity or 2) grouping miscellaneous work items together. Examples of the former were activity “Special Borrow” representing work items “Special Borrow-Excavation” and “Special Borrow-Neat Line” or activity “Plant Mix Surfacing,” representing work items of different mixture grades such as bid item “Plant Mix Surf GR S-3/4 IN.” An example of the latter was activity “Drainage Pipe,” representing drainage-pipe work items of different materials and dimensions, such as work items “Drainage Pipe 450 MM” and “Drainage Pipe 600 MM.”

By further applying a minimum frequency threshold of 5%, activities that appeared in less than ten projects out of the 190 projects were eliminated. A filtered list of 29 activities was obtained and used for further analysis. Table 4.4 shows the filtered list of activities and corresponding frequencies.

**Table 4.4.** List of construction activities

<b>Activity ID</b>	<b>Activity Name</b>	<b>Frequency (Projects)</b>	<b>Start-Event ID</b>	<b>Finish-Event ID</b>
100	Asphalt Cement	82	101	102
110	Bridge Deck Crack Seal	14	111	112
120	Bridge Deck Repair	30	121	122
130	Concrete Sidewalks	42	131	132
140	Conduits and Pull Boxes	24	141	142
150	Cover	29	151	152
160	Crushed Aggregate Course	39	161	162
170	Curbs and Gutters	55	171	172
180	Drainage Pipe	22	181	182
190	Emulsified Asphalt	42	191	192
200	Excavation and Embankment	31	201	202
210	Fences	19	211	212
220	Final Sweep and Broom	28	221	222
230	Geotextiles	30	231	232
240	Guardrail and Concrete Barrier Rail	109	241	242
250	Milling and Pulverization	76	251	252
260	Pavement Marking Application	187	261	262
270	Plant Mix Surfacing	100	271	272
280	Remove Pipe Culvert	10	281	282
290	Remove, Reset, And Adjust Facilities	24	291	292
300	Revise Bridge Rail	24	301	302
310	Roadside Re-Vegetation	49	311	312
320	Shoulder Gravel	26	321	322
330	Signs and Delineators	142	331	332
340	Special Borrow	18	341	342
350	Temporary Erosion Control	26	351	352
360	Topsoil-Salvaging and Placing	10	361	362
370	Traffic Control	190	371	372
380	Traffic Signals and Lighting	29	381	382

#### **4.6.2. Step 2: Develop a sequence database of the selected project work type from the DWR data**

Each activity identified in Step 1 was given an activity ID, as shown in the first column of Table 4.4. The start and finish events of each activity were also given event

IDs, the activity ID added by 1 and 2, respectively, as shown in Table 4.4. For each project, relevant activities and events were identified. Each event's date was calculated based on the linked work items' recorded dates in the DWR data. Each project was then transformed into an ordered sequence of events based on the calculated event dates. For example, project 4071005000 started with event 371 (the start of activity "Traffic Control") and ended with event 262 (the finish of activity "Pavement Marking Application"). At the end of Step 2, a sequence database of 190 sequences/projects was available for further analysis.

#### **4.6.3. Step 3: Apply an SPM algorithm to the sequence database to extract all two-event subsequences and their occurrence frequencies**

All two-event subsequences, along with their support, were obtained by applying the PrefixSpan algorithm in the Sequential Pattern Mining Framework (SPMF) library (Fournier-Viger et al. 2017). Table 4.5 shows several examples of the obtained subsequences. A pair of two events  $A$  and  $B$  can have at most three subsequences:  $\langle \{A\}, \{B\} \rangle$  (event  $B$  happened at a later day than event  $A$ ),  $\langle \{A, B\} \rangle$  (events  $A$  &  $B$  happened at the same day), and  $\langle \{B\}, \{A\} \rangle$  (event  $A$  happened at a later day than event  $B$ ). For instance, pairs 1, 3, and 6 have one, two, and three subsequences.

The subsequence and support of pair 1 indicate that 22 projects in the database contained both activities "Asphalt Cement" and "Cover" and that event 151 (the start of "Cover") happened at a later day than event 101 (the start of "Asphalt Cement") in all 22 projects. In other words, there was a Start-Start relationship between the two activities. The start and finish events of the same activity (e.g., pair 2) can have at most two

subsequences: the finish event happened on the same day with or later than the start event.

**Table 4.5.** Examples of pairs of events and their subsequences

<b>Pair No.</b>	<b>Pair of Events</b>	<b>Subsequence</b>	<b>Support (Projects)</b>
1	Event 101 (the start of “Asphalt Cement”) Event 151 (the start of “Cover”)	{101}, {151}	22
2	Event 101 (the start of “Asphalt Cement”) Event 102 (the finish of “Asphalt Cement”)	{101, 102} {101}, {102}	33 49
3	Event 101 (the start of “Asphalt Cement”) Event 171 (the start of “Curbs and Gutters”)	{101}, {171} {171}, {101}	14 5
4	Event 361 (the start of “Topsoil-Salvaging and Placing”) Event 371 (the start of “Traffic Control”)	{361}, {371} {371}, {361}	1 9
5	Event 251 (the start of “Milling and Pulverization”) Event 271 (the start of “Plant Mix Surfacing”)	{251, 271} {251}, {271} {271}, {251}	3 35 12
6	Event 252 (the finish of “Milling and Pulverization”) Event 271 (the start of “Plant Mix Surfacing”)	{252, 271} {252}, {271} {271}, {252}	3 24 23
7	Event 262 (the finish of “Pavement Marking Application”) Event 272 (the finish of “Plant Mix Surfacing”)	{262, 272} {262}, {272} {272}, {262}	7 10 82

**4.6.4. Step 4: Determine a representative sequential pattern for each pair of events from the two events’ subsequences**

The McNemar exact binomial test was applied to compare the support values of two possible contradictory subsequences of a given pair of events to choose the most probable one. The selected minimum required support was five, as some activities had a



frequency as low as ten (see Table 4.4). Table 4.6 shows the results of applying the proposed process of Step 4 to the examples in Table 4.5.

**Table 4.6.** Examples of pairs of events and their representative sequential patterns

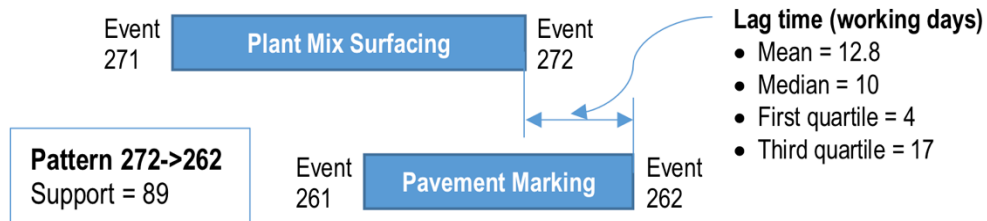
Pair No.	Event A	Event B	Support			p-value (McNemar test, $\alpha = 0.05$ )	Pattern	Probability
			{A, B}	{A}, {B}	{B}, {A}			
1	101	151		22		N/A	101->151	1
2	101	102	33	49		N/A	101->102	1
3	101	171		14	5	0.032 (S)	101->171	14/19
4	361	371		1	9	0.011 (S)	371->361	9/10
5	251	271	3	35	12	0.001 (S)	251->271	38/50
6	252	271	3	24	23	0.500	No pattern	
7	262	272	7	10	82	0.000 (S)	272->262	89/99

Note: S = Significant difference, N/A = Not applicable

Pairs 1 and 2 did not require the statistical test due to no contrary subsequences and the support values larger than the minimum threshold (i.e., 5). Pairs 3 to 7 required the test. The differences in the support values (between  $\langle\{A\}, \{B\}\rangle$  and  $\langle\{B\}, \{A}\rangle$ ) with these pairs were significant except for pair 6. As a result, no patterns were found for pair 6 of event 252 (the finish of “Milling and Pulverization”) and event 271 (the start of “Plant Mix Surfacing”). The pattern and probability of other pairs were given in the last two columns of Table 4.6.

The lag time for each pattern can also be calculated. Fig. 4.8 shows an example of pattern “Event 272 -> Event 262” or a Finish-Finish relationship between activity “Plant Mix Surfacing” and activity “Pavement Marking.” The median value of the lag

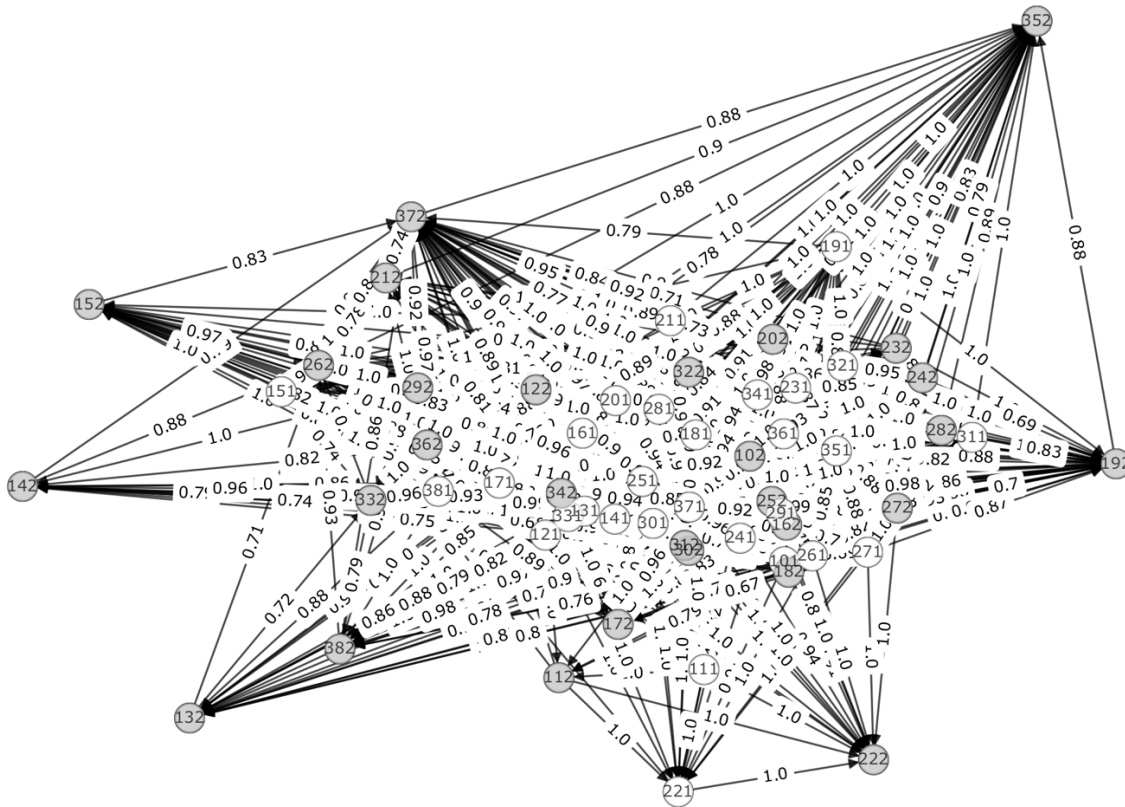
time of 89 projects containing the pattern was ten working days. At the end of this step, a total of 699 patterns were identified.



**Fig. 4.8.** An example of the lag time of a pattern

#### 4.6.5. Step 5: Interlink the identified sequential patterns and develop a network of as-built construction sequence patterns

The 699 identified patterns were used to develop a knowledge network of construction sequence patterns for the overlay project work type using a Python Package - NetworkX. A directed edge was used to illustrate a sequential relationship with the relationship's probability shown on the edge. Start and finish events were respectively represented by white and grey nodes with an event ID as a node number to visually show different types of patterns: Start-Start, Start-Finish, Finish-Start, and Finish-Finish relationships. Fig. 4.11 is the network of the 699 identified patterns, including 58 nodes representing the start and finish events of 29 activities and 699 edges representing 699 patterns.



**Fig. 4.9.** A network of construction sequence patterns of overlay projects

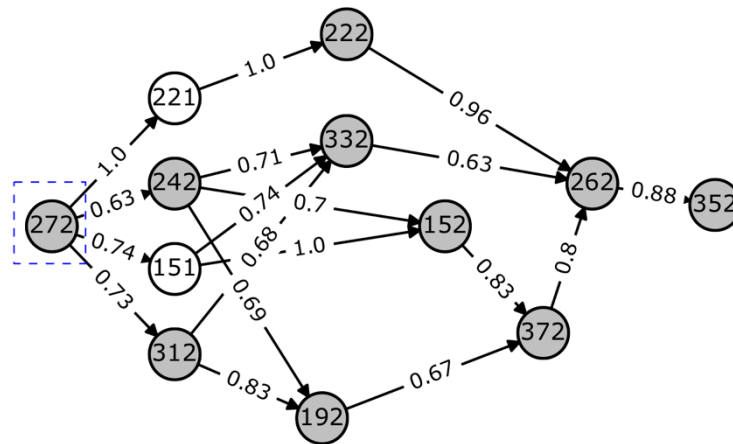
**4.6.6. Step 6: Implement algorithms to apply the developed network for sequencing a new project**

Below give examples of applying the proposed algorithms S1, S2, and S3 to the developed knowledge network in Step 5.

***Algorithm S1: Find the successors of an event***

Fig. 4.10 shows the result of applying algorithm S1 to find the successors of event 272, the finish of the plant mix surfacing activity, to answer the question: “what happens after the plant mix surfacing activity is finished?” The output of the algorithm is a simplified subnetwork of the network developed in Step 5. It contains event 272 and its

11 successors. Without the development and analysis of the network, SPM can still provide the list of successors. However, the subnetwork provides even more information than that. It suggests that four events, 221, 242, 151, and 312, have a more direct relationship with event 272 than the others due to the direct edge between each of them and 272. For example, edge 272 → 221 (probability = 1) indicates that in 100% of the past projects containing both the plant mix surfacing activity and the final sweep and broom activity, the finish of the plant mix surfacing activity was followed by the start of the final sweep and broom activity. In other words, there was a Finish-Start relationship between the two activities. Event 372 (i.e., the finish of the traffic control activity) is an example of an indirect successor. Even though 372 happened after event 272, other events occurred between them.

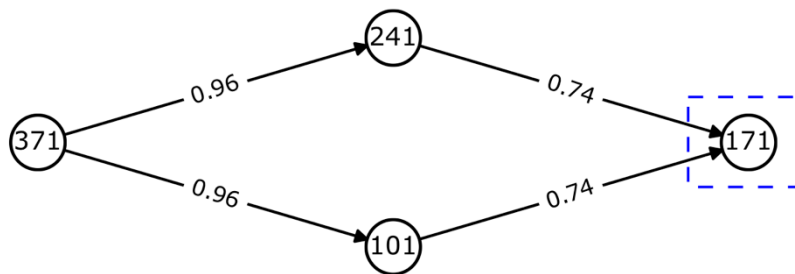


**Fig. 4.10.** Successors of event 272 – the finish of the pavement mix surfacing activity

**Algorithm S2: Find the predecessors of an event**

Fig. 4.11 shows the result of applying algorithm S2 to find the predecessors of event 171 (i.e., the start of the curbs and gutters activity) to answer the question: “What

happens before the start of the curbs and gutters activity?” The output of the algorithm is a simplified subnetwork of the network developed in Step 5. It contains event 171, its three predecessors, and sequential relationships between them. For example, edge 101 -> 171 (probability = 0.74) indicates that in 74% of the past projects containing both the asphalt cement activity and the curbs and gutters activity, the start of the curbs and gutters activity was after the start of the asphalt cement activity. In other words, there was a Start-Start relationship between the two activities.

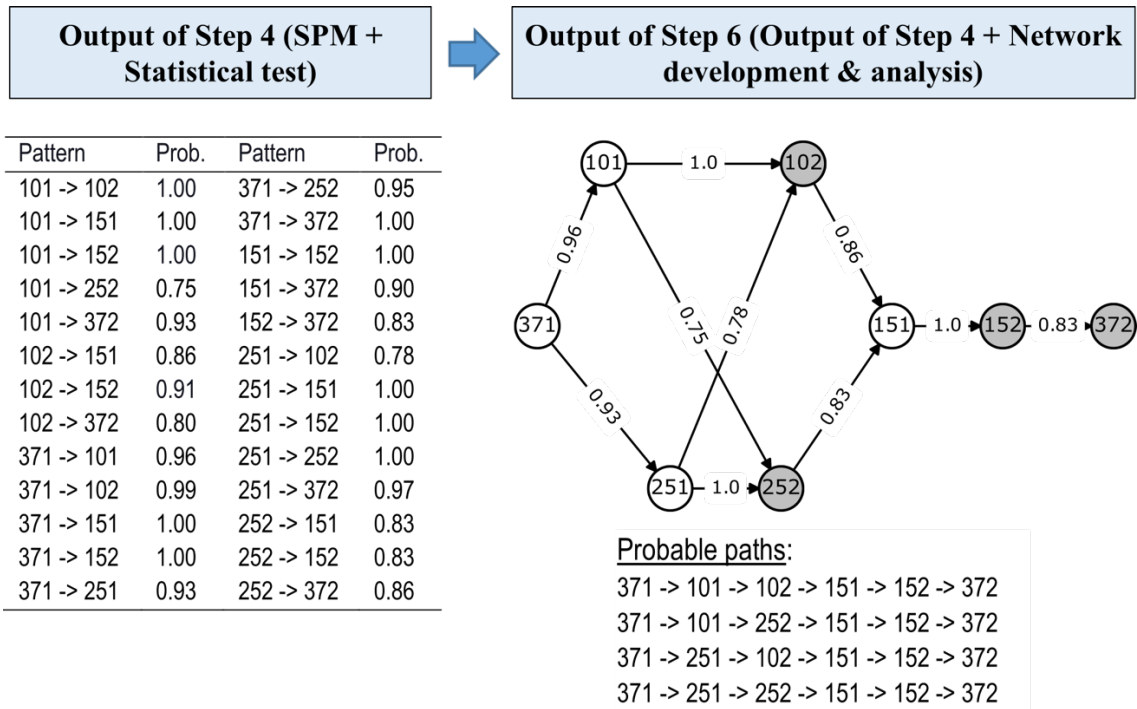


**Fig. 4.11.** Predecessors of event 171 – the start of the curbs and gutters activity

***Algorithm S3: Find sequential relationships among a set of activities of interest***

Fig. 4.12 shows sequential relationships among a sample set of four activities (i.e., 100 – Asphalt Cement, 150 – Cover, 250 – Milling and Pulverization, and 370 – Traffic Control) with four start events and four finish events. The left part of the figure results from applying SPM and statistical tests in Step 4: 26 separated pairwise relationships among the activity events. By applying network development and analysis to the output of Step 4, the 26 separated patterns were connected in a graph, as shown on the right part of the figure, with the exclusion of redundant patterns. Algorithm S3 also provides four probable paths from the first event (i.e., the start of the traffic control

activity) to the last event (i.e., the finish of the traffic control activity) of the four activities.



**Fig. 4.12.** Sequential relationships among a sample set of activities

#### 4.7. Discussion and practical implications

The case study results show that applying an SPM algorithm and statistical analysis to the DWR data can detect pairwise sequential patterns or logical relationships between construction activities. The relationship types used for sequencing are not affected by subjective expert judgments as in the current practice. Instead, there can be multiple logical relationships of different types (i.e., Start-Start, Start-Finish, Finish-Start, and Finish-Finish) between two activities. Unlike the deterministic logic templates developed in previous studies, the probability of occurrence associated with each

relationship indicates the flexibility and variability of construction sequencing due to the effect of influential factors on sequences and the existence of more than one right way to construct any project. As only one primary influential factor (i.e., project work type) was considered in the case study, the probabilities of the occurrence of some patterns were significantly smaller than 1 to reflect variations in construction sequences of a given project work type. The range of lag time for each pattern used in past projects can also be calculated. Lag time is not fixed but changes project-by-project.

However, a discovered logical relationship between two activities can become redundant in sequencing a project even if the project contains both activities. Suppose two sequential patterns,  $A \rightarrow B$  and  $B \rightarrow C$ , are detected from past projects' DWR data, then  $A \rightarrow C$  is also reported as a pattern. If a new project contains all  $A$ ,  $B$ , and  $C$ , pattern  $A \rightarrow C$  becomes redundant given  $A \rightarrow B$  and  $B \rightarrow C$ . If the project includes  $A$  and  $C$  but not  $B$ , then pattern  $A \rightarrow C$  is necessary. It is also challenging to determine whether an activity is an immediate predecessor or successor of another activity by looking at the list of separated discovered sequential patterns. The application of network modeling and analysis helps to deal with the mentioned issues and interlink the patterns/relationships while eliminating redundant relationships.

Since most DOTs have limited guidance on construction sequencing, the proposed approach can significantly enhance the sequencing practice by applying it to their digital DWR systems without collecting any additional data. The analysis of as-built construction data can help alleviate differences between sequences developed by DOTs for bidding and contracting purposes and actual orders of work performed by

contractors. The detected logical relationships provide a good start for DOT schedulers in logic development. Schedulers may still need to modify the provided relationships depending on specific project conditions. However, the probability associated with each relationship can provide schedulers, especially less experienced ones, with a confidence level in applying it to a new project. The proposed algorithms also enable a flexible and rapid application of the discovered relationships for future projects. A knowledge network of construction sequencing can be developed for each project work type. For the sequencing of a new project, the algorithms can be applied to the predeveloped network of a corresponding project work type by only providing the new project's list of activities as input. Furthermore, the relationships and their network can be easily updated periodically by applying the proposed approach to a newer DWR dataset.

#### **4.8. Summary and conclusions**

Production rate estimation and logic development are two primary components of estimating and determining the contract time of a highway construction project. While various methods and tools are available for the former task, limited guidance is available to the latter. A highway project's logic development requires a scheduler to have a thorough understanding of the project, scheduling knowledge, and site experience. The scheduler also needs to make reasonable and realistic assumptions on how the project will be constructed, while different contractors may have other ways to perform the required work. Therefore, information from the as-built sequences in past projects can be critically valuable to improving the efficiency and reliability of the owner agency's scheduling and contract time determination process. Most DOTs have as-built data such



as DWRs available but not leveraged for this purpose since the current practice is mostly experience and judgment based.

This study's primary contribution to the body of knowledge is the development of a data-driven process model that can detect as-built pairwise logical relationships in past projects and build construction logic knowledge networks. With a novel application of network theory, each relationship is visually and dynamically represented, along with the probability of the relationship's occurrence showing the certainty level associated with it. The discovered relationships of each project work type are interlinked together to form a unified reference source. Three algorithms are proposed to extract only relevant relationships to a new project, eliminate redundant ones that can be inferred from others, and return a corresponding logic network.

A limitation of this study is the consideration of only one primary influential factor in construction sequencing (i.e., project work type) due to data availability issues. However, the proposed approach's first two steps can be easily adjusted to consider more influential factors on sequencing, such as project phasing and construction methods, if data are available. A collection of data on the additional factors can make knowledge networks and logical relationships more project-specific. However, this required extra effort may hinder some DOTs from applying this approach for enhancing their sequencing practices.

#### 4.9. References

- Abdel-Raheem, M., Torres Cantu, C., and Wang, X. (2020). "Dynamic Contract Time Determination System for Highway Projects." *Transportation Research Record: Journal of the Transportation Research Board*, 2674(5), 381-392.
- ADOT (2015). "Chapter 8: Estimating Contract Time." Arizona Department of Transportation, Arizona.
- Alikhani, H., Le, C., and Jeong, H. D. "Deep Learning Algorithms to Generate Activity Sequences Using Historical As-built Schedule Data." *Proc., Creative Construction e-Conference 2020*, Budapest University of Technology and Economics, 2-6.
- Bruce, R. D., Slattery, D. K., Slattery, K. T., and McCandless, D. (2012). "An Expert Systems Approach to Highway Construction Scheduling." *Technology Interface International Journal*, 13(1), 21-28.
- Carson, C., Oakander, P., and Relyca, C. (2014). "CPM scheduling for construction— Best practices and guidelines." *Project Management Institute, Newton Square, PA*.
- CDOT (2019). "Construction Manual – Section 100: General Provisions." Colorado Department of Transportation, Colorado.
- Cherneff, J., Logcher, R., and Sriram, D. (1991). "Integrating CAD with Construction-Schedule Generation." *Journal of Computing in Civil Engineering*, 5(1), 64-84.

- Chevallier, N. J., and Russell, A. D. (2001). "Developing a Draft Schedule Using Templates and Rules." *Journal of Construction Engineering and Management*, 127(5), 391-398.
- Chua, D. K. H., Nguyen, T. Q., and Yeoh, K. W. (2013). "Automated construction sequencing and scheduling from functional requirements." *Automation in Construction*, 35, 79-88.
- FHWA (2002). "FHWA Guide for Construction Contract Time Determination Procedures." Federal Highway Administration, Washington, D.C.
- Fischer, M. A., and Aalami, F. (1996). "Scheduling with Computer-Interpretable Construction Method Models." *Journal of Construction Engineering and Management*, 122(4), 337-347.
- Florez, L. (2017). "Crew Allocation System for the Masonry Industry." *Computer-Aided Civil and Infrastructure Engineering*, 32(10), 874-889.
- Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., and Thomas, R. (2017). "A survey of sequential pattern mining." *Data Science and Pattern Recognition*, 1(1), 54-77.
- Hagberg, A., Swart, P., and S Chult, D. (2008). "Exploring network structure, dynamics, and function using NetworkX." Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
- Hancher, D. E., McFarland, W., and Alabay, R. T. (1992). "Construction contract time determination."
- Hinze, J. (2012). *Construction planning and scheduling*, Pearson/Prentice Hall.

- Hu, D., and Mohamed, Y. (2014). "A dynamic programming solution to automate fabrication sequencing of industrial construction components." *Automation in Construction*, 40, 9-20.
- Idaho DOT (2011). "Contract Time Determination in Project Development." Idaho Department of Transportation, Idaho.
- IDOT (2017). "Chapter Sixty-six: Contract Processing, Bureau of Design and Environment Manual." Illinois Department of Transportation, Illinois.
- Jang, Y., Jeong, I.-B., Cho, Y. K., and Ahn, Y. (2019). "Predicting Business Failure of Construction Contractors Using Long Short-Term Memory Recurrent Neural Network." *Journal of Construction Engineering and Management*, 145(11).
- Jeong, H. D., Le, C., and Devaguptapu, V. (2019). "Effective Production Rate Estimation Using Construction Daily Work Report Data." Montana. Dept. of Transportation. Research Programs.
- Jeong, H. S., Atreya, S., Oberlender, G. D., and Chung, B. (2009). "Automated contract time determination system for highway projects." *Automation in Construction*, 18(7), 957-965.
- Le, C., and Jeong, H. D. (2020a). "Artificial Intelligence Framework for Developing a Critical Path Schedule Using Historical Daily Work Report Data." *Construction Research Congress 2020*, 565-573.
- Le, C., and Jeong, H. D. (2020b). "A Daily Work Report Based Approach for Schedule Risk Analysis." *CIGOS 2019, Innovation for Sustainable Infrastructure*, Springer, 1131-1136.

- Le, C., Jeong, H. D., Le, T., and Kang, Y. (2020). "Evaluating Contractors' Production Performance in Highway Projects Using Historical Daily Work Report Data." *Journal of Management in Engineering*, 36(3), 04020015.
- Lim, T.-K., Yi, C.-Y., Lee, D.-E., and Arditi, D. (2014). "Concurrent Construction Scheduling Simulation Algorithm." *Computer-Aided Civil and Infrastructure Engineering*, 29(6), 449-463.
- Liu, H., Al-Hussein, M., and Lu, M. (2015). "BIM-based integrated approach for detailed construction scheduling under resource constraints." *Automation in Construction*, 53, 29-43.
- MassDOT (2014). "Construction Contract Time Determination Guidelines for Designers/Planners." Massachusetts Department of Transportation, Massachusetts.
- McCrary, S. W., Corley, M., Leslie, D. A., and Aparajithan, S. (1995). "Evaluation of contract time estimation and contracting procedures for louisiana department of transportation and development construction projects: final report: volume I." Louisiana Tech University. Dept. of Civil Engineering.
- Mooney, C. H., and Roddick, J. F. (2013). "Sequential pattern mining -- approaches and algorithms." *ACM Computing Surveys*, 45(2), 1-39.
- Mubarak, S. A. (2015). *Construction project scheduling and control*, John Wiley & Sons.
- NCDOT (n.d.). "Guidelines for Determining Contract Time." North Carolina Department of Transportation, North Carolina.

- Newitt, J. S. (2009). *Construction Scheduling: Principles and Practices*, Pearson/Prentice Hall, New Jersey.
- Ott, R. L., and Longnecker, M. (2015). *An Introduction to Statistical Methods and Data Analysis*, Cengage Learning.
- Shrestha, K. J., and Jeong, H. D. (2017). "Computational algorithm to automate as-built schedule development using digital daily work reports." *Automation in Construction*, 84, 315-322.
- Shrestha, K. J., Le, C., Jeong, H. D., and Le, T. "Mining Daily Work Report Data for Detecting Patterns of Construction Sequences." *Proc., Creative Construction Conference 2019*, 578-583.
- Tao, S., Wu, C., Hu, S., and Xu, F. (2020). "Construction project scheduling under workspace interference." *Computer-Aided Civil and Infrastructure Engineering*.
- Taylor, T. R. B., Sturgill, R. E., and Li, Y. (2017). *Practices for Establishing Contract Completion Dates for Highway Projects*.
- TxDOT (2018). "Contract Time Determination Guidance." Texas Department of Transportation, Texas.
- Wang, Y., and Yuan, Z. (2017). "Research on BIM-Based Assembly Sequence Planning of Prefabricated Buildings." *ICCREM 2017*, 10-17.
- Werkmeister, R. F., Luscher, B. L., and Hancher, D. E. (2000). "Kentucky Contract Time Determination System." *Transportation Research Record*, 1712(1), 185-195.

## **5. PARETO PRINCIPLE IN SCOPING-PHASE COST ESTIMATING: A MULTI-OBJECTIVE OPTIMIZATION APPROACH FOR SELECTING AND EVALUATING OPTIMAL MAJOR WORK ITEMS**

### **5.1. Overview**

Cost estimation is a critical part of a typical transportation project's development process. Specifically, project owners use cost estimates in the scoping phase to set project budgets for cost management. Due to the lack of detailed design information during scoping, State Transportation Agencies (STAs)' estimators typically apply the Pareto principle in cost estimates by estimating only major high-impact work items and taking the remaining items into account by a percentage. However, the 80/20 rule is a rough rule of thumb, which does not apply to every scenario. Definitive information on the determination of major work items and their application to estimating is necessary to ensure the reliability of the resulting estimates. Nevertheless, STAs' guidance is minimal, and few studies have investigated this research topic. This study proposes a novel application of well-established multi-objective optimization methods to discovering new knowledge about optimal major work items for cost estimation, their contribution to total project cost and relative variation, and the Pareto principle approach's error. The proposed approach applies to different STAs' work breakdown structures and project work types and is illustrated by actual historical bid tabulation data from an STA.

## 5.2. Introduction

Cost estimating is a crucial part of the development process of a transportation project that typically includes four main phases: 1) Planning, 2) Scoping, 3) Preliminary Design, and 4) Final Design (AASHTO 2013; ITD 2020). In the planning phase, needs for new projects are identified and prioritized. Planning-phase cost estimates are necessary to estimate potential funding needs and be one of the main criteria for comparing alternatives (PennDOT 2018). The most critical project needs progress to the scoping phase, where the projects' definitions become clearer with input from various functional groups and stakeholders (MnDOT 2008). Based on a project's definition, a scoping-phase cost estimate is developed and included in the project's definition (MnDOT 2008). If the project is admitted to a transportation improvement program or, in other words, is programmed, the scoping-phase estimate becomes the project's baseline cost (WSDOT 2015). Subsequently, estimates are developed in the preliminary design phase to manage project costs against approved budgets (AASHTO 2013). Plans, Specifications, and Estimates (PS&E) estimating is required in the final design phase for evaluating bids (AASHTO 2013).

Therefore, the reliability of cost estimates affects State Transportation Agencies (STAs) at both agency and project levels (AASHTO 2013; Elmousalami 2020). First, timely project completion within budget is critical as it directly influences public STAs' public image and public satisfaction (Zhang et al. 2017). Second, unreliable cost estimating affects budget-related communications, budgeting decisions, and the use of agencies' resources negatively (WSDOT 2015). For example, too conservative estimates



lead to fewer projects being developed, or conversely, too low forecasts can result in project cost overruns during construction (Gardner et al. 2017). Specifically, for a specific project, the scoping phase's cost estimate is essential. The project owner considers it the baseline to set the project budget for cost management; cost estimates in later stages are compared against it (WSDOT 2015).

Due to the difference in project maturity levels in different project development phases, the methods used for scoping-phase cost estimating are not the same as those used for planning-phase or PS&E estimating. For planning-phase cost estimating, STAs typically use simple parametric methods such as applying cost per parameter (e.g., dollars per centerline mile or square foot of bridge deck) of past similar projects for their estimations of new projects due to minimal available project information and scope definition (AASHTO 2013; MnDOT 2008). Numerous research studies have also developed statistical modeling- or advanced artificial intelligence-based parametric models for predicting total project construction cost to improve estimation accuracy and overcome the lack of work item-level information (Elmousalami 2020; Gardner et al. 2017; Karaca et al. 2020; Zhang et al. 2017). Work item-level information (e.g., approximate quantities of major items) becomes available or can be reasonably determined from the scoping phase, and STAs often use the historical bid-based method to estimate work items' costs (AASHTO 2013; ITD 2020). Unlike the final design phase, estimators in the scoping stage (about 10% to 30% of project definition completed) do not have detailed design drawings to determine all work items entailed in a project and their quantities (WSDOT 2015). They typically focus on high-cost impact work items, as

suggested by the Pareto principle, that approximately 20% of the work items comprise 80% of a project's total cost (Olumide et al. 2010; PennDOT 2018; TxDOT n.d.). A percentage or minor item allowance is used to consider the remaining work items (CTDOT 2019; PennDOT 2018).

STAs' application of the Pareto principle (a.k.a. the 80/20 rule) to scoping-phase cost estimating faces challenges. First, the numbers 80 and 20 do not necessarily hold in cost estimating, and the guidelines are missing or vary across STAs. While some STAs state that 20% of the work items can account for 80% of the cost, others have different ideas about the contribution of the 20% items, such as 70% (Iowa DOT 2012; TxDOT n.d.). Second, the cost contributions of the same set of major items in even similar projects are expected to fluctuate, but little is currently known about the variation. Third, STAs do not have detailed guidance on which items should be used for estimating in the scoping phase and rely on estimators' judgments and experiences to select major items for their estimation. Fourth, high-cost impact work items vary with project work types and work-item breakdown structures, but STAs' guidance or research studies on these dynamics are very limited. Last, the error of applying the Pareto principle to estimating only major items' unit prices and accounting for minor items by a percentage compared with estimating all work items in a project has not been investigated. All of the issues mentioned above can significantly affect the accuracy of scoping-phase cost estimates.

A few studies have applied the Pareto principle to cost estimating but only identified the major work items or influential factors affecting total project costs (Le et al. 2019; Sayed et al. 2020; Shehab and Meisami-Fard 2013), which is not sufficient to

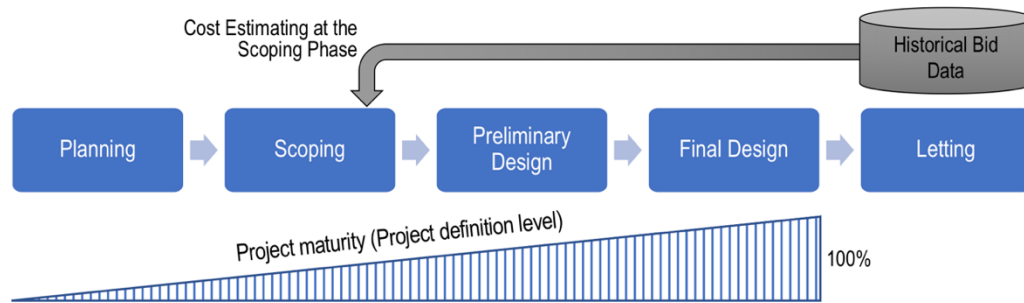
address the issues. For example, Le et al. (2019) identified the top five work items for unit price visualization, and Sayed et al. (2020) determined the nine most critical influential factors from 29 factors influencing construction cost estimates (e.g., site conditions and estimators' experience).

In this research, historical bid data (i.e., cost estimates submitted by bidders for past projects in the letting phase) are leveraged to address the issues. A past project's bid data consist of all work items identified by designers/estimators along with item quantities calculated from detailed plans at the end of the final design phase, which includes both major and minor work items. Also, unit prices are available and enable the investigation of the cost contributions of different work items.

As numerous sets of major items can be used, selecting an optimal set is desirable. Apart from accuracy, another critical aspect of cost forecasting is the amount of effort spent on developing cost estimates (Cao et al. 2018) because of the limited time allowed for estimating (Alroomi et al. 2012; ITD 2020). In this research, a novel application of multi-objective optimization methods is proposed to automatically find optimal major work items for cost estimation for different project work types and work-item breakdown structures. Optimization objectives include 1) maximizing the mean or the median of cost percentages of major items over total project cost (e.g., maximizing the average cost contribution of top 20% items: it is 80% or much higher), 2) minimizing the coefficient of variance of the percentages for uncertainty reduction, and 3) minimizing the error of applying the Pareto principle in scoping-phase cost estimating. Comparisons between different numbers of major work items are also conducted.

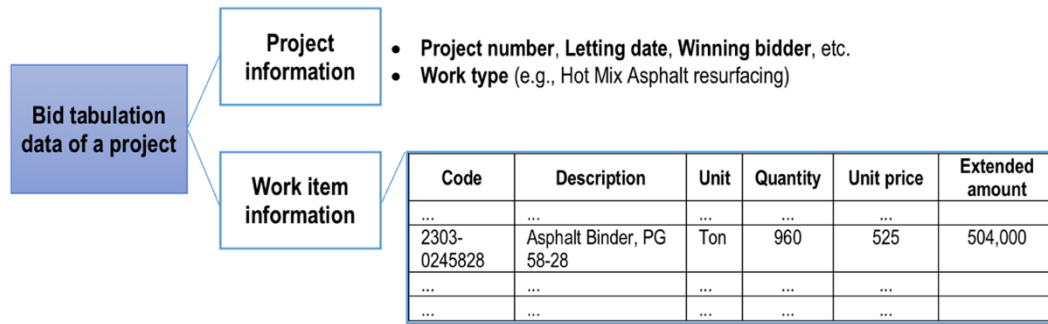
### 5.3. Research scope and objective

Cost estimates are necessary for each development phase (AASHTO 2013). Due to the differences in the amount of input information available for cost estimating and the purpose and required accuracy of the estimates, the methodologies adopted for cost estimation are different between the development phases (PennDOT 2018; WSDOT 2015). This research focuses on cost estimating at the scoping phase due to its importance to budget approval and project cost management (see Fig. 5.1).



**Fig. 5.1.** Timing of scoping-phase cost estimating in the project development phases

STAs often apply the Pareto principle to scoping-phase cost estimating (ITD 2020; Olumide et al. 2010). Specifically, they use the historical bid-based estimating method for major quantifiable work items and then apply a percentage to account for the remaining minor items. Historical bid-based estimating is an approach that relies on the bid tabulation data of similar projects in past recent years to estimate unit prices for a new project with possible modifications by estimators to account for unique project characteristics (Le et al. 2019) (see Fig. 5.2). However, STAs' guidance on applying the Pareto principle is minimal (see Table 5.1). Issues mentioned in the Introduction can significantly affect estimating accuracy.



**Fig. 5.2.** Data attributes of historical bid data

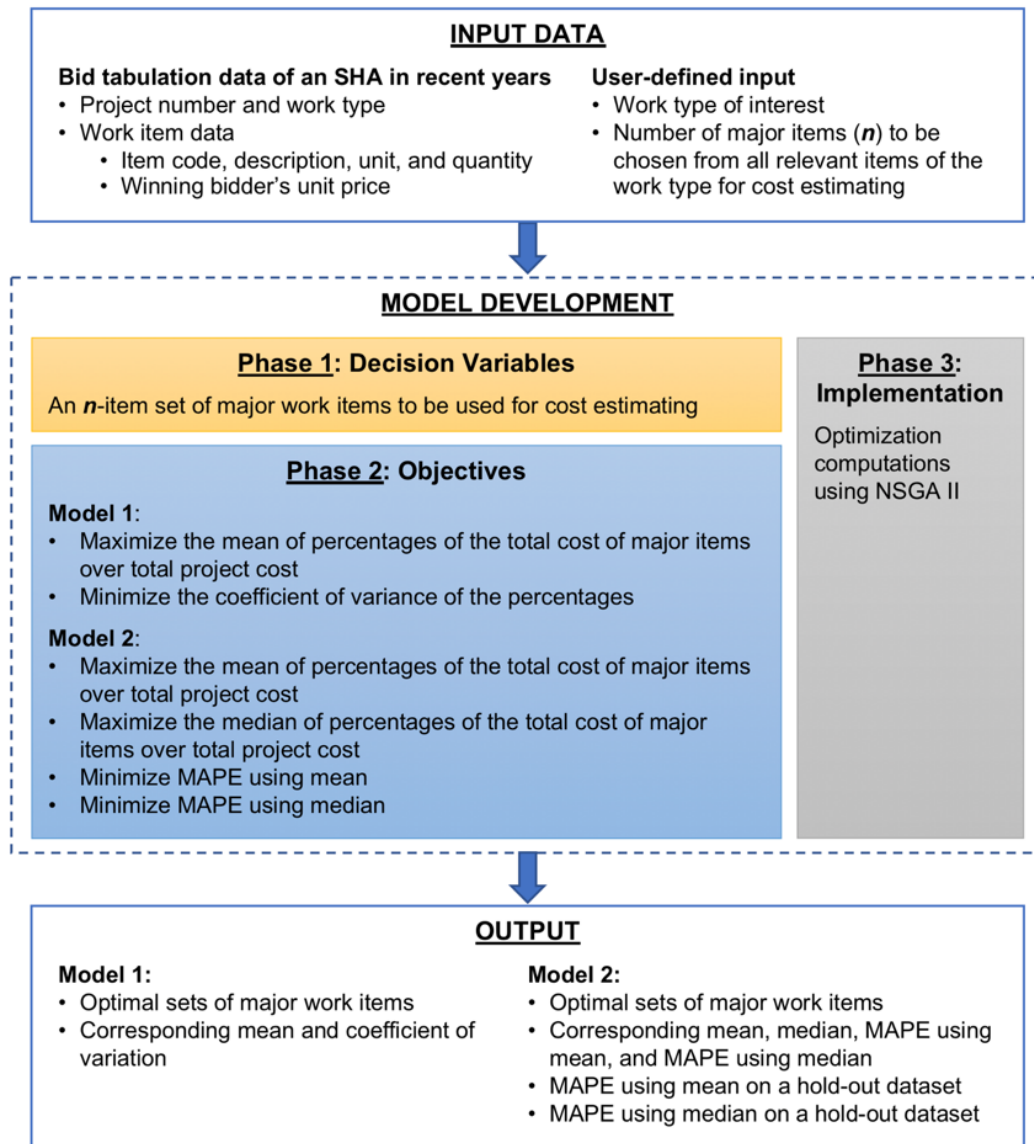
**Table 5.1.** STAs' guidance on applying the Pareto principle to cost estimating

No.	Guidance	Project phase	Reference
1	Minor items: 20 – 30% of the cost	Scoping	WSDOT (2015)
2	Minor items: 15 – 30% of the major item cost	Scoping	(CTDOT 2019)
3	20% of work items: 70% of the cost	Early phases	(Iowa DOT 2012)
4	20% of work items: 80% of the cost	Scoping	(MnDOT 2008)
5	20% of work items: 80% of the cost	Not specified	(PennDOT 2018)
6	20% of work items: 80% of the cost	Scoping	(TxDOT n.d.)
7	Major items: 65 – 85% of the cost	Not specified	(MDT 2016)
8	Major items: most of the cost	Not specified	(ITD 2020)

The objective of this paper is to develop an innovative multi-objective approach for selecting optimal major work items for cost estimating in the scoping phase by leveraging STAs' currently available historical bid tabulation data with the considerations of different project work types, work-item breakdown structures, and multiple objectives. For each optimal set of major work items, definitive and relevant information to apply the itemset for future projects is determined. The expected error of using the Pareto principle with the itemset is also discovered.

## 5.4. Methodology

This section presents the development of two multi-objective optimization models to support the application of the Pareto principle to cost estimating in the scoping phase. Fig. 5.3 gives an overview of the proposed models, including three main components: 1) Input data, 2) Model development, and 3) Model output.



**Fig. 5.3.** Proposed multi-objective optimization models

### 5.4.1. Input data

Historical bid data in recent years of an STA are used as input of the proposed models. Each project's data attributes used in the models include the project work type and the winning bidder's extended amounts of all work items in the project (see Fig. 5.2). Additionally, two user-defined input variables are necessary, including 1) Project work type of interest and 2) Number of major items ( $n$ ) selected from all items relevant to the work type that need to be estimated.

- The project work type variable is needed because projects of different work types have significantly different lists of work items and cost distributions. Table 5.2 shows the top five work items of four different work types. The top items of Hot Mix Asphalt (HMA) resurfacing projects are entirely different from those of Portland Cement Concrete (PCC) pavement projects. Major work items used for cost estimating, therefore, vary with project work types.
- The number of major items ( $n$ ), on the other hand, is related to the amount of time and effort spent on scoping-phase cost estimating.

**Table 5.2.** Top five work items of four different work types

No.	Work type code	Work type description	Top five work items
1	1523	HMA resurfacing	<ol style="list-style-type: none"> <li>1. Asphalt binder, PG 58-28</li> <li>2. Asphalt binder, PG 64-22</li> <li>3. Asphalt binder, PG 64-28</li> <li>4. HMA mixture (300,000 ESAL), Intermediate</li> <li>5. Granular shoulders, Type B</li> </ol>
2	1524	HMA resurfacing with mill	<ol style="list-style-type: none"> <li>1. Asphalt binder, PG 58-28</li> <li>2. Asphalt binder, PG 64-22</li> <li>3. HMA mixture (3,000,000 ESAL), Surface course</li> <li>4. Asphalt binder, PG 64-28</li> <li>5. Pavement scarification</li> </ol>
3	1525	HMA resurfacing/Cold in-place recycled	<ol style="list-style-type: none"> <li>1. Asphalt binder, PG 58-28</li> <li>2. HMA mixture (1,000,000 ESAL), Surface course</li> <li>3. Asphalt stabilizing agent (Foamed asphalt)</li> <li>4. Cold in-place recycled asphalt pavement</li> <li>5. Asphalt binder, PG 64-22</li> </ol>
4	1014	PCC pavement – Grade/New	<ol style="list-style-type: none"> <li>1. Standard or slip-form PCC pavement</li> <li>2. Mobilization</li> <li>3. Special backfill</li> <li>4. Excavation, Class 10, Roadway and borrow</li> <li>5. Removal of pavement</li> </ol>

Note: Top five work items in terms of the total extended amount of the item of all projects of the work type in a historical bid dataset

#### 5.4.2. Model development

The model development process consists of three phases: 1) Decision variables, 2) Objectives, and 3) Implementation.

##### 5.4.2.1. Phase 1: Decision variables

Assume the past projects of the work type of interest involve  $m$  work items (from *Item 1* to *Item m*);  $m$  can be hundreds. However, not all work items can be used for cost estimating in the scoping phase due to the lack of detailed design information and design plans. With the user-defined input  $n$ , the models need to identify optimal  $n$ -item sets



from  $m$  work items for future cost estimating in the scoping phase. The selection of  $n$  items from  $m$  work items is modeled by vector  $I$  [see Eq. (5.1) to (5.3)].

$$I = (I_1, I_2, I_3, \dots, I_m) \quad (5.1)$$

$$I_i = \begin{cases} 1, & \text{if Item } i \text{ is selected} \\ 0, & \text{if Item } i \text{ is not selected} \end{cases} \quad (5.2)$$

$$\sum_{i=1}^m I_i = n \quad (5.3)$$

#### 5.4.2.2. Phase 2: Objectives

Assume there are  $k$  projects of the work type of interest in the input bid tabulation data (from *Prj 1* to *Prj k*). Given a set of  $n$  items, the cost percentage of the work items in the itemset over total project cost in each project is calculated, forming a sample of  $k$  cost percentages for the  $k$  projects:  $P_j$  ( $j = 1, k$ ) [see Eq. (5.4) to (5.6)].

$$\text{Total project cost of Prj } j = PC_j = \sum_{i=1}^m EA_{ij} \quad (5.4)$$

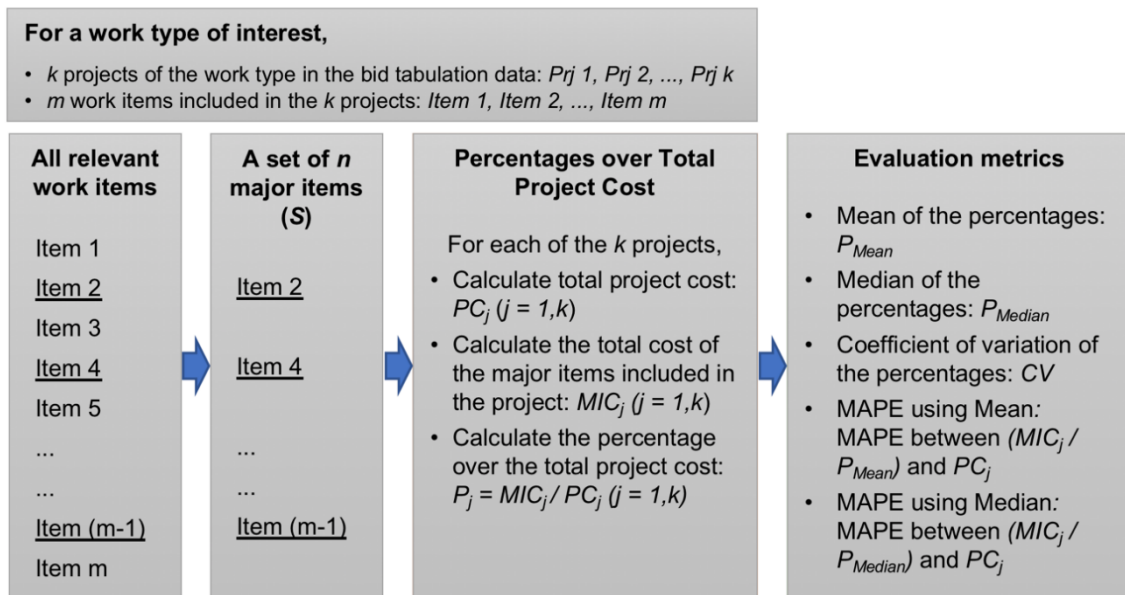
$$\text{Total cost of selected items in Prj } j = MIC_j = \sum_{i, I_i=1} EA_{ij} \quad (5.5)$$

$$\text{Cost percentage of selected items over total project cost} = P_j = \frac{MIC_j}{PC_j} \quad (5.6)$$

Where  $EA_{ij}$  = the extended amount of *Item i* in *Prj j*. Statistical measures of the cost percentages are then calculated (see Fig. 5.4).

- Mean of the percentages ( $P_{Mean}$ ): the average of the percentages, a measure of central tendency (Ott and Longnecker 2015).
- Median of the percentages ( $P_{Median}$ ): the middle value when the percentages are ordered from the lowest to the highest, another measure of central tendency. Compared to the mean, the median is less sensitive to skewness and outliers (Ott and Longnecker 2015).

- Coefficient of variance ( $CV$ ): the standard deviation divided by the mean of the percentages. While standard deviation is commonly used to measure population spread,  $CV$  is more appropriate in comparing the variability between populations (i.e., between different sets of work items) because it reflects variation over the baseline mean value (Ott and Longnecker 2015).



**Fig. 5.4.** Calculations of evaluation metrics

A set of  $n$  major work items is associated with two measures of the center of cost percentages:  $P_{Mean}$  and  $P_{Median}$ . Therefore, there are two strategies for applying the itemset to estimating the total cost of a new project.

- Strategy 1 involves calculating the total cost of the major items included in the project (denoted as  $MIC$ ) and then dividing  $MIC$  by  $P_{Mean}$  to obtain a total project cost estimate. The Mean Absolute Percentage Error (MAPE) between  $MIC/P_{Mean}$  values and total project costs reflects the error of applying Strategy

1 to cost estimating, compared with the estimation of all work items in a project. The resulted MAPE is called "MAPE using mean" in shorthand.

$$MAPE \text{ using mean} = \frac{1}{k} \sum_{j=1}^k \left| \frac{\frac{MIC_j - PC_j}{P_{Mean}}}{PC_j} \right| \quad (5.7)$$

- Strategy 2, similarly, involves calculating the total cost of the major items included in the project and then dividing it by  $P_{Median}$  to obtain a total project cost estimate. The MAPE between  $MIC/P_{Median}$  values and total project costs reflects the error of applying Strategy 2 to cost estimating, compared with the estimation of all work items in a project. The resulted MAPE is called "MAPE using median" in shorthand.

$$MAPE \text{ using median} = \frac{1}{k} \sum_{j=1}^k \left| \frac{\frac{MIC_j - PC_j}{P_{Median}}}{PC_j} \right| \quad (5.8)$$

Model 1 includes two objectives: 1) Maximizing  $P_{Mean}$  and 2) Minimizing  $CV$ .

This model is designed to examine whether the numbers 80 and 20 in the 80/20 rule hold in cost estimating and assess the variation of the ratio.

Model 2 includes four objectives: 1) Maximizing  $P_{Mean}$ , 2) Maximizing  $P_{Median}$ , 3) Minimizing MAPE using mean, and 4) Minimizing MAPE using median. The model is designed to examine the errors of applying the Pareto principle to scoping-phase cost estimating and compare two application strategies: using mean (Strategy 1) or median (Strategy 2) to represent the cost contributions of major work items over total project cost in past projects.

### 5.4.2.3. Phase 3: Implementation

The models are implemented using the Non-Dominated Sorting Genetic Algorithm (NSGA-II) due to its popularity and capability to solve a variety of multi-objective optimization problems and its ability to consider all objectives simultaneously without the need to pre-define weights for the objectives (Deb et al. 2002). Examples of applying NSGA-II to construction decision-making problems are multi-objective scheduling and planning (El-Abbasy et al. 2017; Halabya and El-Rayes 2020; Jeong and Abraham 2006; Peralta et al. 2018), design optimization (Dino and Üçoluk 2017; Hyari et al. 2016), and optimal construction layout or work zone design and development (Abdelmohsen and El-Rayes 2016; Abdelmohsen and El-Rayes 2018; Schuldt and El-Rayes 2018). The models are developed with the support of the Distributed Evolutionary Algorithms in Python (DEAP) toolbox (Fortin et al. 2012).

The NSGA-II computations in the models include four primary tasks:

- 1) An initialization task that randomly creates an initial population of sets of  $n$  work items from all work items relevant to the project work type of interest [see Eq. (5.1) to (5.3)],
- 2) A fitness evaluation task that calculates model evaluation metrics for each generated  $n$ -item set (see Fig. 5.4),
- 3) A ranking task that sorts the item sets using nondomination ranks and crowding distances (Deb et al. 2002), and
- 4) An evolution task of generating new populations with selection, crossover, and mutation operations.

Tasks 2 to 4 iterate until a stop criterion (e.g., a maximum number of iterations) is met.

### 5.4.3. Model output

Given the work type of interest and the user-defined number of major items ( $n$ ) for cost estimating, the output of Model 1 is optimal  $n$ -item sets, trade-off solutions between maximizing  $P_{Mean}$  while minimizing  $CV$ . The item sets and their corresponding measures are provided for further comparisons and analyses.

Similarly, the output of Model 2 is optimal  $n$ -item sets with four defined objectives: 1) Maximizing  $P_{Mean}$ , 2) Maximizing  $P_{Median}$ , 3) Minimizing MAPE using mean, and 4) Minimizing MAPE using median. The MAPE values illustrate the errors of applying the Pareto principle to cost estimating.

However, the MAPE values from the output of Model 2 are calculated from the same projects used for optimization, which probably make the errors underestimated. Therefore, the optimal sets of  $n$  major work items are applied to the cost estimation of each of the projects in a hold-out dataset with the two defined strategies (i.e., Strategy 1: using mean and Strategy 2: using median). Assume there are  $l$  projects in the hold-out dataset [from  $Prj(k+1)$  to  $Prj(k+l)$ ].

$$MAPE \text{ using mean on the holdout dataset} = \frac{1}{l} \sum_{j=k+1}^{k+l} \left| \frac{\frac{MIC_j - PC_j}{P_{Mean}}}{PC_j} \right| \quad (5.9)$$

$$MAPE \text{ using median on the holdout dataset} = \frac{1}{l} \sum_{j=k+1}^{k+l} \left| \frac{\frac{MIC_j - PC_j}{P_{Median}}}{PC_j} \right| \quad (5.10)$$

The MAPE values on the hold-out dataset are expected to provide more realistic and reliable error estimates.

## 5.5. Data analysis and results

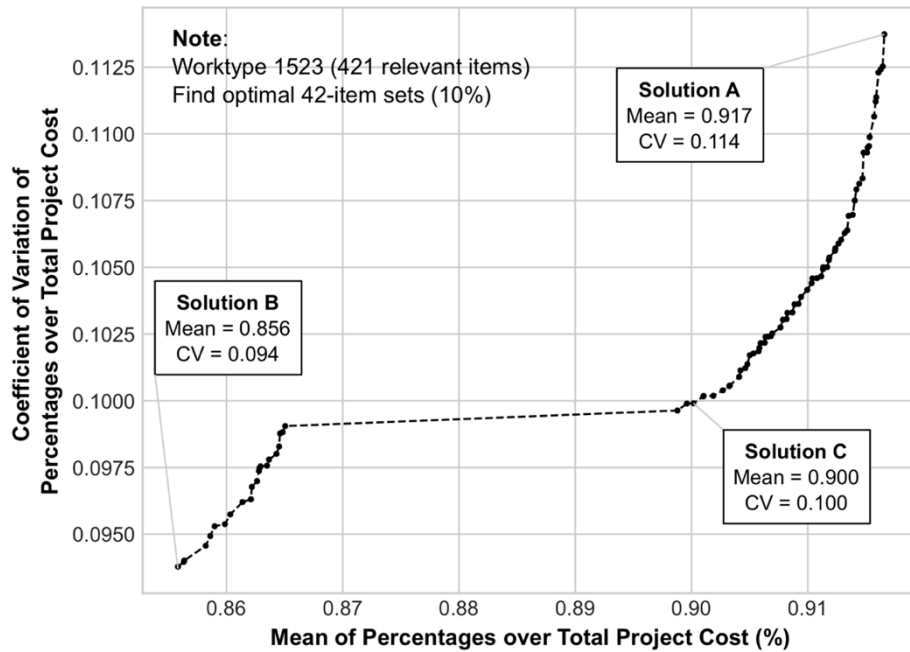
Bid tabulation data of 181 HMA resurfacing projects (Work type code: 1523) and 95 projects of the work type PCC pavement – Grade/New (Code: 1014) were obtained from an STA and used as input for the proposed models.

### 5.5.1. Model 1

The input data include 181 HMA resurfacing projects (Work type code: 1523). A total of 421 work items were used in the letting stage of these projects. According to the Pareto principle, STAs suggest estimating only high-cost impact work items in the scoping phase, only a portion of all relevant work items. As the proposed models allow users to define the number of major items they want to use for scoping-phase cost estimating, optimal sets of various numbers of work items can be obtained. Fig. 5.5 depicts a wide range of optimal solutions representing the optimal sets of 42 work items (10% of the total number of work items) for the estimating of HMA resurfacing projects.

On one end of the spectrum, Solution A represents an optimal 42-item set that results in the highest  $P_{Mean}$ . On average, the 42 items of Solution A account for 91.7% of the total project cost. However, that percentage is also associated with the highest variation among the generated solutions, which may not be a desirable feature from cost estimators' perspectives. Solution B corresponds to an optimal 42-item set at the other end of the spectrum, resulting in the lowest  $CV$  but a  $P_{Mean}$  value significantly lower than that of Solution A (i.e., 85.6% compared with 91.7%). Between the two ends of the spectrum, the model provides other trade-offs between the two defined objectives: 1) Maximizing  $P_{Mean}$  and 2) Minimizing  $CV$ . Of those, Solution C seems to be a

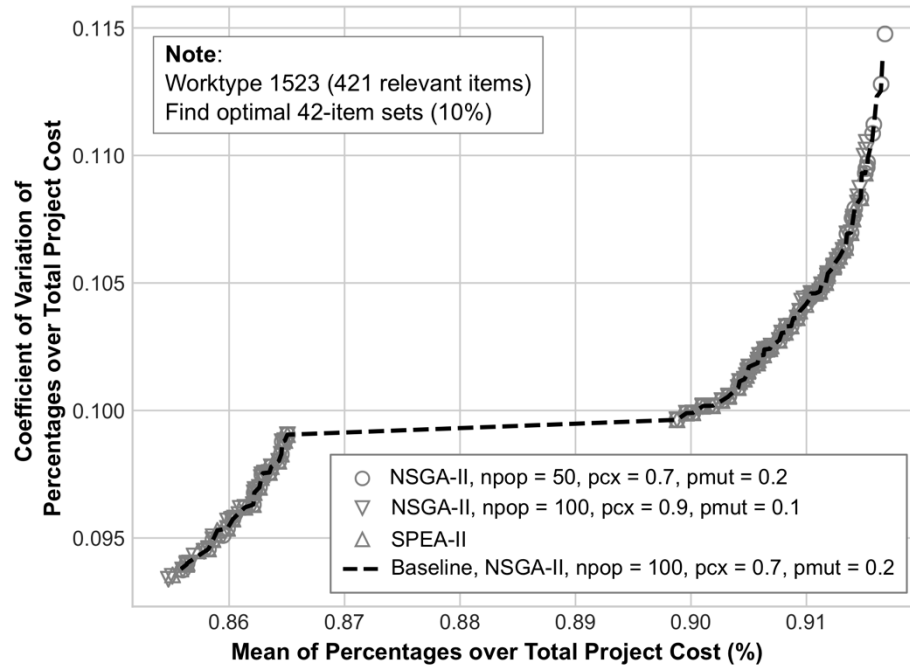
harmonious solution between Solution A and Solution B, using the elbow method. In fact, the ratio 90/10, not 80/20, applies to Solution C.



**Fig. 5.5.** Optimal trade-offs between mean and *CV* of cost percentages of a 42-major-item set over total project cost in HMA resurfacing projects (Work type 1523)

The solutions correspond to the following setting: population size ( $n_{pop}$ ) = 100, two-point crossover with the probability that an offspring is produced by crossover ( $p_{cx}$ ) = 0.7, two-point swapping mutation with the probability that an offspring is produced by mutation ( $p_{mut}$ ) = 0.2, and the maximum number of iterations = 2,000. The solutions are compared with the results of two other NSGA-II settings and the solutions obtained by another popular multi-objective optimization method, i.e., the Strength Pareto Evolutionary Algorithm II (SPEA-II). Fig. 5.6 shows that the solutions from the

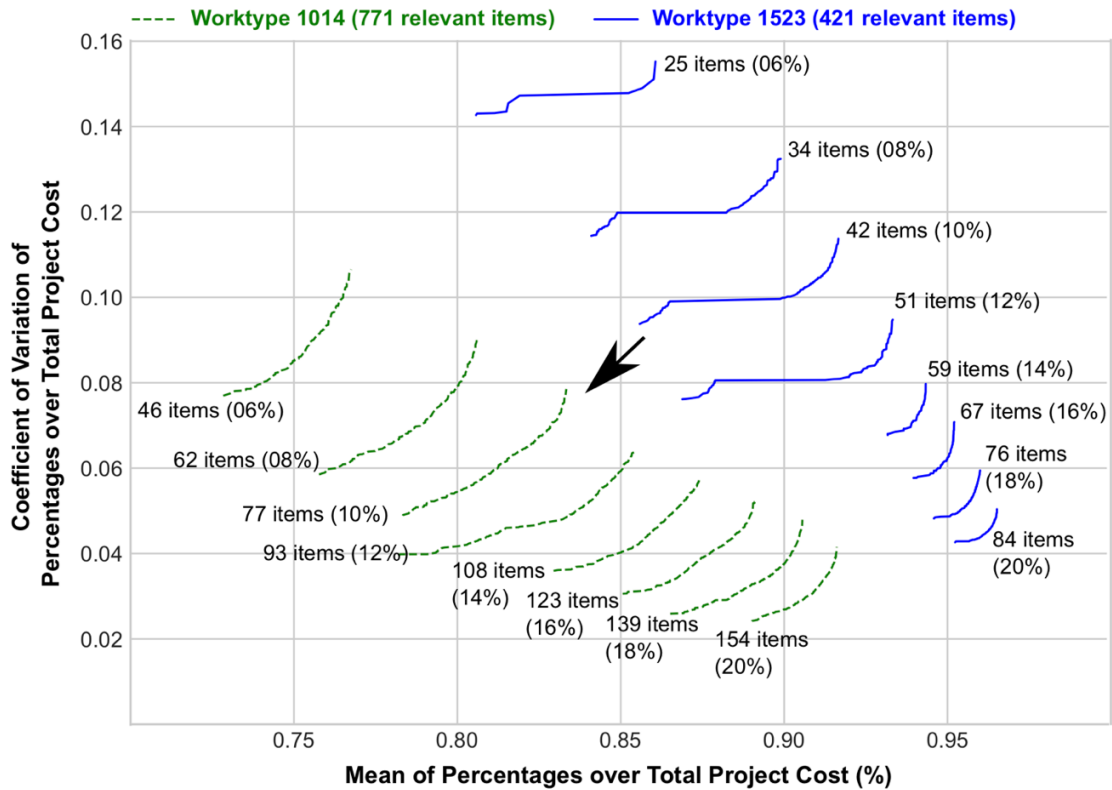
alternatives lie very close to the baseline, illustrating the quality and convergence of the baseline solutions.



**Fig. 5.6.** Convergence of optimal trade-offs solutions

When the number of major items used for scoping-phase cost estimating increases, the effort required for the estimation increases, which naturally results in improvements in both objectives (see Fig. 5.7). However, the improvements decrease as the number of major items increases due to the uneven cost distribution among work items. As shown in the figure, for HMA resurfacing projects (Work type code: 1523), the improvements in the objectives from 67 items to 76 items are significantly smaller than those from 25 items to 34 items. A similar trend applies to PCC pavement projects.





**Fig. 5.7.** Optimal trade-offs between mean and  $CV$  of cost percentages of a major item set over total project cost in different projects: A comparison between two work types (1523 — HMA resurfacing and 1014 — PCC pavement)

Fig. 5.7 also demonstrates the necessity of applying the Pareto principle to different project work types separately. While 421 items were used in the bid tabulation data of the past 181 HMA resurfacing projects, 771 items were used in those of 95 PCC pavement projects. The evaluation metrics (i.e.,  $P_{Mean}$  and  $CV$ ) of the two work types are also substantially different for the same ratio of major items. Take the ratio of 10% as an example. While 10% of the work items of HMA resurfacing projects can account for up to 91.7% of the total project cost on average, the counterpart only contributes up to an average of 83.3% of the total cost in their corresponding projects. Conversely, the variations of the cost percentages in PCC pavement projects are significantly smaller

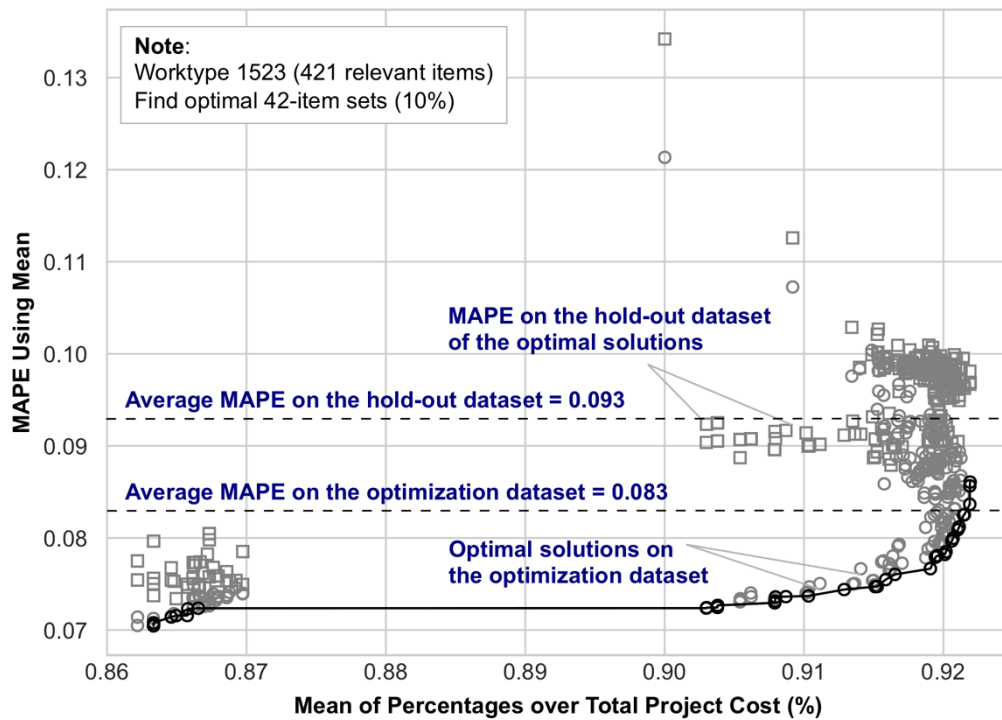
than that in HMA resurfacing projects for the same major item ratios (i.e., from 6% to 20% with an increment of 2%).

### 5.5.2. Model 2

For a specific set of major work items, the major items' contribution to total project cost still varies among projects, even with the projects of the same work type. A measure of the center of cost contributions (i.e., mean or median) is necessary to apply the itemset for future projects. Model 2 can enable comparison between using the mean or the median of the cost percentages in past projects for future estimating.

The bid tabulation data of the 181 HMA resurfacing projects were randomly divided into two datasets: optimization (75% of the projects) and hold-out (25% of the projects). With the optimization dataset as input, Model 2 can generate optimal solutions for different user-defined numbers of work items and their evaluation metrics (i.e.,  $P_{Mean}$ ,  $P_{Median}$ , MAPE using mean, and MAPE using median). The generated optimal sets of work items were subsequently applied to the hold-out dataset [see Eq. (5.9) to (5.10)]. For each optimal set, the total cost of its items in each hold-out project was calculated. It was then divided by the corresponding  $P_{Mean}$  or  $P_{Median}$  and then compared with the total project cost (i.e., the sum of the extended amounts of all work items in the project). Collectively, the relative differences were used to obtain MAPE using mean or MAPE using median on the hold-out dataset. The obtained MAPE values reflected the expected errors of applying the Pareto principle with the item set for future projects alone, not yet considering other factors influencing the accuracy of a cost estimate (e.g., inaccuracies in quantity takeoffs and unit price estimates).

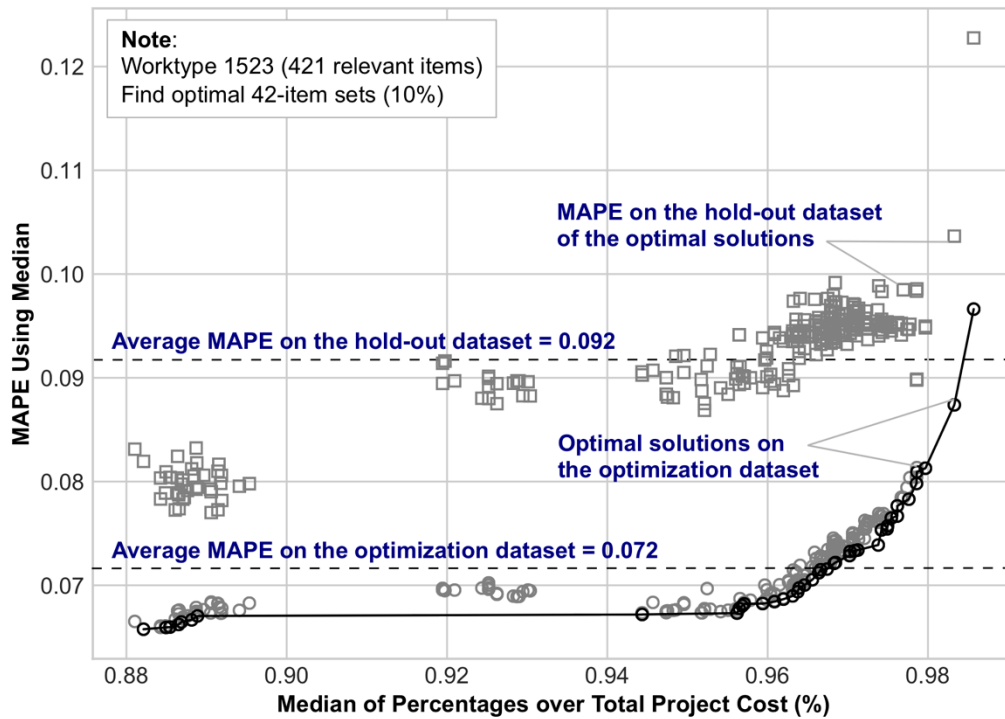
Fig. 5.8 shows the optimal 42-item solutions generated from Model 2 with the optimization dataset as input and their corresponding 1) mean and 2) MAPE using mean on the optimization dataset and 3) MAPE using mean on the hold-out dataset. MAPE values on the optimization dataset are generally smaller than MAPE values on the hold-out dataset, justifying the need for splitting the original data into two datasets as performed. The average error of using an optimal set of 42 major work items with mean as the center measure (i.e., Strategy 1) for scoping-phase cost estimating of HMA resurfacing projects alone is 9.3%.



**Fig. 5.8.** Optimal trade-offs between the mean of cost percentages of a 42-major-item set over total project cost in HMA resurfacing projects and MAPE using mean

Similarly, Fig. 5.9 shows the optimal 42-item solutions generated from Model 2 with the optimization dataset as input and their corresponding 1) median and 2) MAPE

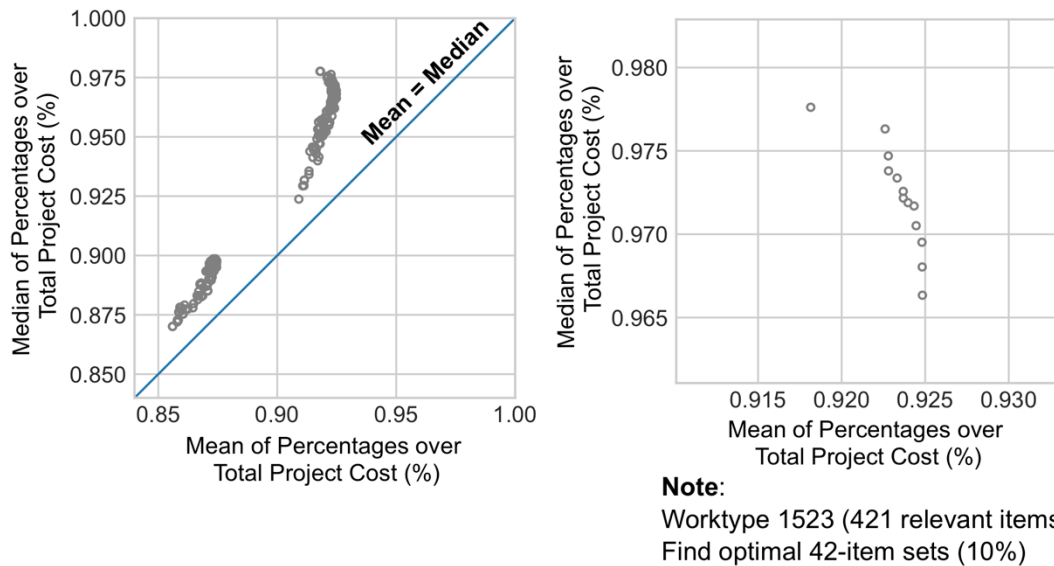
using median on the optimization dataset and 3) MAPE using median on the hold-out dataset. The average error of using an optimal set of 42 major work items with median as the center measure (i.e., Strategy 2) for scoping-phase cost estimating of HMA resurfacing projects alone is 9.2%.



**Fig. 5.9.** Optimal trade-offs between the median of cost percentages of a 42-major-item set over total project cost in HMA resurfacing projects and MAPE using median

For each scenario of user-defined input (i.e., the project work type of interest and the number of major items used for cost estimating), a comparison between MAPE using mean and MAPE using median is necessary to decide whether mean or median should better be used as the represented cost contribution of an optimal major-item set over total project cost. Fig. 5.10 provides a comparison between the mean and the median of the optimal 42-item sets of HMA resurfacing projects generated by Model 2. The right part

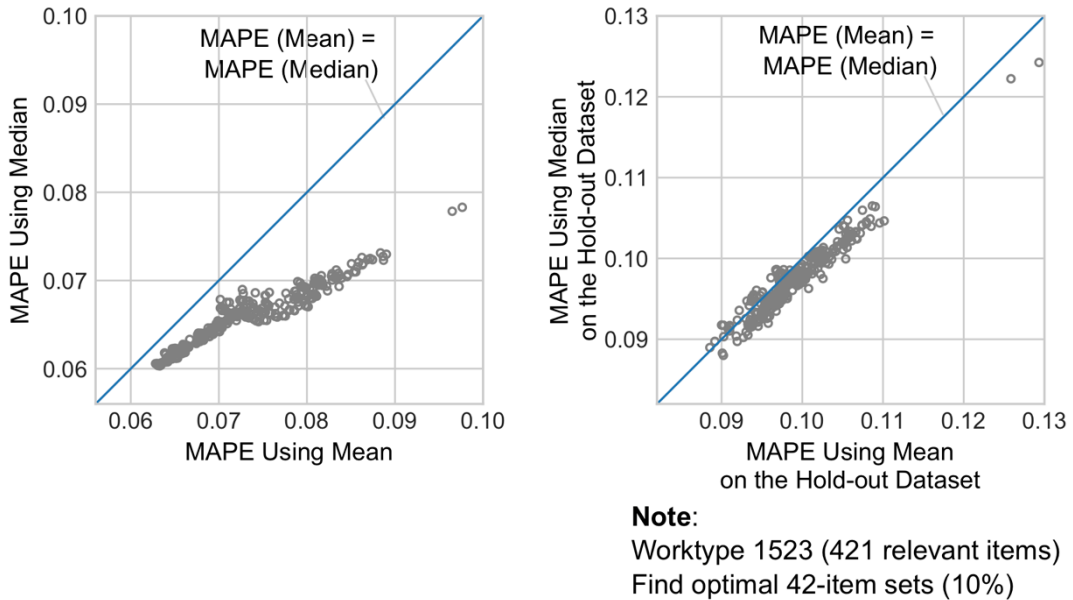
of the figure shows some trade-offs between maximizing  $P_{Mean}$  and maximizing  $P_{Median}$ , demonstrating that the two objectives should be separated as originally defined in Model 2. The left part of the figure illustrates that the mean is smaller than the median in all generated optimal solutions, indicating a left-skewed distribution of cost percentages.



**Fig. 5.10.** Comparison between the mean and median of cost percentages of a 42-major-item set over total project cost in HMA resurfacing projects

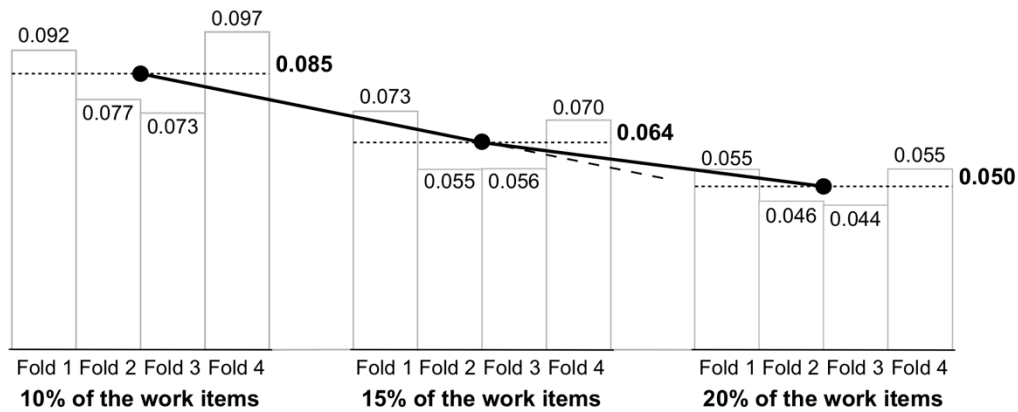
Each generated set of 42 major work items is associated with four MAPE values: MAPE using mean and MAPE using median, from the output of Model 2, and MAPE using mean and MAPE using median when applying the itemset to the hold-out dataset. The left part of Fig. 5.11 shows that MAPE using median (Strategy 2) is smaller than MAPE using mean (Strategy 1) on the optimization dataset. On the hold-out dataset, MAPE using median is also smaller than MAPE using mean for most solutions (see the right part of Fig. 5.11). Collectively, the figure suggests that median (Strategy 2) is a better center measure than mean (Strategy 1) in this case, as it produces smaller errors.

The result agrees with the common suggestion that median is more preferred to mean for skewed distributions.



**Fig. 5.11.** Comparison between MAPE using mean and MAPE using median on the optimization dataset and the hold-out dataset

As previously shown in Fig. 5.9, for optimal sets of 42 major work items, the MAPE using median on the hold-out dataset has an average value of 9.2%. To examine the changes when different subsets of the data were used for optimization, other 25% of project groups were left aside as the hold-out dataset while the remaining corresponding 75% of the projects were used for optimization, similar to four-fold cross-validation. The average errors in the four cases are 9.2%, 7.7%, 7.3%, and 9.7%, and the four-fold average error is 8.5% (see Fig. 5.12).



**Fig. 5.12.** Changes in average MAPE using median with four-fold cross-validation and increases in the number of major items

In order to obtain a more accurate cost estimate or a lower error of applying the Pareto principle to cost estimating itself, an obvious solution is to increase the number of major work items to be used for scoping-phase cost estimating. However, the effectiveness in reducing errors is not linear with the increase in the number of major items (see Fig. 5.12).

### 5.6. Discussion and practical implications

The proposed models' outputs and further analyses have addressed the five issues stated in the Introduction section about the applications of the Pareto principle to scoping-phase cost estimating by STAs.

First, the numbers 80 and 20 in the 80/20 rule and STAs' guidance summarized in Table 5.1 are not likely to hold in cost estimating. For example, Solution C in Fig. 5.5 corresponds to an optimal set of 42 major work items (i.e., 10% of all relevant items) that accounts for on average 90% of the total cost of an HMA resurfacing project.

Additionally, 20% of the work items can contribute up to 96.5% of the total project cost

on average (see Fig. 5.7). In these cases, 90/10 or 96/20 (not 80/20) apply. Second, Model 1 provides a measure of the variability of the cost percentages of a major item set over total project cost, which is not available in previous studies or STAs' guidance. Various trade-off solutions between maximizing the average cost percentage and minimizing the *CV* of cost percentages are also provided. Third, with an STA's historical bid tabulation dataset as input, the proposed models can automatically determine various optimal sets of major items, which helps avoid STAs' reliance on estimators' judgments and experiences in selecting work items for cost estimating.

Fourth, the applications of the Pareto principle to different project work types can be substantially different (see Fig. 5.7). Yet, STAs' guidance is the same for all projects regardless of project work types, which can cause significant errors in cost estimation. The proposed models can flexibly be applied to different STAs (i.e., work breakdown structures) and project work types (e.g., HMA resurfacing or PCC pavement) to obtain corresponding optimal sets of major work items for cost estimating. Last, the errors of applying the Pareto principle itself were not known but are now discovered by Model 2. Two strategies of using the generated optimal solutions to future estimating (i.e., mean or median of cost percentages of the major items in past projects) can also be compared using the output of Model 2. For example, in the case of HMA resurfacing projects and 42-item sets, Strategy 2 is more preferred due to its smaller MAPE values.

Since most STAs have very limited guidance on the application of the Pareto principle to cost estimating, the proposed models can significantly enhance the current practices by applying them to their historical bid tabulation data without collecting any



additional data. The generated optimal solutions and their corresponding measures (e.g.,  $P_{Mean}$ ,  $P_{Median}$ ,  $CV$ , MAPE using mean, and MAPE using median) provide enough detailed data-back information for their applications to future estimates. The error of the approach itself is also available. As the number of major work items used for estimating increases, the required effort increases, and the error of the approach decreases. Cost estimators can rely on an expected error of the approach to select the number of work items and specific work items for which they need to estimate unit prices. As the required accuracy of scoping-phase cost estimating is not high, with a required range from -30% to +50% (AASHTO 2013), an error of 8.5% of the approach of using the Pareto principle to cost estimating (see Fig. 5.12) seems acceptable, allowing mistakes caused by other factors (e.g., inaccuracies in quantity takeoffs and unit price estimates).

### **5.7. Summary and conclusions**

Cost estimates in the scoping phase are critical to the development of a typical transportation project. Project owner agencies use them to set the budget for project cost management. Due to the lack of detailed design plans in the scoping phase and the limited time allowed for estimating, STA cost estimators often apply the Pareto principle in their estimation. They focus time and effort on estimating major high-cost impact work items and account for the remaining items by a percentage or a minor item allowance. However, STAs have minimal guidance on this approach. Besides, few previous studies have investigated the issues associated with applying the Pareto principle, such as major item determination, variances among projects and project work types, or the error of the approach itself.

This study's primary contribution to the body of knowledge is the novel application of multi-objective optimization methods to address the issues and discover new knowledge of using the Pareto principle in cost estimating. From an STA's historical bid tabulation dataset, the proposed models can automatically determine various optimal sets of major work items for different project work types and numbers of work items. The output measures of each solution also provide definitive information for applying the optimal work item set for future projects, such as the distribution of cost percentages over total project cost with mean, median, and  $CV$  and the expected error associated with using the mean or the median for a new project, compared with estimating all work items relevant to the project and summing them up. The study has also explored the differences in applying the Pareto principle to different project work types and increasing the number of major work items used for cost estimation.

Due to data availability issues, this study is limited by considering only one primary factor influencing the list of work items and cost distribution in a project. Projects of the same project work type are similar to each other than projects of different work types. However, variations still exist. Considering extra factors may help create more uniform groups of projects. However, it also requires extra effort from STAs in collecting additional data, which may impede practical applications. Furthermore, the proposed models provide measures of variations and errors to help STA cost estimators make an informed data-back decision. Although this research focused on transportation projects, the proposed approach also applies to other construction sectors provided that a

systematic and consistent work breakdown structure is in use and historical bid tabulation data are available.

## 5.8. References

AASHTO (2013). "Practical guide to cost estimating." AASHTO Washington, DC.

Abdelmohsen, A. Z., and El-Rayes, K. (2016). "Optimal Trade-Offs between Construction Cost and Traffic Delay for Highway Work Zones." *Journal of Construction Engineering and Management*, 142(7), 05016004.

Abdelmohsen, A. Z., and El-Rayes, K. (2018). "Optimizing the Planning of Highway Work Zones to Maximize Safety and Mobility." *Journal of Management in Engineering*, 34(1).

Alroomi, A., Jeong, D. H. S., and Oberlender, G. D. (2012). "Analysis of Cost-Estimating Competencies Using Criticality Matrix and Factor Analysis." *Journal of Construction Engineering and Management*, 138(11), 1270-1280.

Cao, Y., Ashuri, B., and Baek, M. (2018). "Prediction of Unit Price Bids of Resurfacing Highway Projects through Ensemble Machine Learning." *Journal of Computing in Civil Engineering*, 32(5).

CTDOT (2019). "Connecticut Department of Transportation 2019 Estimating Guidelines." Connecticut Department of Transportation.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). "A fast and elitist multiobjective genetic algorithm: NSGA-II." *IEEE Transactions on Evolutionary Computation*, 6(2), 182-197.

- Dino, I. G., and Üçoluk, G. (2017). "Multiobjective Design Optimization of Building Space Layout, Energy, and Daylighting Performance." *Journal of Computing in Civil Engineering*, 31(5), 04017025.
- El-Abbasy, M. S., Elazouni, A., and Zayed, T. (2017). "Generic Scheduling Optimization Model for Multiple Construction Projects." *Journal of Computing in Civil Engineering*, 31(4).
- Elmousalami, H. H. (2020). "Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review." *Journal of Construction Engineering and Management*, 146(1).
- Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A. G., Parizeau, M., and Gagné, C. (2012). "DEAP: Evolutionary algorithms made easy." *The Journal of Machine Learning Research*, 13(1), 2171-2175.
- Gardner, B. J., Gransberg, D. D., and Rueda, J. A. (2017). "Stochastic Conceptual Cost Estimating of Highway Projects to Communicate Uncertainty Using Bootstrap Sampling." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3(3), 05016002.
- Halabya, A., and El-Rayes, K. (2020). "Optimizing the Planning of Pedestrian Facilities Upgrade Projects to Maximize Accessibility for People with Disabilities." *Journal of Construction Engineering and Management*, 146(1).
- Hyari, K. H., Khelifi, A., and Katkhuda, H. (2016). "Multiobjective Optimization of Roadway Lighting Projects." *Journal of Transportation Engineering*, 142(7).
- Iowa DOT (2012). "Design Manual." Iowa Department of Transportation.

- ITD (2020). "Construction Cost Estimating Guide." Idaho Transportation Department.
- Jeong, H. S., and Abraham, D. M. (2006). "Operational Response Model for Physically Attacked Water Networks Using NSGA-II." *Journal of Computing in Civil Engineering*, 20(5), 328-338.
- Karaca, I., Gransberg, D. D., and Jeong, H. D. (2020). "Improving the Accuracy of Early Cost Estimates on Transportation Infrastructure Projects." *Journal of Management in Engineering*, 36(5).
- Le, C., Le, T., Jeong, H. D., and Lee, E.-B. (2019). "Geographic Information System–Based Framework for Estimating and Visualizing Unit Prices of Highway Work Items." *Journal of Construction Engineering and Management*, 145(8), 04019044.
- MDT (2016). "Cost Estimation Procedure for Highway Design Projects." Montana Department of Transportation.
- MnDOT (2008). "Cost Estimation and Cost Management - Technical Reference Manual." Minnesota Department of Transportation.
- Olumide, A. O., Anderson, S. D., and Molenaar, K. R. (2010). "Sliding-Scale Contingency for Project Development Process." *Transportation Research Record*, 2151(1), 21-27.
- Ott, R. L., and Longnecker, M. (2015). *An Introduction to Statistical Methods and Data Analysis*, Cengage Learning.
- PennDOT (2018). "Estimating Manual." Pennsylvania Department of Transportation.

- Peralta, D., Bergmeir, C., Krone, M., Galende, M., Menéndez, M., Sainz-Palmero, G. I., Bertrand, C. M., Klawonn, F., and Benitez, J. M. (2018). "Multiobjective Optimization for Railway Maintenance Plans." *Journal of Computing in Civil Engineering*, 32(3), 04018014.
- Sayed, M., Abdel-Hamid, M., and El-Dash, K. (2020). "Improving cost estimation in construction projects." *International Journal of Construction Management*, 1-20.
- Schuldt, S., and El-Rayes, K. (2018). "Optimizing the Planning of Remote Construction Sites to Minimize Facility Destruction from Explosive Attacks." *Journal of Construction Engineering and Management*, 144(5).
- Shehab, T., and Meisami-Fard, I. (2013). "Cost-Estimating Model for Rubberized Asphalt Pavement Rehabilitation Projects." *Journal of Infrastructure Systems*, 19(4), 496-502.
- TxDOT (n.d.). "Risk-Based Construction Cost Estimating - Reference Guide." Texas Department of Transportation.
- WSDOT (2015). "Cost Estimating Manual for Projects." Washington State Department of Transportation.
- Zhang, Y., Minchin, R. E., and Agdas, D. (2017). "Forecasting Completed Cost of Highway Construction Projects Using LASSO Regularized Regression." *Journal of Construction Engineering and Management*, 143(10), 04017071.

## 6. CONCLUSIONS

Besides contractors' operations during construction, many decisions made by project owners before the start of construction (e.g., contractor selection, contract time estimation, and construction cost estimation) affect project outcomes. However, these decisions are typically made under uncertainty due to the lack of information in different project development phases and stem from decision makers' subjective judgments and experiences in a time-consuming process, particularly by State Highway Agencies (SHAs) for highway projects. Also, historical project data collected by SHAs are primarily for administrative purposes despite their potentials for enhancing future projects' decision-making. This study leveraged SHAs' preexisting data and developed novel data-driven approaches and frameworks for improving and complementing the current practices of project duration and cost-related decision making.

The first paper presented a Daily Work Report (DWR)-based approach for evaluating contractors' past production performance. For each controlling activity, the effects of four main contractor-independent factors, i.e., location, project budget, weather, and quantity of work, were tested to form different project condition groups. The activity's past production rates with the same condition group were compared and classified into three performance levels (i.e., high, medium, and low) with cut-off points determined by using classification techniques, distribution fitting, and Monte Carlo simulation. Performance indexes for individual controlling activities and their combination were also proposed to compare contractors in post-qualification or compare

against the thresholds predetermined by SHAs in pre-qualification. With the increasing use of DWRs, SHAs can easily apply the proposed approach to enhancing their practices without the need or extra effort to collect additional data.

In the second and third papers, an alternative and complementary approach to construction sequencing, a primary task of contract time estimation, was developed to alleviate the heavy and sole dependence of SHAs on planners' or schedulers' knowledge and experience. The primary contribution of the second paper is a sequential pattern mining (SPM)-based approach that allows for the automated development of knowledge bases of construction sequence patterns from DWR data for different project condition groups. Apart from the SPM's standard measure (i.e., support), domain measures, such as sequencing confidence, were proposed to evaluate the discovered sequential patterns among construction activities and enable a formal way to validate the effect of an influential factor (e.g., project work types) on construction sequencing. The most probable pattern of a given set of activities can also be suggested by comparing relevant alternatives with statistical tests.

Building upon the second paper, the third one employed the network theory in a novel way to visualize and interlink the pairwise logical relationships among construction activities (i.e., Start-Start, Start-Finish, Finish-Start, and Finish-Finish) discovered from DWR data. The proposed process model allows for the automated establishment of a construction logic knowledge network for a common project work type and the rapid application of the developed network to sequencing a new project. Three algorithms were designed to find the intermediate successor or predecessor of an



activity or sequence a set of activities that all were or even were not commonly occurred together in past projects. The algorithms also eliminate the redundant relationships that can be inferred from the others to return a simplified network of only relevant and necessary logics for the sequencing. Statistical measures of the lag time of a logical relationship were also determined.

Compared to contract time estimation, SHAs are more mature in applying historical project data to construction cost estimation, a closely related task. However, there is still significant room for improvement. SHAs apply the Pareto rule to cost estimating to establish cost baselines for management and monitoring in the scoping phase, but minimal guidance and information are available. A bid tabulation data-based approach was developed to identify optimal major item sets for different SHAs' work breakdown structures and project work types. It also provides necessary information for applying each item set to the scoping-phase cost estimating of a new project. A major item set's cost percentages over total project cost in past projects were presented by the mean, median, and coefficient of variance measures. The errors of the Pareto principle itself vary with the number of major items used for cost estimating. This relationship was also explored to help cost estimators make informed and data-back decisions.

Overall, this study developed novel data-driven approaches and frameworks to enhance three critical project duration- and cost-related decision-making practices by SHAs, i.e., contractor selection, contract time estimation, and construction cost estimation. Since all of the proposed approaches and frameworks only require SHAs' preexisting data (i.e., bid tabulation data and DWR data), the agencies can quickly

implement them to build data-driven decision-making systems without collecting additional data. However, this practical feature comes with a primary limitation, i.e., the consideration of only major influential factors on the decisions due to data availability issues. Nevertheless, if more data attributes are available, the proposed approaches and frameworks can be easily adjusted to consider them.