Enhancing Data-driven Applications in Construction

by

Lingzi Wu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Construction Engineering and Management

Department of Civil and Environmental Engineering

University of Alberta

ABSTRACT

Digital transformation of the construction industry has been slow and challenging. With continuously improving information and communication technology, increasing amounts of construction data are automatically generated throughout the stages of construction for all construction management functions. However, due to the complex nature of construction, collected data are noisy, fragmented, and discordant, consisting of observational and subjective as well as structured and unstructured information. These types of data form natural barriers for use in any data-driven applications, limiting their ability to provide reliable, timely, and informed decision support. How to fully exploit the value of "big data"—specifically, learn as much as we can from the raw construction data that we collect—is a challenge the entire construction industry is facing.

This research investigated this problem by addressing three specific challenges that hinder the digital transformation of the construction industry: 1) low automation for integrating and pre-processing fragmented construction data for project-level decision support; 2) lack of means for fusing information derived from various origins for data-driven simulation in real-time; and 3) slow implementation of machine learning, resulting in organizations 'drowning' in a flood of data.

This research adopted methods from applied mathematics and statistics, data science, and computing science to develop methodologies capable of addressing these challenges. This research better exploits the value of construction data and improves its conversion into informed project decision support. Specifically,

> **Through** the development of an enhanced data-driven application framework with two embedded custom functions to automate key data preprocessing steps for data

aggregation and merging, this research increases information flow between segmented data sets, thus enhancing data-driven simulation and analytics in general;

Through the proposal of two methods for enabling real-time input model calibration for simulation, this research establishes a foundation of dynamic data-driven simulations to incorporate real-time data of diverse origins, extending their applications to all stages of a project's life cycle and potential connections with multiple project stakeholders;

Through the development of a data solution to improve preliminary resource planning in industrial construction, this research not only provides vital decision support—a scientific and data-driven resource plan at the early planning stage—but also demonstrates the practicality of integrating unsupervised and supervised learning for large, unlabeled, and noisy construction data.

This research has achieved the goal of bridging low-quality construction data to a real-time data solution and contributed to the academic literature and construction industry by: 1) proposing a novel framework for enhanced data-driven applications built upon fragmented construction data; 2) developing and generalizing functions to automate and streamline the otherwise manual data pre-processing steps; 3) proposing a numerical-based Bayesian inference method for systematically updating input models (any given univariate continuous probability distribution) of simulations as new observations become available; 4) proposing a Markov chain Monte Carlo-based weighted geometric average method to effectively fuse information generated from diverse sources (both subjective and objective) for stochastic simulation inputs; and 5) developing a data solution to scientifically plan project resources with incomplete engineering.

PREFACE

This thesis is an original work by Lingzi Wu and follows a paper-based format. Various chapters, or portions thereof, have been published or are in revision in peer-reviewed journals.

A version of **Chapter 2** has been published as Wu, L., Li, Z., and AbouRizk, S., (2020) "Automation in extraction and sharing information between BIM and project management databases" *Proceedings of the International Conference on Construction and Real Estate Management* (ICCREM), Stockholm, Sweden. It is reprinted with permission of the American Society of Civil Engineers and is accessible online: https://doi.org/10.1061/9780784483237.005. Lingzi Wu was the lead investigator and was responsible for concept formation, model development, data analysis, and manuscript composition. Zuofu Li was an undergraduate research student involved in model development. Dr. Simaan AbouRizk was the supervisory authority and contributed to concept formation and manuscript composition.

A version of **Chapter 3** has been published as Wu, L., Ji, W., and AbouRizk, S. M. (2020). "Bayesian inference with Markov chain Monte Carlo–based numerical approach for input model updating." *Journal of Computing in Civil Engineering*, *34*(1), 04019043. It is reprinted with permission of the American Society of Civil Engineers and is accessible online: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000862. Lingzi Wu was responsible for concept formation, model development, case study analysis, and manuscript composition. Dr. Simaan AbouRizk and Dr. Wenying Ji had supervisory roles and contributed to concept formation and manuscript composition.

A version of **Chapter 4** has been published as Wu, L., and AbouRizk, S. M. "A numerical-based approach for updating simulation input in real time." *Journal of Computing in Civil*

Engineering,35(2), 04020067. It is reprinted with permission of the American Society of Civil Engineers and is accessible online: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000948. Lingzi Wu was responsible for concept formation, model development, case study analysis, and manuscript composition. Dr. Simaan AbouRizk had a supervisory role and contributed to concept formation and manuscript composition.

A version of **Chapter 5** has been submitted as Wu, L., Ji, W., Feng, B., Hermann U., and AbouRizk, S., "Intelligent Data-Driven Approach for enhancing preliminary resource planning in industrial construction." *Automation in Construction*, [Submitted, Nov 2020]. Lingzi Wu was the lead investigator and was responsible for concept formation, model development, data analysis, and manuscript composition. Baoli Feng was an undergraduate research student involved in data analysis. Dr. Simaan AbouRizk was the supervisory authority and contributed to concept formation and manuscript composition. Ulrich (Rick) Hermann is a long-time industry partner. He provided the data used for the case study and validated the result of data analysis.

ACKNOWLEDGEMENTS

To my supervisor, mentor, and role model, Dr. Simaan M. AbouRizk.

To my committee members: Dr. Aminah Robinson Fayek, Dr. Yasser Mohamed, Dr. Farook Hamzeh, Dr. Yashar Pourrahimian, and Dr. Xinming Li

To my external examiner: Dr. Mounir El Asmar

To my colleagues: Brenda Penner, Catherine Pretzlaw, Mickey Richards, and Stephen Hague.

To my industrial partners: Rick Hermann, Brian Gue, and Di Hu.

To the undergraduate research students: Baoli Feng, and Zuofu Li

To my beloved partner: R. Berkley Laurin.

To my parents: Jianlin Wu and Zhenli Xu.

To my furry babies: Kimi, Gigi, Mammoet, and Riceball.

To my deceased grandparents: Jing Xu, Shiding Wang, Yindi Mei, and Lianda Wu.

Table of Content

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
3D	Three-dimensional
ANN	Artificial neural networks
API	Application programming interface
AutoCAD	Automatic computer aided design
BIM	Building information modeling
CERP	Construction enterprise resource planning
CII	Construction industry institute
CO	Cumulative observations
CPM	Critical path method
DDDAS	Dynamic data driven application system
DES	Discrete event simulation
detecte3Dr	A custom R function detects 3D object relationships
DP	Dynamic programming
ERP	Enterprise resource planning
GA	Geometric average
GPS	Global positioning system
HDI	High density interval
HDR	High density region
HLA	High-level architecture
ICCREM	International conference on construction and real estate management
ICT	Information and communication technology
ID	Identification document
IT	Information technology
KNN	K-nearest neighbors
LCStr	A custom R function detects the longest common substring
LOD	Level of detail
MANOVA	Multivariate analysis of variance
MC	Monte Carlo
MCMC	Markov chain Monte Carlo

MD	Mixture density
MH	Metropolis-Hastings
ODBC	Open database connectivity
PC	Principal components
PCA	Principal components analysis
PDF	Probability density function
PM	Proposed methodology
RODBC	An R package of ODBC database interface.
SQL	Structured query language
SVM	Support vector machines
UD	Underlying distribution
US	United states of America
WAAM	Weighted arithmetic average method

1. CHAPTER 1: INTRODUCTION

1.1. BACKGROUND

Managing a construction project involves effective planning, organizing, executing, monitoring, and controlling (Ahuja et al. 1994). The introduction of quantitative analysis in assisting these project management functions has been slowly shifting the construction industry from a craft-oriented culture to a data-driven one.

The digital transformation in the construction industry, however, has been slow and challenging (Wu and AbouRizk 2021a). Unlike its increasingly digitized customers, the construction industry seems to "stuck in the analog era" (Koeleman et al. 2019). The low level of digitalization and painful transformation is caused by many inherent characteristics of construction:

Fragmentation. The lifecycle of a construction project involves a collection of large and small contractors, subcontractors, and numerous specialized trades (Neelamkavil 2009). Each agency in this dynamic network has adopted various degrees of technological advancements to suit its operation, capital, and culture (Rezgui et al. 2011, Sardroud 2015). This is especially true for independent subcontractors and suppliers, who have little incentive to embrace advanced technologies during the brief periods when they are on the project (Koeleman et al. 2019). Standardizing and implementing information technology across a project, which requires buy-in from all stakeholders, is extremely challenging.

Complex and Unique. Unlike many other industries, construction projects are almost always one-of-a-kind and complex (Behzadan et al. 2015, Lu et al. 2015, Caldas and Soibelman 2003). Many project management processes are in place to fulfill specific contractual requirements for a specific project. The industry's inherent

complexity and uniqueness challenge information technologies to provide universal solutions that apply to various stakeholders and diverse projects and that cross multiple project phases.

Transience. Workforce turnover is extremely high in construction at both the project and company levels (Koeleman et al. 2019). This results in a corporation's reluctance to invest in employee training, which is critical in digitalization. Additionally, a significant amount of project data is exchanged verbally or in unstructured documents among only the involved personnel (Caldas et al. 2002; Al Qady and Kandil 2013). These data often include experts' opinions of the current project conditions and forecasts of future performance, shedding light on important factors for critical project decision-making. Unfortunately, very few of these valuable assets are captured or shared to enhance future projects in a company—they remain in the mind of the individuals and leave the company when the person departs (Martínez-Rojas et al. 2016).

Remote and Harsh Environments. Unlike manufacturing, which takes place in a well-controlled environment, construction sites are often in remote and harsh environments (Sidawi and Alsudairi 2014). The harsh and remote site conditions pose extra challenges to hardware and software development, including limited IT support, which impedes the data quality. Construction data is often embedded with missing values, human errors, and outliers.

These characteristics of the construction sector make it particularly difficult to develop a standardized commercial digital solution at both the industry and corporation levels. Indeed, many large general contractors, such as our industry partners, reported a large portion of in-house developed information technologies, such as APIs (application programming interface),

databases, and data warehouses. Without coordination, the competing systems exist not only among companies but also within a single firm (Koeleman et al. 2019). Especially for large companies, who are often highly federated, project data from different departments are commonly maintained and stored separately at different databases, either commercially- or in-house-developed.

Reflecting the low level of digitalization—reported among one of the world's least digitized industry by McKinsey Global Institute (Manyika et al. 2015)—construction data are noisy, both "soft" and "hard", both structured and unstructured, and segmented. These types of data form natural barriers for use in any data-driven applications to provide reliable, timely, and informed decision supports.

1.2. DEFINITION OF DATA

Before diving into research-related details, clarification of the following concepts will greatly assist readers in grasping the magnitude of the problem and the significance of this research.

Data are defined as "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in the electronic form that can be stored and used by a computer" (Cambridge Business English Dictionary 2020). Within the data boundary, hard data are defined as "information such as numbers or facts that can be proved" (Cambridge Business English Dictionary 2020). In other words, hard data are in the form of numbers or graphs (McGraw-Hill Dictionary of Scientific & Technical Terms 2003). Opposed to hard data, soft data are defined as "information about things that are difficult to measure such as people's opinions or feelings" (Cambridge Business English Dictionary 2020).

Both hard and soft data prevailingly exist in the construction industry. The development and implementation of various sensor technologies (e.g. radio frequency identification, global positioning system, laser, and vision-based detection) in construction has drastically improved the efficiency of the data collection process in industry (Zhang et al. 2017), resulting in larger volumes of hard data. These data, also referred to as observational data in this research, are commonly stored in relational databases, as a single spreadsheet quickly reaches its capacity. On the other hand, seemingly repetitive operations in construction can drastically differ due to uncertainties and external factors, such as location, weather, labor skills, morale, and utilization of technology (Seresht and Robinson Fayek 2018). The majority of such important project data—which capture these external factors—are stored in unstructured text documents or exchanged verbally among the professionals involved, rendering the terms "soft data" and difficult to apply (Martínez-Rojas et al. 2016, Caldas et al. 2002, Al Qady and Kandil 2013).

1.3. PROBLEM STATEMENT

How to fully exploit the value of construction data—specifically, to learn as much as we can from the raw construction data that we collect and covert them into informed decision support—is a grand challenge the entire construction industry is facing. Attempting to answer this grand problem, this research identified the following three bottlenecks that prevail in the construction industry, preventing data from being properly converted into informed decision supports. Exploring solutions to these identified bottlenecks, this research removes some systemic barriers to increase data usage and enhance data-driven applications in construction, thus bridging fragmented construction data with a real-time data-driven application.

Challenge 1: Low automation in integrating and pre-processing segmented construction data for project-level decision support

The existing standalone data management systems support a limited number of data analysis functions and decision support tasks (Ng et al. 2017), but they often fail to provide insights in connections among various data sets or provide a high-level integrated view. For instance, a safety database can summarize, report, and visualize safety incidents at various detail levels throughout the entire project. But, to discover the potential correlation between safety indices and various project conditions, or a specific incident's immediate and delayed effect on the project requires merging, process, and cleaning data that are stored in separate databases to a central location.

Because each of the databases is developed individually and often cater to a specific construction management function (as illustrated in the above case), the level of detail and structure of the data can differ drastically among different databases, creating difficulties in data sharing, syncing, and aggregating. The resulting project data are segregated like individual islands without channels to flow and integrate freely, thus requiring routine data manipulation manually. Repetitive and mundane manual manipulation is not only an inefficient use of experts' time but also introduces human error further lowering the data quality (Wu et al. 2020a).

Challenge 2: Lack of means in fusing information derived from various origins for data-driven simulation in real-time

Data-driven simulations have been widely used in project management to plan, schedule, and control in a variety of construction projects. However, many documented challenges, especially failing to effectively reflect the project conditions based on real-time project data,

limit this powerful tool to mostly planning stages (Martínez-Rojas et al. 2016, Abdelmegid et al. 2020, Leite et al. 2016).

The success of a simulation model is highly dependent on accurately modeling the inputs, particularly in construction where a considerable number of inputs (each imbued with a wide variety of uncertainties) all relate to the underlying random process of various activities and tasks (Wu et al. 2020b, Wu and AbouRizk 2021b). Modeling inputs as probability distributions, in a process known as stochastic or Monte Carlo simulation, has been widely studied and used in the construction industry due to its success at incorporating the randomness and various uncertainties inherent to construction activities. Nevertheless, these input probability distributions are often rigid (e.g. a distribution fitted from historical data or experts' judgments) and lack reliable or effective solutions for fusing actual performance and subjective opinions with the original input distribution to achieve real-time updating (Akhavian and Behzadan 2013).

Challenge 3: Behind the curve in implementing machine learning—drowning in a flood of data

The traditionally craft-oriented culture and processes in the construction industry pose additional barriers to the adoption of data-driven applications, as the industry tends to put trust in individual experience and expertise over empirics. Very few companies have data analysts on staff who can take ownership of advanced analytics initiatives (Hovnanian et al. 2019).

On the other hand, a majority of marching learning techniques are like black-boxes, where the paths from input to output are too complicated for any human to comprehend (Rudin 2019). Additionally, the typical machine learning process focuses on selecting a model solely

on performance metrics, such as accuracy rate, while neglecting the interpretability and actionable insights of the data (Krause et al. 2016).

The often inaccurate predictions from these machine learned models, as a result of low-quality data, have exacerbated the clash between craft-oriented culture in construction and the implementation of black-box approaches. Reflecting the highly-fragmented, stressful, and dynamic environment—like most raw data—construction data require extensive data preprocessing, such as identifying outliers, labelling, and formatting to produce a robust, predictive model (Han et al. 2011). Due to a lack of trained staff and constant pressure to deliver, this critical step—a key to the success of gaining insights into raw facts—is often skipped or inadequately performed during project execution. Consequently, the results of these data solutions often demonstrate large deviations and mismatch with human intuition, which further diminishes trust of machine-learned models by the construction industry.

To conclude, most construction companies are only capable of summarizing the quick, easy, factual information from a large amount of construction data they collect, and they miss the potentially critical project information such as connections, correlations, and causal relationships among large data sets that are often hidden and harder to discover— rendering the majority of the construction agents drowning in a flood of data (Leite et al, 2016).

1.4. RESEARCH AIM AND OBJECTIVES

Aiming to better exploit the value of construction data and convert data into informed project decision supports, this research explored and adopted methods from applied mathematics and statistics, data science, and computing science. The research goal of bridging low-quality construction data to a real-time data solution was achieved by addressing the above-mentioned challenges and accomplishing the following objectives:

1) Increasing information flow among separate databases and automating critical data pre-processing tasks (Chapter 2). This objective focuses on addressing the challenges of the common, labor-intensive, raw data preparation steps that prevail in the industry. Data sets that are hard to integrate due to drastically different structure can be easily merged and prepared to the required level for the next step analysis by automating two major data pre-processing tasks. Further generalizing these two functions as public *R* library, this research developed a framework that linked original raw data stored in separate databases with the ultimate managerial product, data-driven analytics and/or simulation model.
2) Improving the input modeling process to incorporate real-time "soft" and "hard" data (Chapters 3 and 4). This objective examines the state-of-art methods in fusing data generated from various origins and adopts methodologies from applied statistics to improve the way we model simulation inputs in real-time. It also enables a new generation of decision-support systems capable of incorporating real-time updates based on as-built data, integrating different data sources (both subjective and objective), and consolidating all critical information into the input of data-driven simulation models.
3) Demonstrating how to use a variety of machine learning algorithms to augment the data parsing and labelling process, learn critical design information from large available historical construction data sets, and form a data-driven decision support system in future projects (Chapter 5). This research objective examines a critical construction management function, namely preliminary resource planning—a preconstruction planning function traditionally carried out in an ad hoc way—and proposes a combination of supervised and unsupervised (or semi-supervised) machine learning methods to reveal critical design information from a low level-of-detail 3D

model while engineering is incomplete, thereby bridging certain design to its historical resource requirement. This research component is the first academic study for deriving a set of data-driven preliminary resource planning indices given limited engineering. By demonstrating how unsupervised learning can be combined with supervised learning, this research component also promotes analysis of large construction data sets from a data-driven perspective, and effectively brings forward "hidden" project information without the labor-intensive process of labelling and parsing data.

Segmented, raw construction data deter the majority of data-driven applications. As depicted in Figure 1-1, the first research objective addresses this bottleneck and lays a solid foundation for the following research activities. The second and third research objectives further discuss how to incorporate both "hard" and "soft" construction data in real-time for enhanced decision supports.

1.5. RESEARCH METHODOLOGY AND ACTIVITIES

This research explored and adopted interdisciplinary methods such as dynamic programming, Bayesian inference, Monte Carlo and Markov chain Monte Carlo (MCMC) methods, weighted geometric averages, and supervised and unsupervised machine learning algorithms to achieve the research objectives of connecting raw construction data to real-time data solution.

The goal and objectives of this research was achieved through the following activities (Figure 1-1):

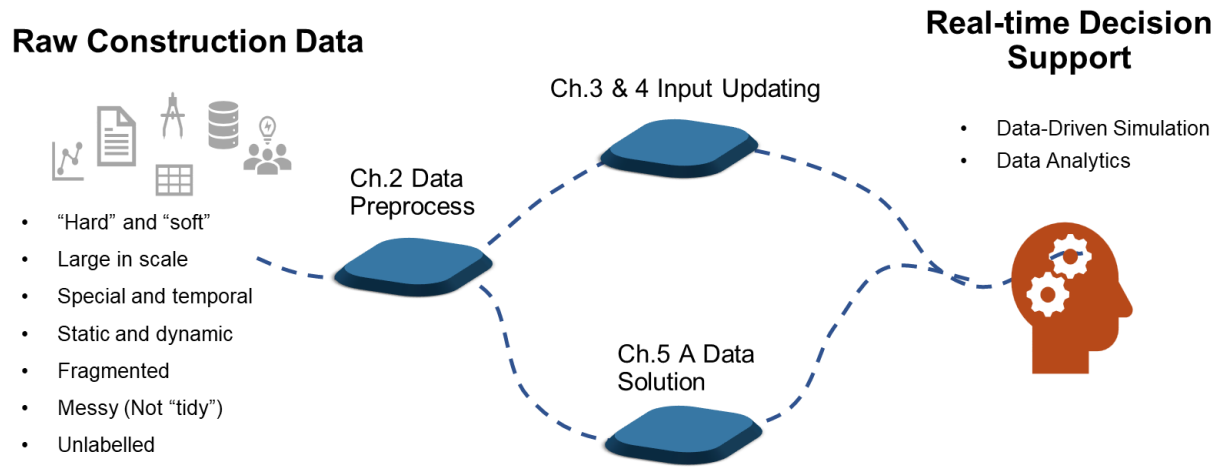


Figure 1-1 Roadmap of the research

1.5.1. Enhancing real-time information flow for data-driven applications

This research activity in Chapter 2 was developed to address the challenges of low automation in data preprocessing steps for highly fragmented and non-integrated raw construction data. Specifically, this research developed two custom functions (published as *R* libraries) to automate key data preprocessing steps for data aggregating and merging: the dynamic programming-based longest common substring algorithm (Li et al. 2019) and the interval-based 3D object relationship detection algorithm (Wu and AbouRizk 2020). Building on these two functions, this research proposed a framework to automatically prepare data generated from various origins with no obvious path to a tidy format (where each variable is a column, and each observation is a row (Wickham 2014)) and can be supplied to a wide range of applications in real-time. The proposed framework increases the information flow among segmented data sets, thus enhancing data-driven simulation and analytics in general.

1.5.2. Input modelling for dynamic data-driven simulations in Construction

Targeting the challenges of rigid parameters, structure, and assumptions that limit data-driven simulations' application mostly to the planning stage, this research activity systematically studied the way we model simulation inputs and proposed two methods in enabling real-time input model calibration for simulation.

The first study in Chapter 3 coupled the MCMC method with Bayesian inference to enable input model updating with real-time observations for any given univariate distribution. This research not only demonstrates Bayesian inference updates and approximates the underlying probability distribution despite noisy data, in a real-time manner, but also extends its application to cases when analytical solutions do not exist. The second study in Chapter 4 has proposed a MCMC-based, weighted geometric averaging method for fusing data generated from diverse sources—including both subjective and objective—to update inputs of simulation models in real-time. This research addresses the practical challenges associated with fusing observational and subjective project information in real-time, which is extremely common in construction, as experts' opinions are critical during a decision-making process.

The methodological improvement of enabling data-driven simulations to incorporate real-time data of diverse origins extends their applications to all stages of a project's life cycle and potential connections with multiple project stakeholders. Additionally, both studies implemented MCMC-based numerical methods, demonstrating their great potential for approximating complex and arbitrary probability in the engineering discipline.

1.5.3. Data solution to improve industrial construction preliminary resource planning

This research in Chapter 5 proposed a data-driven approach for preliminary resource planning in industrial construction projects. This approach deploys semi-supervised machine learning techniques (clustering and classification) to learn from historical projects, predict module types, and generate key resource planning indices based on incomplete, segmented—yet early available—data for a future project. This research not only provides vital decision support, as it is the first to provide a scientific and data-driven resource plan at the early planning stage with limited engineering, but it also demonstrates the practicality of integrating unsupervised and supervised learning for large unlabeled noisy construction data.

1.6. THESIS ORGANIZATION

This thesis is organized following a paper-based format, and the remainder of the thesis is organized as described in Table 1-1.

Table 1-1. Thesis organization

Chapter	Research Activity	Publication
2	1.4.1"Enhancing real-time information flow for data-driven applications"	Wu, L., Li, Z., and AbouRizk, S., (2020) "Automation in extraction and sharing information between BIM and project management databases" *Proceedings of the International Conference on Construction and Real Estate Management (ICCREM) 2020:* 37-46, Stockholm, Sweden
3	1.4.2 "Input modelling for dynamic data-driven simulations in Construction"	Wu, L., Ji, W., and AbouRizk, S. M. (2020). "Bayesian inference with Markov chain Monte Carlo–based numerical approach for input model updating." *Journal of Computing in Civil Engineering*, *34*(1), 04019043
4		Wu, L., and AbouRizk, S. (2021). Numerical-Based Approach for Updating Simulation Input in Real Time. *Journal of Computing in Civil Engineering*, *35*(2), 04020067.
5	1.4.3 "Data solution to improve industrial construction preliminary resource planning"	Wu, L., Ji, W., Feng, B., Hermann U., and AbouRizk, S., "Intelligent Data-Driven Approach for enhancing preliminary resource planning in industrial construction." *Automation in Construction*, [Revision]
6	Conclusion	NA

1.7. REFERENCES

Abdelmegid, M. A., González, V. A., Poshdar, M., O'Sullivan, M., Walker, C. G., and Ying, F. 2020. "Barriers to adopting simulation modelling in construction industry." *Automation in Construction*, *111*, 103046.

Ahuja, H. N., Dozzi, S. P., and Abourizk, S. M. 1994. *Project management: techniques in planning and controlling construction projects*. John Wiley & Sons.

Al Qady, M., and Kandil, A. 2013. "Document discourse for managing construction project documents." *Journal of Computing in Civil Engineering*, *27*(5), 466-475.

Akhavian, R., and Behzadan, A. H. 2013. "Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining." *Journal of Construction Engineering and Management*, 139(11), 04013021.

Behzadan, A. H., Menassa, C. C., and Pradhan, A. R. 2015. "Enabling real time simulation of architecture, engineering, construction, and facility management (AEC/FM) systems: a review of formalism, model architecture, and data representation." *ITcon*. 20: 1-23.

Caldas, C. H., and Soibelman, L. 2003. "Automating hierarchical document classification for construction management information systems." *Automation in Construction*, *12*(4), 395-406.

Caldas, C. H., Soibelman, L., and Han, J. 2002. "Automated classification of construction project documents." *Journal of Computing in Civil Engineering*, *16*(4), 234-243.

Data, Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/data Retrieved December 13, 2020

Han, J., Pei, J., and Kamber, M. 2011. *Data mining: concepts and techniques*. Elsevier.

Hard data. (n.d.) McGraw-Hill Dictionary of Scientific & Technical Terms, 6E. 2003. https://encyclopedia2.thefreedictionary.com/hard+data Retrieved December 13, 2020

Hard data, Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/hard-data Retrieved December 13, 2020

Hovnanian, G., Kroll, K., and Sjödin, E. 2019. How analytics can drive smarter engineering and construction decisions. *McKinsey & Company Capital Projects & Infrastructure*.

Krause, J., Perer, A., and Ng, K. 2016. "Interacting with predictions: Visual inspection of black-box machine learning models." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* 5686-5697.

Koeleman, J., Ribeirinho, M. J., Rockhill, D., Sjödin, E., and Strube, G. 2019. "Decoding digital transformation in construction." *McKinsey and Company: Chicago, IL, USA*.

Leite, F., Cho, Y., Behzadan, A. H., Lee, S., Choe, S., Fang, Y., and Hwang, S. 2016. "Visualization, information modeling, and simulation: Grand challenges in the construction industry." *Journal of Computing in Civil Engineering*, 30(6): 04016035.

Li, Z., Wu, L., and AbouRizk, S. 2019. XiaomoLing/LongestCommonSubString: First Release of Longest Common SubString R Library (Version v1.0.0). Zenodo, http://doi.org/10.5281/zenodo.4057067

Lu, Y., Li, Y., Skibniewski, M., Wu, Z., Wang, R., and Le, Y. 2015. "Information and communication technology applications in architecture, engineering, and construction organizations: A 15-year review." *Journal of Management in Engineering*, *31*(1), A4014010.

Manyika, J., Ramaswamy, S., Khanna, S., Sarrazin, H., Pinkus, G., Sethupathy, G., and Yaffe, A. 2015. "Digital America: A tale of the haves and have-mores" *McKinsey Global Institute*.

Martínez-Rojas, M., Marín, N., and Vila, M. A. 2016. "The role of information technologies to address data handling in construction project management." *Journal of Computing in Civil Engineering*, 30(4): 04015064.

Neelamkavil, J. 2009. "Automation in the prefab and modular construction industry." In *26th Symposium on Construction Robotics ISARC*.

Ng, S. T., Xu, F. J., Yang, Y., and Lu, M. 2017. "A master data management solution to unlock the value of big infrastructure data for smart, sustainable and resilient city planning." *Procedia Engineering*, *196*, 939-947.

Rezgui, Y., Boddy, S., Wetherill, M., and Cooper, G. 2011. "Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction?" *Computer-Aided Design*, *43*(5), 502-515.

Rudin, C. 2019 "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nat Mach Intell,* 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sardroud, J. M. 2015. "Perceptions of automated data collection technology use in the construction industry." *Journal of Civil Engineering and Management*, *21*(1), 54-66.

Seresht, N. G., and Robinson Fayek, A. 2018. "Dynamic modeling of multifactor construction productivity for equipment-intensive activities." *Journal of Construction Engineering and Management*, *144*(9): 04018091.

Sidawi, B., and Alsudairi, A. 2014. "The potentials of and barriers to the utilization of advanced computer systems in remote construction projects: case of the Kingdom of Saudi Arabia." *Visualization in Engineering*, *2*(1), 3.

Soft data, Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/soft-data Retrieved December 13, 2020

Wickham, H. 2014. "Tidy data." *Journal of Statistical Software*. 59(10): 1-23. http://theta.edu.pl/wp-content/uploads/2012/10/v59i10.pdf, last accessed Nov 02, 2020.

Wu, L., and AbouRizk, S., 2020. XiaomoLing/Detect3DRelation: First Release of the Detect 3D Relation function (Version v1.0.0). Zenodo, http://doi.org/10.5281/zenodo.4058576

Wu, L., and AbouRizk, S. 2021a. "Towards construction's digital future: a roadmap for enhancing data value" *9th CSCE International Construction Specialty Conference,* Niagara Falls, Canada [In Press]

Wu, L., and AbouRizk, S. 2021b. "Numerical-Based Approach for Updating Simulation Input in Real Time." *Journal of Computing in Civil Engineering*, *35*(2), 04020067.

Wu, L., Li, Z., and AbouRizk, S., 2020a. "Automation in extraction and sharing information between BIM and project management databases" *Proceedings of the International Conference on Construction and Real Estate Management* (ICCREM), Stockholm, Sweden

Wu, L., Ji, W., and AbouRizk, S. M. 2020b. "Bayesian inference with Markov chain Monte Carlo–based numerical approach for input model updating." *Journal of Computing in Civil Engineering*, *34*(1), 04019043.

Zhang, M., Cao, T., and Zhao, X. 2017. "Applying sensor-based technology to improve construction safety management." *Sensors*, 17(8): 1841.

2. CHAPTER 2: ENHANCING REAL-TIME INFORMATION FLOW FOR DATA-DRIVEN APPLICATIONS IN INDUSTRIAL CONSTRUCTION

2.1. INTRODUCTION

With the digitalization—the rapid evolution of digital technologies enabling both generating and storing data electronically—more data than ever are being created and stored during construction (Soibelman et al. 2008, Soibelman and Kim 2002). Benefiting from information learned through these data, construction agencies have been slowly shifting from a craft-oriented culture to a data-driven one. Ultimately the more a company can learn from its data, the better its chances of identifying potential risks, increasing profitability, and staying competitive (McGee et al 1993).

With increasing data generated through automated data collection technology, such as radio frequency tags, barcode, and sensors (Sardroud 2015), a single spreadsheet quickly reaches its maximum capacity to properly store, transmit, and manage data. In the past few decades, construction agencies have invested time, effort, and money in information and communication technology (ICT) to manage the increasing amount of data (Martínez-Rojas et al. 2016, Sardroud 2015). As a result, the majority of observational construction data is stored in large, complex databases, either developed commercially or in-house.

Nevertheless, the value of these construction data has been significantly under exploited (Leite et l. 2016, Sardroud 2015, Barbosa et al. 2017). The construction industry is one of the world's least digitized sectors (Manyika et al. 2015), according to *McKinsey Global Institute*, and it seems to "stuck in the analog era" (Koeleman et al. 2019).

Many researchers have studied and documented the barriers and struggles to standardizing and fully-implementing digital solutions in construction (Martínez-Rojas et al. 2016, Sardroud 2015, Lu et al. 2015, Adriaanse et al. 2010). Hindering digitalization are the inherent characteristics of the construction industry: 1) fragmentation along the value chain

(Nitithamyong and Skibniewski 2004, Rezgui et al. 2011, Lu et al. 2015, Koeleman et al. 2019), 2) uniqueness and complexity (Behzadan et al. 2015, Adriaanse et al. 2010), 3) transient and temporariness (Koeleman et al. 2019), and 4) open and uncertain conditions (Bowden et al. 2006; Behzadan et al. 2008). These characteristics of the industry further create secondary barriers, such as the data's low quality leading to lost trust (Soibelman et al. 2008), inadequate employee trainings (Lu et al. 2015, Viljamaa and Peltomaa 2014), and low research and development budgets (Agarwal et al. 2016).

These challenges are magnified when it comes to industrial construction industry, which is more complex; often located in remote, harsh environments; and usually involves multiple layers of contractors, suppliers, specialized trades and subcontractors at each phase of the project lifecycle (Koeleman et al. 2019). As such the construction data in a large industrial construction project is highly segmented—1) each construction agency along the value chain adopts a certain degree of ICT that matches its own organizational structure, culture, and capital; and 2) each department within one large construction company, functioning as one federate, invests in ICT independently without coordinating with other departments. The lack of industry-wide standardization and the lack of coordination during development and implementation create physical barriers to freely share project data among departments and corporations (Forcada et al. 2007, Chassiakos and Sakellaropoulos 2008).

These standalone data management systems usually support a limited number of data analysis functions and decision support tasks (Ng et al. 2017), but often fail to provide insights at a high-level with an integrated view, or to be easily shared and reused by downstream stakeholders. In practice, domain experts routinely (e.g. weekly, biweekly, monthly) perform manual data manipulation, including data extraction, merging, filtering, re-entery, summarizing, and wrangling. Multiple iterations of manual data pre-processing

not only is inefficient and labor-costly, but it also introduces human error, further lowering the data quality. The low level of information integration occupies experts' essential time, introduces potential errors, and hinders the success of construction projects (Mitchell 2006, Saraf et al, 2007, Berteaux and Jacernick-Will 2015).

The initiative of this research came directly from a multi-national general contractor, whose portfolio covers heavy industrial, infrastructure, and commercial/residential construction, along with special projects. As an industry leader, they have implemented many commercial information systems, such as Autodesk Navisworks, Oracle's JD Edwards EnterpriseOne, Primavera P6 Enterprise Project Portfolio Management, as well as dozens of in-house developed information systems, from a small scaffolding request system to a large progress tracking database. They have been struggling with the aforementioned challenges at full scale in past industrial construction projects.

This research developed two custom functions to automate common data pre-processing tasks and address this prevailing challenge in the industrial construction industry: 1) automating data mapping based on auto-detected keywords; 2) auto-identifying overlapped and/or included relationships between two 3D objects. Wrapped around these two core functions is a framework that incorporates data adopter(s) for real-time access of segmented data sets, structured query language (SQL) for effective data merging, and various simulation and machine learning algorithms for advanced data analysis. The proposed framework demonstrates the capability in effectively and efficiently bridging raw, segmented construction data with various data analytics through automated data pre-processing. The proposed framework significantly reduces the manual data manipulation, improves the data quality, streamlines the processes between raw, segmented data and data solutions, thus enhance data-driven applications in general.

After validating the two core custom functions, the proposed framework was implemented in a case study where historical project data was used for auto-generating S-curves for project controlling and managing tasks. The proposed framework provides a universal solution to increase automation in data pre-processing for all fragmented construction data. As a result, it promotes information flow, improves data quality, and ultimately, increases data utilization for critical project decision supports. The rest of the chapter is organized as follows: a systematic literature review on integrating information management systems in construction, the methodology of the proposed framework, validation of the custom functions, and an industrial case study.

2.2. RESEARCH BACKGROUND

Construction projects—regardless of size—generate a large amount of raw data throughout their lifecycles (Tatari et al. 2004, Martínez-Rojas et al. 2016). As outlined in Figure 2-1, the project data generated at one project phase are often critical for stakeholders within this phase, as well as the following phase(s) (Hu 2008). However, due to the fundamental causes and derivative barriers listed above, the data flow has been rocky, causing information loss and inefficiency—either among departments within a corporation or among organizations.

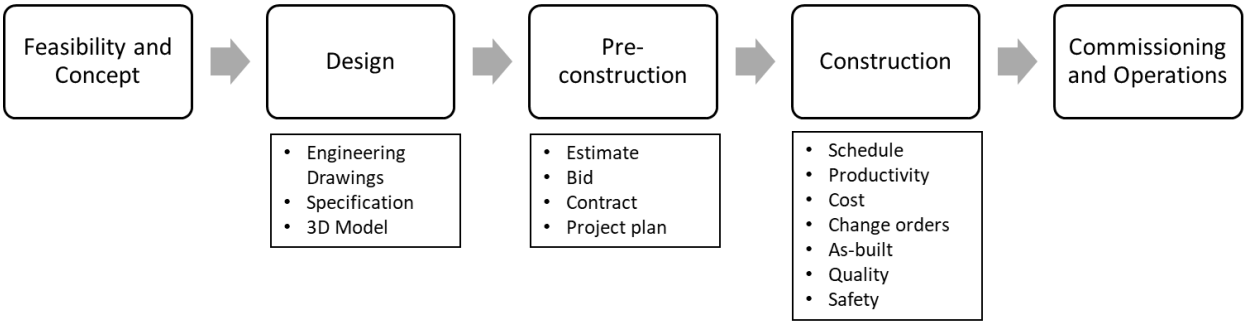


Figure 2-1 Construction project phases and typical construction data involved

Previous research has proposed many solutions and frameworks in the attempt to address fragmented construction data, improve information flow, and achieve integrated decision support (Yuan et al. 2019). The three subsections below review such research efforts including the most promising ICTs: construction enterprise resource planning (CERP) systems and building information modeling (BIM). Regardless of the degree of implementation, they are unarguably the most widely adopted ICT with the capability of integrating fragmented data.

2.2.1. Construction Enterprise Resource Planning

Originating from manufacturing, enterprise resource planning (ERP) systems are highly complex information systems (Fizgerald 1992, Umble, 2003). More than a traditional database, these information systems provide an enterprise with a platform to gather and store business data among departments and perform integrated critical management functions (Shi and Halpin, 2003, Thompson 1996; Gibson and Holland 1999; Tinham 1999). With widely recognized benefits including financial, operational, managerial, strategic, organizational and IT (Fan, 2018), an ERP system in manufacturing commonly includes functions such as production planning, purchasing, inventory control, sales, marketing, financial, human resources (Umble, 2003, Zhang et al, 2005).

To extend ERP's benefits to the construction industry, Shi and Halpin (2003) first studied and proposed CERP knowledge base systems and summarised 6 features—"project-oriented, paralleled and distributed, open and expandable, scalable, remotely accessible, and reliable and robust"—as necessities for a CERP system to suit construction industry's needs. Later, a few qualitative analyses of implementation, performance, and recommendation of CERP systems were conducted: Tatari et al (2008) employed a qualitative system dynamics model to explore and describe the implementation and performance of CERP in two construction

agencies; Chung et al (2008) identified factors associated with the success/failure of CERP systems.

Despite the great potential benefits, the costly implementation and high demands of corporation's time, effort, and resources (Umble 2003, Voordijk et al. 2003) has deterred the majority of general contractors. Consequently, many construction firms have implemented their CERP systems partially, where only the financial management processes were adopted (Tatari et al. 2007). Other software packages for the rest of the construction processes, commercial or in-house developed, are then integrated with the partial CERP (Chang et al 2008).

2.2.2. Building Information Modeling

Building information modeling (BIM), as an extension and enhancement of conventional computer-aided design tools, has been widely adopted in construction industry to support design, design evaluation, quantity take-off, construction planning and control (Kaner et al 2008).

Defined as "a digital representation of physical and functional characteristics of a facility" (US National Building Information Model Standard), BIM has the potential to integrate design information with project data in a digital format throughout the project's lifecycle (Penttila, 2006). Regardless of the documented benefits of using BIM to integrate graphical and non-graphical data for different construction functions (Jung and Gibson, 1999; Sanvido and Medeiros, 1990; Teicholz and Fischer, 1994, Azhar 2011, Kimmance 2002), the full utilization of BIM beyond design stage has been very limited. In particular, Santos et al. (2017) provides a systematic review and analysis of BIM in assisting various aspects of project decision-making.

A series of studies have examined, evaluated, and documented the challenges in adopting BIM as an information hub to integrate design data and various construction data (Guerra and Leite, 2020, Arshad et al. 2019, Hamdi and Leite 2013, Alwash et al. 2017, Olatunji 2015, Ashcraft 2008). First, such attempt is a costly, multi-organizational endeavor, which requires buy-in from all stakeholders along the project value chain (Eastman et al 2009, Solihin et al. 2017, Solihin and Eastman 2015). Second, the lack of industry-wide standards and legal status of 3D models—not perceived as a contractual deliverable—causes great contractual risks and legal concerns for many major players (Arshad et al. 2019, Ashcraft 2008, Olatunji 2015). To reduce risks and avoid potential legal issues, BIM (or 3D model) has been excluded from contracts as a source of deliverables, and is widely perceived to be unnecessary by downstream stakeholders, such as the general contractor and subcontractors (Guerra and Leite, 2020).

To conclude, although it has great potential, such as a handful of demonstrations of BIM as a link to connect pre-existing ERP systems with engineering drawings (Babic et al, 2010), BIM has not been the universal solution to the fragmented construction data.

2.2.3. Other Solutions of Integrated Data Analysis on Segregated Data

As noted above, none of the current information systems has the capacity to meet the demanding requirements for integrating construction project data among the various stakeholders. Research efforts have been made to address the challenges with the segmented data, while not interfering with the current organizational data structure. Most research suggests exporting all required data through data adaptors into a single location and then supplying the data to the required analysis systems (Ji and AbouRizk 2018). Although a data adaptor enables information systems to access data freely, it does not replace the manual pre-processing of the fragmented raw data. Pereira, et al. (2019) demonstrated the

employment of high-level architecture (HLA) or distributed simulations to connect to various data sources for a data-driven simulations in real time. Nevertheless, the initial construction of HLA is costly. Further, it requires additional services for ongoing maintenance to keep up with the updates from all its connected data sources.

Although a few studies have proposed solutions for the structure and techniques on constructing data-driven analytics from fragmented data, the challenge of repetitive manual manipulation of the non-integrated raw data remains. This research mainly focuses on the automation of critical data pre-processing steps, aiming to bridge the gap between the non-integrated raw construction data with data-driven analytics.

2.3. METHODOLOGY

2.3.1. Framework

The proposed framework is depicted in Figure 2-2.

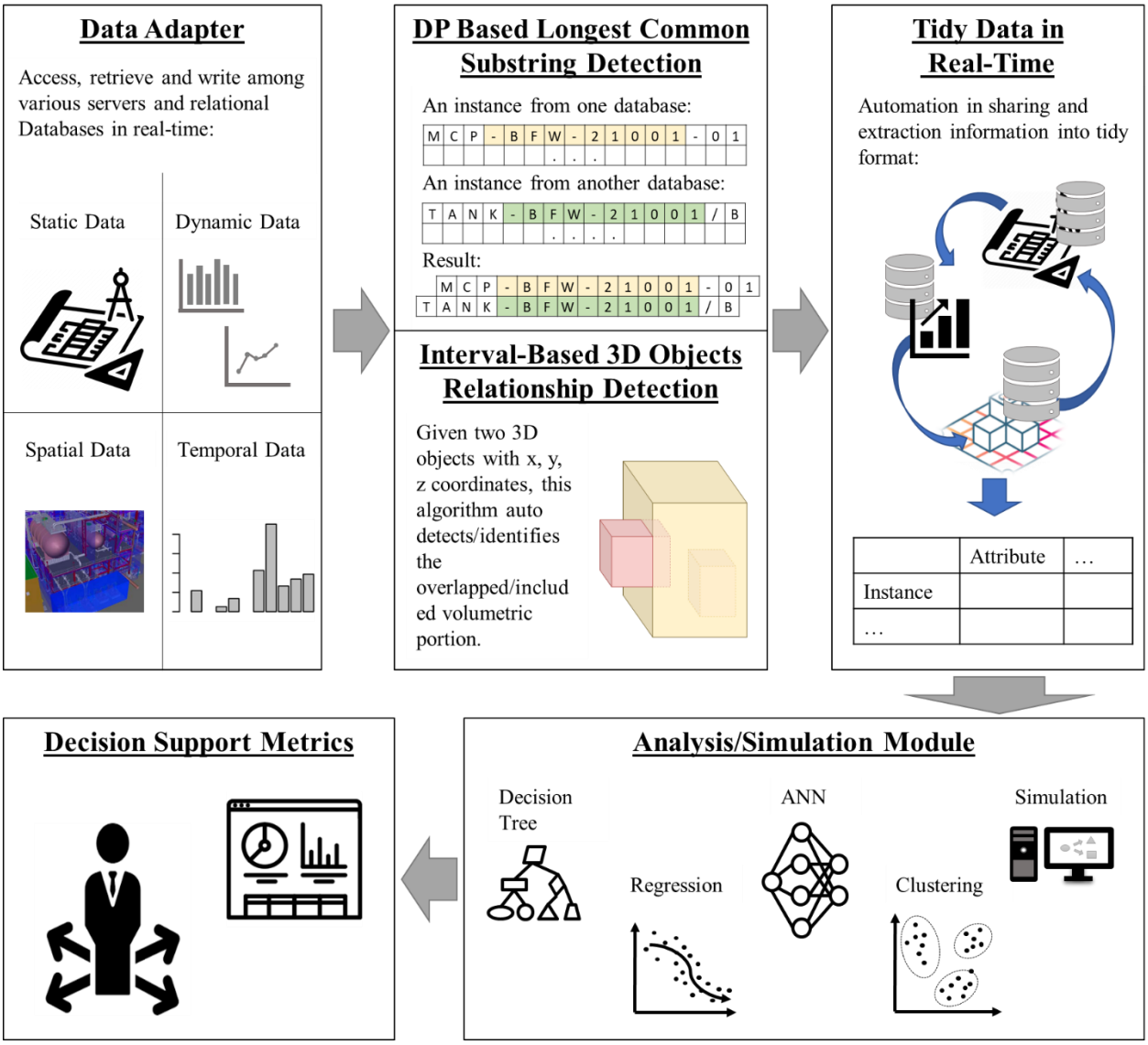


Figure 2-2 The proposed framework

2.3.2. Data Adaptor

As outlined in Figure 2-1, the construction phase relies on data from previous phase(s) (such as specifications and drawings, which are static data as they are less frequently modified); at the same time it generates a large amount of dynamic data (such as cost, safety, quality, and progress).

To access various data sets freely and retrieve relevant tables and store them in a central location, data adaptors are an essential first step within this framework. In this research, a data adaptor was designed using *R* package *RODBC* (v1.3-16; Ripley and Lapsley 2019). Package *RODBC* provides access to various types of databases through an open database connectivity (ODBC) interface. With the data adaptor, data from various sources can be gathered in real-time for further cleaning, wrangling, and mining. Thus, at a very low cost, the company could continuously rely on its existing information systems, as well as enjoy the benefit of real-time access to any given database.

2.3.3. DP-Based Longest Common Substring Algorithm

Although data from different databases is gathered to a central location through the data adaptor, the data structure, level of details, and different naming convention present challenges for further data aggregation. In an industrial construction project, for instance, one section of pipe in the BIM is named as "PIPE-ISO-1234-A1", while it is called "P1-CWA1-ISO-1234-001" in the progress database.

Nevertheless, a technical data set often contains a unique identification document (ID) number or serial number, such as "ISO-1234" from the aforementioned example. Given impossible-to-unify naming conventions due to lack of standardization and coordination, identifying the common serial or ID number and matching data through this shared serial or ID number is a must to create the linkage between various data sets.

As easy as it looks, the process of matching anywhere ten to hundreds of thousands of records from one database to another is labor-intensive. The practical challenge in automating the data aggregation process among various databases comes down to matching objects from one database to another based on shared attributes (or partially-shared attributes).

This problem can be further simplified and categorized into one of the classic problems in string analysis, the longest common substring (Gusfield 1997). For example, if $S_1 = imstr$, and $S_2 = urstr2$, then the longest common substring of S_1 and S_2 is str.

Although the generalized suffix tree is efficient and conceptually simple in finding the longest common substring (Gusfield 1997), this method is difficult to implement. Dynamic programming (DP), first introduced by Richard Bellman (Bellman 1954), has proven effective in solving problems with overlapping subproblems (Nath et al, 2018). With a DP-based algorithm, the problem of finding the longest common substring of S_1 and S_2 with length n and m, respectively, can be solved by filling a $(n+1) \times (m+1)$ matrix.

Specifically, given $S_1 = imstr$, and $S_2 = urstr2$, the detailed steps of the DP algorithm in finding the longest common substring is demonstrated as follows:

Step 1: Insert a place holder character (e.g. #) in front of both strings S_1 and S_2 resulting in $S_1 = \#imstr$, and $S_2 = \#urstr2$.

Step 2: Calculate the length of S_1 and S_2, resulting in 6 and 7, respectively, in this case.

Step 3: Initialize an empty matrix $M_{6\times7}$.

Step 4: Insert 0 for the first column and the first row of matrix M.

Step 5(a): If using i to represent the index of S_1, and j for the index of S_2,

Step 5(b): construct a nested loop. The outer loop starts with $i = 2$, and ends with $i = 6$. The inner loop starts with $j = 2$ and ends with $j = 7$;

Step 5(c): Calculate the matrix $M_{6\times7}$ using the function:

$$M_{6\times7}[i,j] = \begin{cases} M[i-1, j-1] + 1, \\ if\ the\ character\ of\ S_1\ at\ index\ i = the\ character\ of\ S_2\ at\ index\ j \\ 0, otherwise \end{cases}$$

Step 6: Identify the maximum value of the $M_{6\times7}$ together with its indices i, j. The maximum value, which represents the length of the longest common substring, is 3 in this case.

Step 7: Extract the longest common substring from S_1 based the length of the longest common substring (3), and the end index (6), resulting in str.

Figure 2-3 illustrates the results of matrix $M_{6\times7}$ at $i = 4, j = 7$ (a), and $i = 6, j = 7$ (b). As demonstrated, this algorithm is simple, straightforward, and easy to implement. The DP-based algorithm effectively allocates the longest common substring between any two given strings regardless of the length or location of the common substring. As a result, this algorithm was chosen for this research.

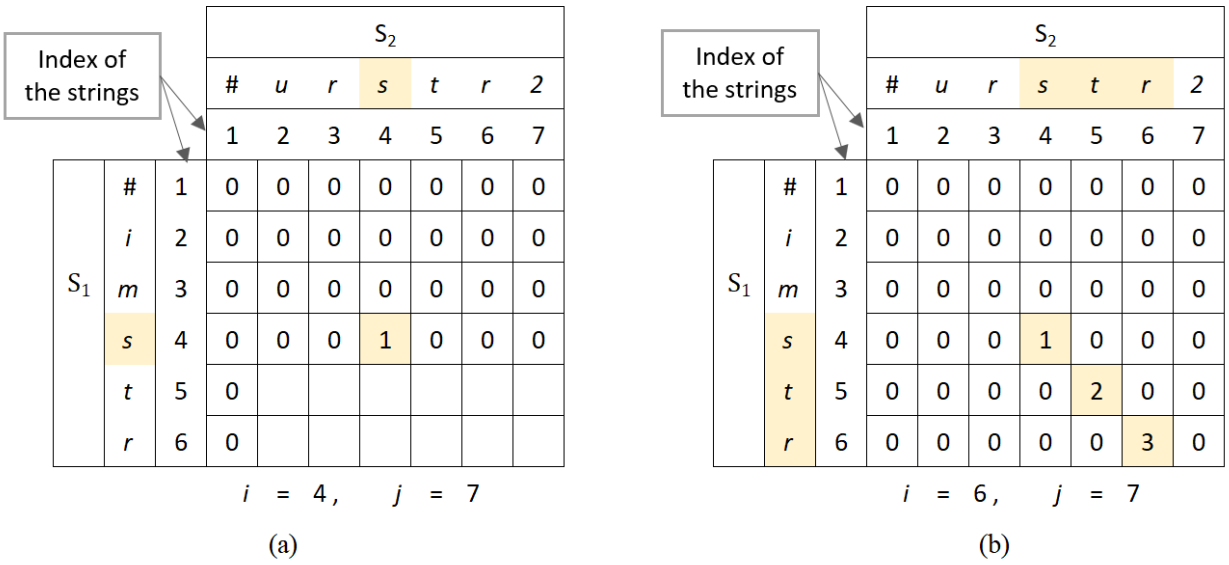


Figure 2-3 Result matrix of the longest common substring example

In replacing the traditionally manual matching object ID task, a DP-based longest common substring algorithm was designed to facilitate the auto-detection of matching key phrases. Since no exact *R* function with the DP-based algorithm existed, a custom *R* function "*LCStr*" was developed. *LCStr* takes three arguments: aString, bString, and minlen; and returns the longest common substring based on the minimum length defined by "minlen". Figure 2-4

presents the pseudocode of this function. This custom function was validated through an artificial dataset, as well as a case study, then wrapped as a library and uploaded to GitHub (Li et al, 2019; Wu et al. 2020).

```
function LCStr(aString, bString, minlen)
    LCS = matrix(nrow = length(aString)+1, ncol = length(bString)+1)
    lengthOfSubstring = -1
    finalIndex = -1
    LCS[1, ] = 0
    LCS[ ,1] = 0
    for (i in range(1, length(aString)+1))
        for (j in range(1,length(bString)+1))
            if (aString[i-1] == bString[j-1])
                LCS[i][j] = LCS[i-1][j-1] + 1
                if (lengthOfSubstring<LCS[i][j])
                    lengthOfSubstring = LCS[i][j]
                    finalIndex = i
            else
                LCS[i][j] = 0

    if (lengthOfSubstring >= minlen)
        return aString[finalIndex - lengthOfSubstring, finalIndex]
    else
        return "no result"
```

Figure 2-4 Pseudocode of custom *R* function *LCStr*

2.3.4. Interval-based 3D Objects Relationship Detection

In addition to temporal data, increasing spatial data (such as 3D models) are shared and used in the construction phase. Spatial information is often stored in object databases, in the form of objects and classes, which resembles object-oriented programming languages. Consequently, this type of data is significantly different from temporal data or dynamic data.

Heavy data pre-processing steps are involved to reuse the 3D model for construction management purposes (Preidel, et al. 2017), including summarizing construction work, progress, and/or engineering objects by contractor-defined physical envelops—a critical project management function. Construction work areas and pre-fabricated modules are common types of physical envelopes that are manually defined later in the project life cycle

by downstream organizations (e.g. contractors, fabricators); as a result, to identify each 3D element from the 3D model to a list of physical envelop boundaries is a common exercise performed by general contractors.

To automate this common spatial data pre-processing task, an interval-based 3D object relationship detection algorithm was designed (Wu and AbouRizk 2020). As any 3D object can be easily exported from the model as its boundaries on three coordinates (i.e. maxima and minima on x, y, and z coordinates), the relationships between 3D objects can be first detected on each coordinate, then conclude based on results on all coordinates. Specifically, 1) if and only if the boundaries (i.e. intervals) on all coordinates of 3D object a are included in the boundaries of the 3D object b, then 3D object a is within 3D object b; 2) if and only if the boundaries of 3D object c on all coordinates overlap the boundaries of the 3D object d on all coordinates, then 3D object c overlaps 3D object d. Further, as inclusion is a special case of overlap, conditions of inclusion will be checked first.

Although no exact *R* function detects 3D object relationships, package *intervals* created by Bourgon (2015) has a full list of functions for working with and comparing sets of intervals. Built upon main functions such as "interval_included" and "interval_overlap" from the package intervals, a custom *R* function "*detecte3Dr*" was developed during this research (Wu and AbouRizk 2020). *detecte3Dr* takes four arguments: tablefrom, tableto, closedfrom, closedto; and returns a result table in the form of tabletowith two additional columns – "Within ID" and "Overlap ID". Input arguments tablefrom and tableto (both in tabular format) each have 6 columns, identifying the boundaries of the 3D objects (i.e. maxima and minima on x, y, and z coordinates). Input arguments closedfrom and closedto each are a two-element vector with either "TRUE" or "FALSE" required to indicate whether endpoints are included (i.e. "TRUE" for including endpoint). Each row of the result table presents an element from

the tableto; and the column “Within ID” and “Overlap ID” indicates which row index of input closedfrom these tableto elements is included or overlapped. Figure 2-5 illustrates the logical process of this function through a process flowchart. This custom function was validated through an artificial dataset first, then demonstrated through a case study. To benefit the greater construction industry, this custom function is wrapped as a *R* library and uploaded to GitHub (Wu and AbouRizk 2020).

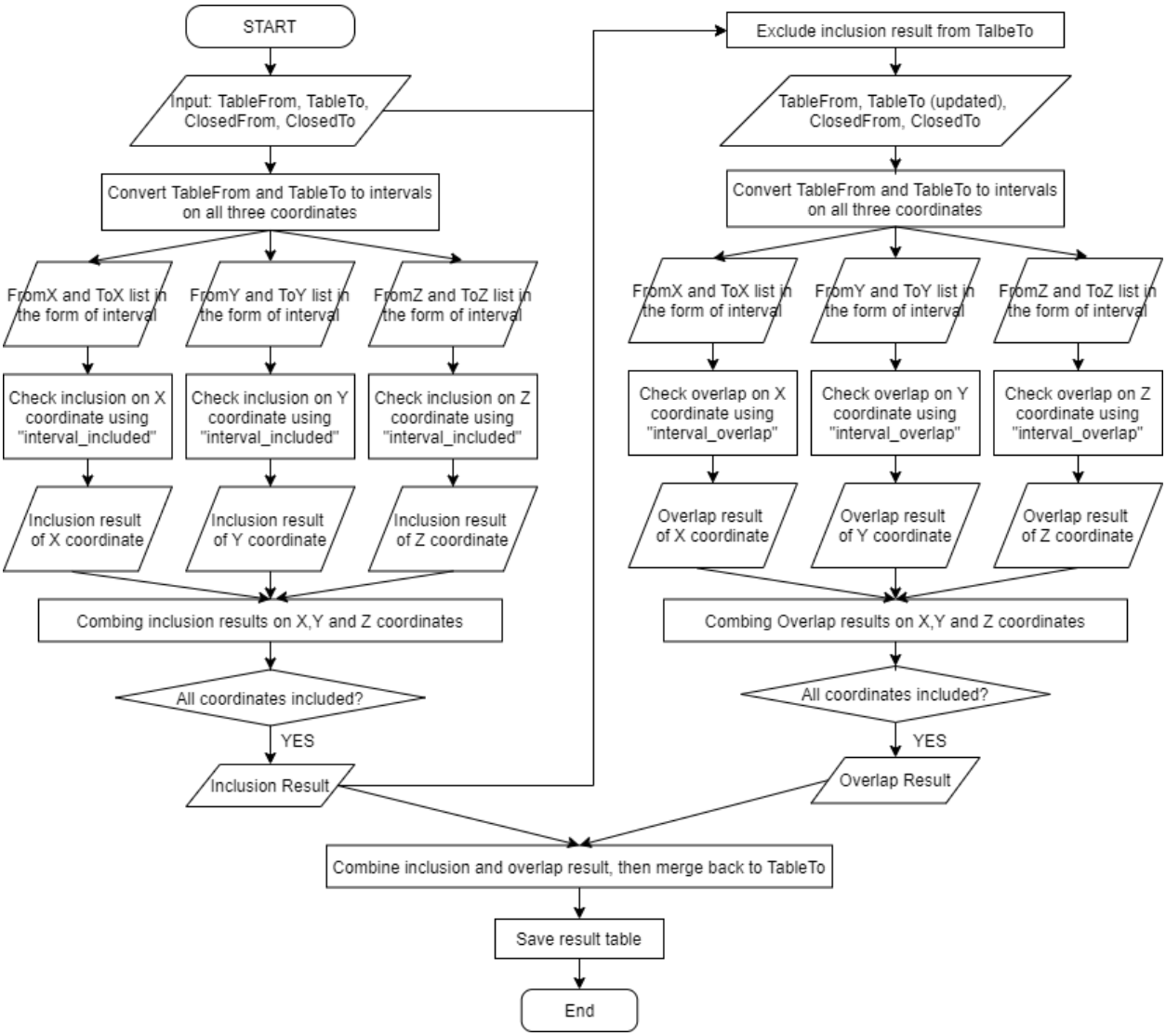


Figure 2-5 Flow chart of custom *R* function *detecte3Dr*

2.3.5. Real-Time Connection Among Relational Database

With the automation of identifying common key words, and additionally—in the case of spatial data—automation of identifying 3D object relationships, the relationship among stand-alone databases (i.e. one-to-one, one-to-many, many-to-many) can be identified without human intervention.

The desired data aggregation thus can be achieved through various SQL commands. In this research, *R* package *dplyr* (v0.8.3; Wickham et al 2020) was used, as it provides various types of joining functions, including mutating joints, filtering joins, and nesting joins, catering to different relationship types and merging requirements.

2.3.6. Analysis Module and Decision Support Matrix

Upon gathering, linking and aggregating fragmented data to a tidy format, desired machine learning algorithms and/or simulation models can be applied (Han et al. 2011). As a result, the proposed framework provides a highly-automated and efficient channel to quickly process raw fragmented data for any data-driven decision supports.

2.4. VALIDATION

Validations of the DP-based longest common substring algorithm and interval-based 3D objects relationship detection algorithms was conducted prior to the case study. Artificial data sets were randomly generated to exhaust all possible cases.

2.4.1. DP-Based Longest Common Substring Algorithm

A list of strings was randomly generated including numbers, symbols, and upper- and lower-case letters as shown in the 1st column in Table 2-1. Based on this list, five types of

modifications – 1) remain identical, 2) addition, 3) subtraction, 4) subtraction and addition, 5) regenerate randomly, were conducted to generate the second list of strings (3rd column in Table 2-1). For example, through "subtraction and addition" type of modification, item "#pLtjgF905jlsg" from column "List 1" is modified to "#pLtjgF9lsg-155", where "05j" was removed and "-155" was added to the end.

Table 2-1 Randomly generated lists of strings for validation

List 1	Modification	List 2	Manual Identified Longest Common Substring
ajM-VN-P0XQ5	addition	1ajM-VN-P0XQ5_2fg	ajM-VN-P0XQ5
Y*a429RAxJc	addition	Y*a429RAx(24)Jc	Y*a429RAx
QH#L&NxJeHma	addition	$we[QH#L&NxJeHma	QH#L&NxJeHma
3VS.TMi7+mdn	addition	3VS.TMi7+mdn1T.d	3VS.TMi7+mdn
dK(dvZM)Rrcn	remain identical	dK(dvZM)Rrcn	dK(dvZM)Rrcn
ia2KzXZS5n_e3	remain identical	ia2KzXZS5n_e3	ia2KzXZS5n_e3
0SAGuE3I0I	subtraction	0SAGuE3I	0SAGuE3I
sfS\3dRGy5m	subtraction	\3dRGy5m	\3dRGy5m
7s8O(5Ng_jlZ	subtraction	7s8O(5_jlZ	7s8O(5
VqdZjOzOyp	addition	Vqd**ZjOzOyp	ZjOzOyp
Z8_wAo-0dGZx	addition	0WZ8_wAo-0dGZx1e5	Z8_wAo-0dGZx
Cg[pux$eYs)Rk	subtraction	pux$eYs)R	pux$eYs)R
PgvdeToxvA	addition	P*gvdeToxvA.1	gvdeToxvA
Lc-Zw@D2cVvS	subtraction and addition	Lc-Zw@D2c.23	Lc-Zw@D2c
qrgmF2YIL4	subtraction and addition	1gw_rgmF2YIL4	rgmF2YIL4
16vz*9n-Q143	subtraction and addition	16*9n-Q143-YI24	*9n-Q143
#pLtjgF905jlsg	subtraction and addition	#pLtjgF9lsg-155	#pLtjgF9
O_Jh(KGk1-gpk5	generate randomly	3QHxqr6%s8afj[	NA
a6sQzj231=gQ	generate randomly	2Pgz%8z)fgW#0	NA

A nested loop was built to check the potential longest common substrings between List 1 and List 2. For each item from List 1, all possible common substrings longer than 5 digits from

List 2 were identified (with argument “minlen” set to be “5”) and vice versa. The results are captured in Table 2-2

Table 2-3, with manually identified longest common substring listed in last column for comparison.

Table 2-2 DP-based longest common substring result table for List 1

Detect	Machine Identified Longest Common Substring	List 1	Manually Identified Longest Common Substring
TRUE	ajM-VN-P0XQ5	ajM-VN-P0XQ5	ajM-VN-P0XQ5
TRUE	Y*a429RAx	Y*a429RAxJc	Y*a429RAx
TRUE	QH#L&NxJeHma	QH#L&NxJeHma	QH#L&NxJeHma
TRUE	3VS.TMi7+mdn	3VS.TMi7+mdn	3VS.TMi7+mdn
TRUE	dK(dvZM)Rrcn	dK(dvZM)Rrcn	dK(dvZM)Rrcn
TRUE	ia2KzXZS5n_e3	ia2KzXZS5n_e3	ia2KzXZS5n_e3
TRUE	0SAGuE3I	0SAGuE3I0I	0SAGuE3I
TRUE	\3dRGy5m	sfS\3dRGy5m	\3dRGy5m
TRUE	7s8O(5	7s8O(5Ng_jlZ	7s8O(5
TRUE	ZjOzOyp	VqdZjOzOyp	ZjOzOyp
TRUE	Z8_wAo-0dGZx	Z8_wAo-0dGZx	Z8_wAo-0dGZx
TRUE	pux$eYs)R	Cg[pux$eYs)Rk	pux$eYs)R
TRUE	gvdeToxvA	PgvdeToxvA	gvdeToxvA
TRUE	Lc-Zw@D2c	Lc-Zw@D2cVvS	Lc-Zw@D2c
TRUE	rgmF2YIL4	qrgmF2YIL4	rgmF2YIL4
TRUE	*9n-Q143	16vz*9n-Q143	*9n-Q143
TRUE	#pLtjgF9	#pLtjgF905jlsg	#pLtjgF9
FALSE		O_Jh(KGk1-gpk5	NA
FALSE		a6sQzj231=gQ	NA

Table 2-3 DP-based longest common substring result table for List 2

Detect	Machine Identified Longest Common Substring	list.2	Manually Identified Longest Common Substring
TRUE	ajM-VN-P0XQ5	1ajM-VN-P0XQ5_2fg	ajM-VN-P0XQ5
TRUE	Y*a429RAx	Y*a429RAx(24)Jc	Y*a429RAx
TRUE	QH#L&NxJeHma	$we[QH#L&NxJeHma	QH#L&NxJeHma
TRUE	3VS.TMi7+mdn	3VS.TMi7+mdn1T.d	3VS.TMi7+mdn
TRUE	dK(dvZM)Rrcn	dK(dvZM)Rrcn	dK(dvZM)Rrcn
TRUE	ia2KzXZS5n_e3	ia2KzXZS5n_e3	ia2KzXZS5n_e3
TRUE	0SAGuE3I	0SAGuE3I	0SAGuE3I
TRUE	\3dRGy5m	\3dRGy5m	\3dRGy5m
TRUE	7s8O(5	7s8O(5_jlZ	7s8O(5
TRUE	ZjOzOyp	Vqd**ZjOzOyp	ZjOzOyp
TRUE	Z8_wAo-0dGZx	0WZ8_wAo-0dGZx1e5	Z8_wAo-0dGZx
TRUE	pux$eYs)R	pux$eYs)R	pux$eYs)R
TRUE	gvdeToxvA	P*gvdeToxvA.1	gvdeToxvA
TRUE	Lc-Zw@D2c	Lc-Zw@D2c.23	Lc-Zw@D2c
TRUE	rgmF2YIL4	1gw_rgmF2YIL4	rgmF2YIL4
TRUE	*9n-Q143	16*9n-Q143-YI24	*9n-Q143
TRUE	#pLtjgF9	#pLtjgF9lsg-155	#pLtjgF9
FALSE		3QHxqr6%s8afj[	NA
FALSE		2Pgz%8z)fgW#0	NA

2.4.2. Interval-Based 3D Objects Relationship Detection

Two sets of 3D object boundaries were designed including real number, integer, positive and negative values (Table 2-4 and Table 2-5). The relationship between every 3D object on List 2 to every 3D object on List 1 was manually examined, and the result is shown in the last column of Table 2-5.

The two tables (Table 2-4 and Table 2-5) were fed into the custom function, "*detecte3Dr*". Table 2-6 presents both the manual detection result (the 7th column) and the machine

detection result (the 8th and 9th columns). If both 8th and 9th columns in Table 2-6 in a given row are "NA", it represents this specific 3D object on List 2 is outside of any of the 3D objects on List 1. If the 8th column is a number (e.g. "1") and 9th column is "NA", it represents this 3D object from List 2 is within the specific 3D object, indicated by the number shown in the 8th column of List 1 (in this case this object from List 2 is within the object with "List1 ID" = "1" from List 1). If the 8th column is "NA" and 9th column is a number (e.g. "2"), it represents this 3D object from List 2 overlaps the 3D object indicated by the number shown in the 9th column of List 1 (in this case this object from List 2 overlaps the object with "List1 ID" = "2").

Table 2-4 List 1 of the boundaries of the 3D object

List1 ID	MIN.X	MAX.X	MIN.Y	MAX.Y	MIN.Z	MAX.Z
1	100	200	100	200	100	200
2	-100.01	-0.01	-100.01	-0.01	-100.01	-0.01

Table 2-5 List 2 of the boundaries of the 3D object

MIN.X	MAX.X	MIN.Y	MAX.Y	MIN.Z	MAX.Z	Manual Result
10	20	10	20	50	150	Outside
100.5	120	105.5	150.5	50.5	150	Overlap of 1
100.5	120	10	20	50	150	Outside
50	250	150	160.1	150.1	160.1	Overlap of 1
150.1	160.1	150.1	160	150	160	Within 1
10	150	150	160	150	160	Overlap of 1
-190	-180	-190	-180	-150	-50	Outside
-99.5	-80	-194.5	-49.5	-149.5	-50	Overlap of 2
-150	50	-50	-39.9	-49.9	-39.9	Overlap of 2
-49.9	-39.9	-49.9	-40	-50	-40	Within 2
-190	-50	-50	-40	-150	-140	Outside
-150	180.5	-50.5	150	-150	150.5	Overlap of 1,2

Table 2-6 Validation result of checking List 2 against List 1

MIN.X	MAX.X	MIN.Y	MAX.Y	MIN.Z	MAX.Z	Manual Result	Within ID	Overlap ID
10	20	10	20	50	150	Outside	NA	NA
100.5	120	105.5	150.5	50.5	150	Overlap of 1	NA	1
100.5	120	10	20	50	150	Outside	NA	NA
50	250	150	160.1	150.1	160.1	Overlap of 1	NA	1
150.1	160.1	150.1	160	150	160	Within 1	1	NA
10	150	150	160	150	160	Overlap of 1	NA	1
-190	-180	-190	-180	-150	-50	Outside	NA	NA
-99.5	-80	-194.5	-49.5	-149.5	-50	Overlap of 2	NA	2
-150	50	-50	-39.9	-49.9	-39.9	Overlap of 2	NA	2
-49.9	-39.9	-49.9	-40	-50	-40	Within 2	2	NA
-190	-50	-50	-40	-150	-140	Outside	NA	NA
-150	180.5	-50.5	150	-150	150.5	Overlap of 1,2	NA	1
-150	180.5	-50.5	150	-150	150.5	Overlap of 1,2	NA	2

2.5. CASE STUDY

2.5.1. Overview

The proposed framework was demonstrated through a case study using historical project data. This project was a multi-billion-dollar industrial construction project located in Alberta, Canada. The project ran from 2014 to 2016; and is a typical-to-the-region oil sand secondary extraction project.

At the request of our industrial partner, the end product was to generate a real-time S-curves for each discipline per contractor-identified module classes. Not as straightforward as it sounds, the information required for the deliverables traditionally are stored in four separate

information systems—including BIM (3D model), an in-house developed progress tracking system, module lift schedule, and a spreadsheet of module class list—without obvious links among one and another.

To meet the specific requirements, the proposed framework was adapted as shown in Figure 2-6. As per the industrial partner's request, the proposed method was applied to the databases by each discipline. For illustration purposes, however, only piping discipline is presented in this chapter.

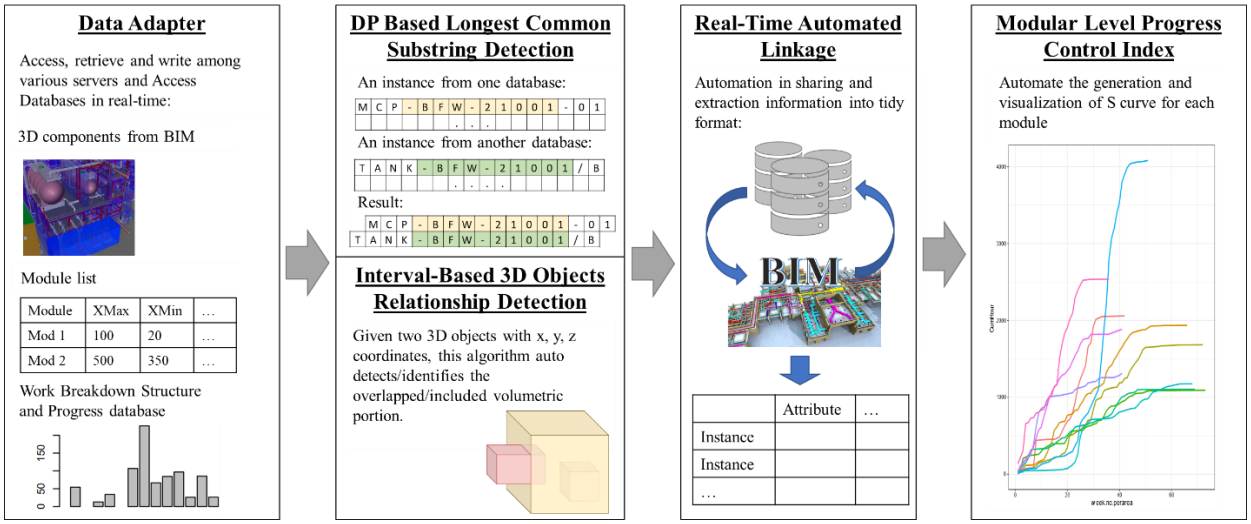


Figure 2-6 Adapted framework of the case study

With the help of the data adaptor, four tables were retrieved from the above-mentioned sources, as shown in Figure 2-7. The progress table (shown as a sample section in Figure 2-8) contains the following information: discipline, progress activity, and weekly records of labor-hours earned for each activity. The component table (shown as a sample section in Figure 2-9) extracted from the 3D model contains the coordinates of each component (the smallest modeling item at the geometry level), the component name (a long string that contains most of the design-related information), and discipline.

Except the one-to-many relationships between the module class table and the module list table, none of the other relationships among these tables as shown in Figure 2-7 existed. Since work breakdown structure does not necessarily follow the division of modules, progress activities as the lower level of work breakdown structure can not be summarized to the modular level. Nevertheless, through 3D model, progress activities can link to BIM components, which has its physical coordinates. Then, by comparing each component's coordinates with each module's coordinates, an indirect link can be generated for progress activities to be summarized to the modular level.

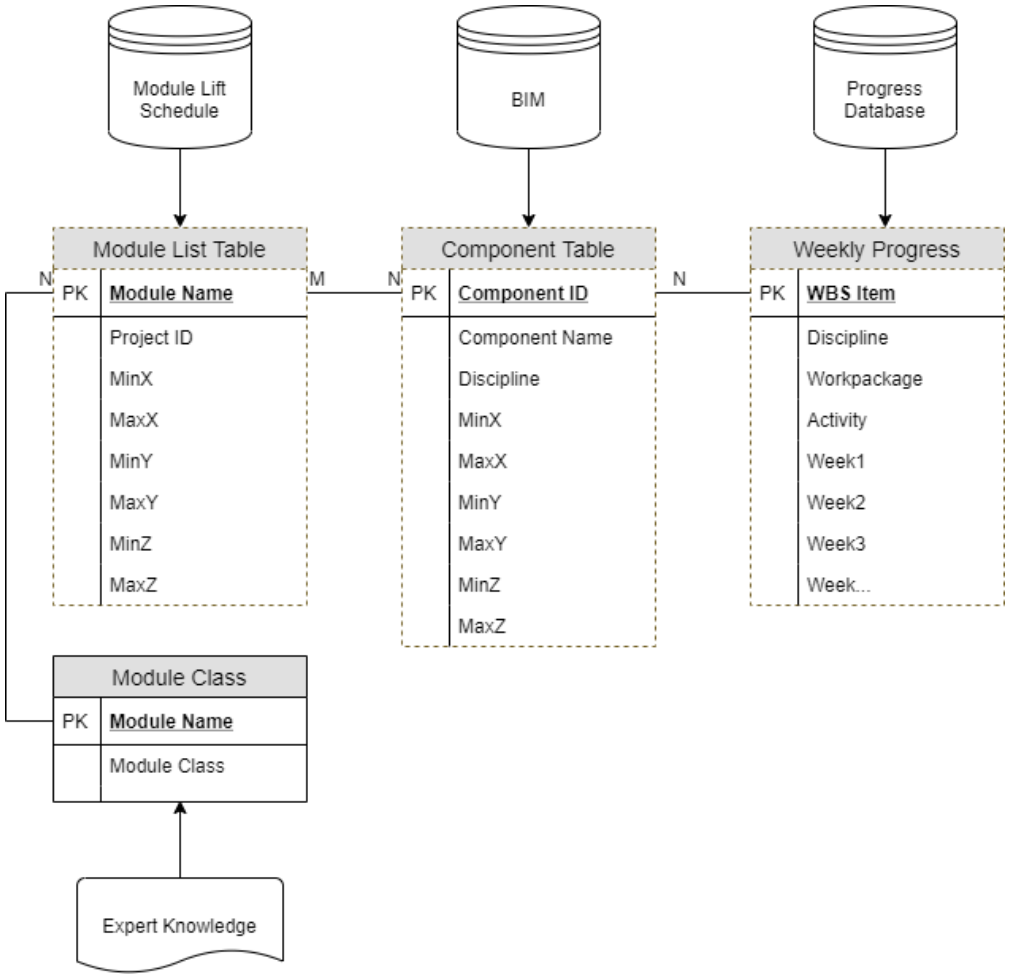


Figure 2-7 Entity relationship diagram of module list table, component table and progress table

Discipline	Progress.Activity	4.5.2015	4.12.2015	4.19.2015	4.26.2015	5.3.2015	5.10.2015
Piping	ISO-MCP-00-BFW-41039-01-00M50						
Piping	ISO-MCP-00-BFW-41039-01-00M56						
Piping	ISO-MCP-00-BFW-41093-01-00M50						
Piping	ISO-MCP-00-BFW-41093-01-00M56						
Piping	ISO-MCP-00-D-16002-01-00A01	19	26	22			2
Piping	ISO-MCP-00-D-16004-01-00A01		65				7
Piping	ISO-MCP-00-D-18001-01-00A01	18	33	25			2
Piping	ISO-MCP-00-D-18003-01-00A01	4	24	37			2
Piping	ISO-MCP-00-D-18005-01-00A01			35			3
Piping	ISO-MCP-00-D-18006-01-00A01		17	18			3
Piping	ISO-MCP-00-D-18008-01-00A01		18				
Piping	ISO-MCP-00-D-18009-01-00A01		1	6			
Piping	ISO-MCP-00-DW-43014-01-00M50						
Piping	ISO-MCP-00-DW-43014-01-00M56						
Piping	ISO-MCP-00-FG-21001-01-00A01						
Piping	ISO-MCP-00-FG-21001-01-00M01		16				8
Piping	ISO-MCP-00-FG-21001-01-00M03					21	
Piping	ISO-MCP-00-FG-21001-01-00M04						
Piping	ISO-MCP-00-FG-21001-01-00M06	4					
Piping	ISO-MCP-00-FG-21001-01-00M52			27	10		

Figure 2-8 Sample section of progress table from progress database

CompName	MinX	MinY	MinZ	MaxX	MaxZ	MaxY	Discipline
ATTACHMENT 1 of BRANCH /G-80904-01-00M52/B1	1577037	5265786	103390	1577547	103990	5265938	Piping
VALVE 1 of BRANCH /PE-99014-01-01A04-V3/B1	1555725	5327853	102598	1555939	103143	5328069	Piping
REDUCER 2 of BRANCH /PG-20004-01-00M02/B1	1566636	5247141	105073	1566992	105479	5247547	Piping
TUBE 3 of BRANCH /WW-44011-01-01A01/B4	1489470	5364608	100950	1489690	101169	5364758	Piping
ATTACHMENT 3 of BRANCH /G-80280-01-01A03/B1	1520763	5312325	103327	1520838	103402	5312595	Piping
TUBE 3 of BRANCH /SB-15038-01-04M52/B2	1550470	5486338	103921	1551291	104742	5486662	Piping
ATTACHMENT 3 of BRANCH /FG-21001-01-00M01/B1	1575900	5262750	103540	1576033	103639	5263050	Piping
GASKET 4 of BRANCH /V-63019-01-01A01/B1	1498600	5360863	106523	1499010	106933	5360866	Piping
TUBE 1 of BRANCH /PW-40001-01-01A03/B1	1508577	5339462	102006	1508850	102256	5339735	Piping
TUBE 11 of BRANCH /PW-40034-01-01A02/B1	1567030	5358570	101151	1567354	102779	5358894	Piping
ELBOW 2 of BRANCH /RG-25001-01-00M03/B1	1849349	5225163	103156	1849867	103673	5225436	Piping
ATTACHMENT 1 of BRANCH /K-70033-01-04A02/B7	1717198	5495923	97900	1717298	97901	5496023	Piping
CLOSURE 1 of BRANCH /SL-12020-01-01A02/B12	1596704	5363161	103239	1596731	103283	5363188	Piping
TUBE 3 of BRANCH /BFW-41003-01-03A02/B1	1546456	5156916	101778	1546780	102102	5157209	Piping
FLANGE 2 of BRANCH /IA-30096-01-03A01/B1	1481392	5183871	105471	1481622	105701	5183948	Piping
TUBE 3 of BRANCH /HPF-60110-01-01A03/B1	1522016	5318416	114776	1522184	115005	5318584	Piping
FBLIND 1 of BRANCH /PW-40005-01-01A03/B2	1528663	5329500	101070	1528943	101096	5329780	Piping

Figure 2-9 Sample section of component prosperity table from BIM

2.5.2. Generate Links Among Segmented Data Sets

The generation of a one-to-many relationship between the progress table and component table started with a careful manual examination of the data. Patterns of naming conventions for both tables were unveiled. For instance, a majority (3,709 out of 4,766 entries) of the piping progress names start with "ISO-MCP-##-." The pattern is directly followed by a serial identification code combined with letters, numbers, and symbols such as "PW-40005-01-01A03." The remaining 1,057 entries include field-run pipe and piping budget plugs where no obvious patterns could be determined. In the meantime, the piping serial identification code (e.g. "PW-40005-01-01A03") prevailingly exists (105,503 out of 105,716) in the component name field.

With the unveiled naming convention, the tasks of linking the progress table with the component table could be simplified as identifying the longest common substring between the "component name" from the component table with "Progress activity" column from the progress table through the longest common substrings. With the minimal length of the common substring set as 5, 3,075 progress activities were mapped to 98,640 components.

For additional validation, manual extraction of the piping serial identification code from the progress database was performed, followed by a partially match algorithm developed to partially match each piping serial identification code that exists in the progress table with component names. The manual linkage allocated 2,697 progress activities to 90,224 components.

By careful comparison of the manual linkage result and the proposed DP algorithm result, the following were discovered: first, all of the linkages found through manual extraction were also found through the proposed algorithm; second, the DP-based algorithm detected extra

linkages that did not follow the pattern, "ISO-MCP-##-."; third the increased linkages (378 progress activities to 8,416 BIM components) were validated through manual inspection. Further, any progress activities that could not find a link to BIM elements were removed, since without location information, these activities would not be able to summarize into the modular level.

The many-to-many relationship between module list table and the component table was generated by checking each component's coordinates with the module's using the customized function "*detecte3Dr* ".

With the many-to-many relationship between the module list and the component table, and the many-to-one relationship between component table and weekly progress table created, the four data sources were linked. Thus, progress could be summarized to module-level without repetitive manual manipulation in every reporting period.

In this case study, the 208 modules were manually classified into 26 classes (e.g. pipe rack module, electrical building, boilers, pumps, etc.). This information was kept in a stand-alone spreadsheet, based on the experts' knowledge on design/type of the structure (Figure 2-7). With S curves plotted side-by-side for modules belong to the same class, the experts can effectively evaluate the performance, progress, and hence proactively plan for the future. For instance, as shown in Figure 2-10, the S curves for all the modules that belong to high-density pipe rack are plotted on the single chart.

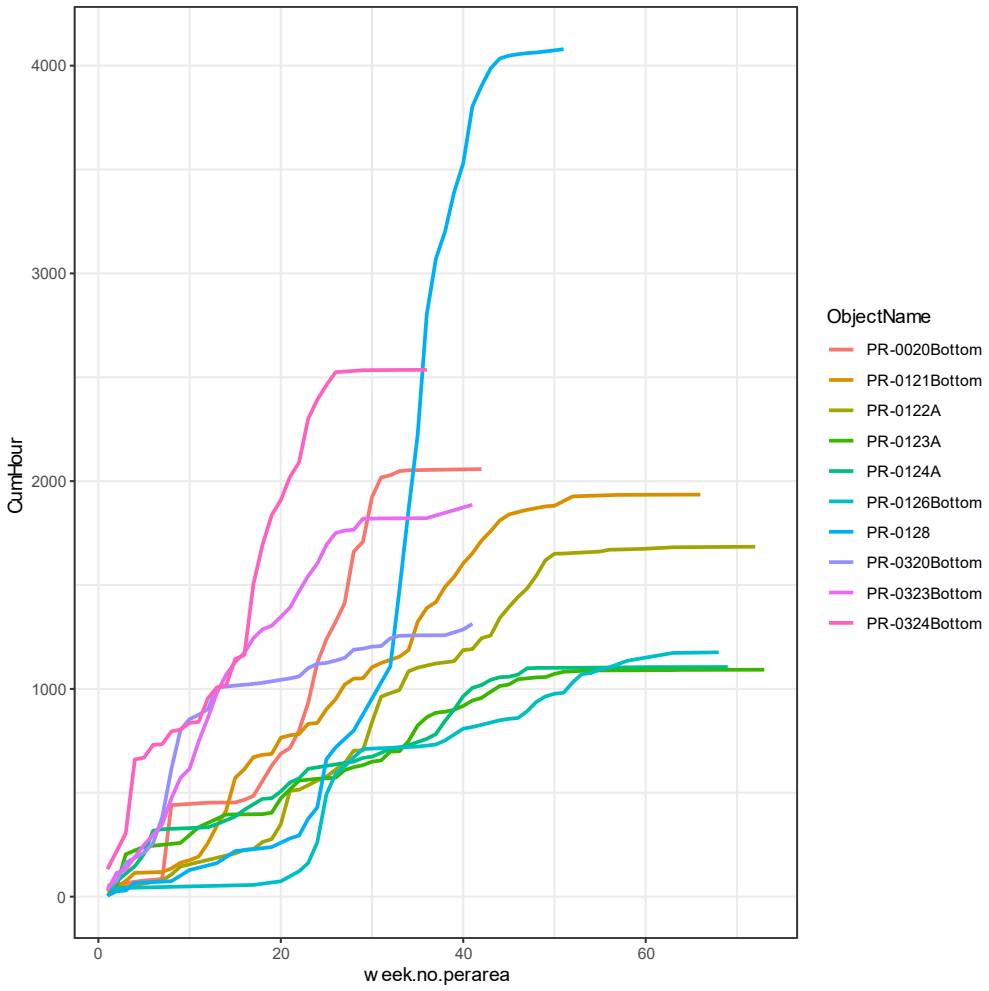


Figure 2-10 S curves for high-density pipe rack module class

2.5.3. Synopsis

In summary, the proposed framework effectively and efficiently addressed the practical challenges within the industrial construction regarding fragmented construction data. Specifically, the DP-based longest common substring algorithm successfully auto-detected the common keywords between the selected attributes from different databases. The interval-based 3D objects relationship detection algorithm effectively identified the relationship between two sets of 3D objects. Once the relationship types (i.e. one-to-one, one-to-many, many-to-many) among databases were identified, the desired aggregation of information

could be achieved through various SQL commands. Thus, it freed the domain experts from the periodic manual manipulation of the non-integrated construction data and allowed them to spend their time in understanding the data and producing meaningful matrices/indices for critical decision support.

2.6. CONCLUSION

This research proposes a framework to effectively process fragmented construction raw data in achieving meaningful construction decision supports. Within the proposed framework, the data adaptor provides real-time database access; the DP-based longest common substring algorithm and the interval-based 3D objects relationship detection algorithm automates two commonly presented data pre-process tasks for identifying relationship types among databases; SQL functions integrate information stored from various databases into a tidy format; lastly, various data mining techniques and simulation models can effectively process the tidy data and produce meaningful decision supports indices.

The proposed framework achieves the automation in information sharing and aggregating regardless of the data origins or data types. Through the two custom functions, spatial data and temporal data stored in various types of databases can be effectively aggregated to the desired level. Thus, it significantly reduces the repetitive manual data pre-processing, and increases information flow in a real-time manner without drastically modifying the existing information system structure. Ultimately, the proposed framework improved the data quality and promotes construction data utilization for critical decision-support process.

The core functions (i.e. DP-based longest common substring algorithm and the interval-based 3D objects relationship detection algorithm) were validated through artificial data sets. Further, the practicability and feasibility of the framework have been demonstrated through

a mega-sized industrial construction project. The custom R functions were successfully performed on the large data set (over 100,000 records) and produced the desired results. Additionally, these two custom functions written in R code for the DP-based longest common substring algorithm and interval-based 3D objects relationship detection algorithm have been generalized as public R libraries—"Chrisfufu/LongestCommonSubString" and "XiaomoLing/Detect3DRelation" respectively. These packages can be accessed from GitHub for the public.

Despite advancements, this research can be furthered upon addressing the following aspects. First, optimizing the longest common substring algorithm to reduce the running time. The running time can be improved by replacing the DP-based algorithm with a generalized suffix tree. Reduced computational complexity from $O(m \times n)$ to $O(m + n)$ is expected given string lengths m, and n. Second, identifying other common repetitive manual data manipulation tasks and applying computer science, applied mathematics, and/or applied statistic algorithms to increase automation. Third, although the proposed framework provides effective solutions to the common challenges of non-integrated construction raw data, it did not address the root cause of the fragmentation of the information systems in construction industry.

2.7. ACKNOWLEDGMENTS

This project was supported by a Collaborative Research and Development Grant (CRDPJ 492657) from the Natural Sciences and Engineering Council of Canada.

2.8. REFERENCES

Adriaanse, A., Voordijk, H., and Dewulf, G. 2010. "Adoption and use of interorganizational ICT in a construction project." Journal of Construction Engineering and Management, 136(9): 1003-1014.

Agarwal, R., Chandrasekaran, S., and Sridhar, M. 2016. "Imagining construction's digital future." *McKinsey and Company*.

Arshad, M. F., Thaheem, M. J., Nasir, A. R., and Malik, M. S. A. 2019. "Contractual risks of building information modeling: Toward a standardized legal framework for design-bid-build projects." *Journal of Construction Engineering and Management*, 145(4): 04019010. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001617

Ashcraft, H. W. 2008. "Building information modeling: A framework for collaboration." *Construction Lawyer* 28(3): 1–14. Accessed May 12, 2021, https://heinonline.org/HOL/Page?collection=journals&handle=hein.journals/conlaw28&id=126&men_tab=srchresults

Azhar, S. 2011. "Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry." *Leadership and Management in Engineering*, 11(3), 241-252.

Babic, N. Č., Podbreznik, P., & Rebolj, D. (2010). Integrating resource production and construction using BIM. *Automation in Construction*, 19(5), 539-543. https://doi.org/10.1016/j.autcon.2009.11.005

Barbosa, F., Woetzel, J., Mischke, J., Ribeirinho, M. J., Sridhar, M., Parsons, M., Bertram, N. and Brown, S. 2017. "Reinventing construction through a productivity revolution." *McKinsey Global Institute*.

Behzadan, A. H., Aziz, Z., Anumba, C. J., and Kamat, V. R. 2008. "Ubiquitous location tracking for context-specific information delivery on construction sites." *Automation in Construction*, *17*(6), 737-748.

Behzadan, A. H., Menassa, C. C., and Pradhan, A. R. 2015. "Enabling real time simulation of architecture, engineering, construction, and facility management (AEC/FM) systems: a review of formalism, model architecture, and data representation." *ITcon*. 20: 1-23.

Bellman, R. (1954). "The theory of dynamic programming." *Bulletin of the American Mathematical Society*, *60*(6), 503-515.

Berteaux, F., and Javernick-Will, A. 2015. "Adaptation and integration for multinational project-based organizations." *Journal of Management in Engineering*, 31(6), 04015008.

Bourgon, R., 2015. *intervals: Tools for Working with Points and Intervals*. R package version 0.15.1. https://CRAN.R-project.org/package=intervals

Bowden, S., Dorr, A., Thorpe, T., and Anumba, C. 2006. "Mobile ICT support for construction process improvement." *Automation in Construction*, *15*(5), 664-676.

Chassiakos, A. P., and Sakellaropoulos, S. P. 2008. "A web-based system for managing construction information." *Advances in Engineering Software*, 39(11), 865-876.

Chung, B. Y., Skibniewski, M. J., Lucas Jr, H. C., and Kwak, Y. H. 2008. "Analyzing enterprise resource planning system implementation success factors in the engineering–construction industry." *Journal of Computing in Civil Engineering*, 22(6), 373-382. https://doi.org/10.1061/(ASCE)0887-3801(2008)22:6(373)

Eastman, C., Lee, J. M., Jeong, Y. S., and Lee, J. K. 2009. "Automatic rule-based checking of building designs." *Automation in Construction*, 18(8), 1011-1033. https://doi.org/10.1016/j.autcon.2009.07.002

Fan, G. G. 2018. "Customized Manufacturing Enterprise Resource Planning System for Offsite Modular Light Gauge Steel Construction."

Forcada, N., Casals, M., Roca, X., and Gangolells, M. 2007. "Adoption of web databases for document management in SMEs of the construction sector in Spain." *Automation in Construction*, *16*(4), 411-424.

"Frequently Asked Questions About the National BIM Standard-United States - National BIM Standard - United States". Nationalbimstandard.org. Archived from the original on 16 October 2014. Retrieved 17 October 2014.

Gibson, N., Holland, C. P., and Light, B. (1999, January). Enterprise resource planning: a business approach to systems development. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences*. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers (pp. 9-pp). IEEE.

Guerra, B. C., and Leite, F. 2020. "Bridging the Gap between Engineering and Construction 3D Models in Support of Advanced Work Packaging." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *12*(3), 04520029. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000419

Gusfield, D. 1997 *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511574931

Hamdi, O., and Leite, F. 2014. "Conflicting side of building information modeling implementation in the construction industry." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 6(3): 03013004. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000137

Han, J., Pei, J., and Kamber, M. 2011. *Data mining: concepts and techniques*. Elsevier.

Hu, W. 2008, November. Information lifecycle modeling framework for construction project lifecycle management. In *2008 International Seminar on Future Information Technology and Management Engineering* (pp. 372-375). IEEE.

Ji, W., and AbouRizk, S. M. 2018. Simulation-based analytics for quality control decision support: A pipe welding case study. *Journal of Computing in Civil Engineering*, 32(3): 05018002.

Jung, Y., and Gibson, G. E. 1999. Planning for computer integrated construction. *Journal of computing in civil engineering*, 13(4): 217-225

Kaner, I., Sacks, R., Kassian, W., and Quitt, T. 2008. "Case studies of BIM adoption for precast concrete design by mid-sized structural engineering firms." *Journal of Information Technology in Construction (ITcon)*, 13(21), 303-323.

Kimmance, A. G. 2002. "An integrated product and process information modelling system for on-site construction" (Doctoral dissertation, © Andrew George Kimmance).

Koeleman, J., Ribeirinho, M. J., Rockhill, D., Sjödin, E., and Strube, G. 2019. "Decoding digital transformation in construction." *McKinsey and Company: Chicago, IL, USA*.

Leite, F., Cho, Y., Behzadan, A. H., Lee, S., Choe, S., Fang, Y., and Hwang, S. 2016. "Visualization, information modeling, and simulation: Grand challenges in the construction industry." *Journal of Computing in Civil Engineering*, 30(6): 04016035.

Li, Z., Wu, L., and AbouRizk, S. 2019. XiaomoLing/LongestCommonSubString: First Release of Longest Common SubString R Library (Version v1.0.0). Zenodo, http://doi.org/10.5281/zenodo.4057067

Lu, Y., Li, Y., Skibniewski, M., Wu, Z., Wang, R., and Le, Y. 2015. "Information and communication technology applications in architecture, engineering, and construction

organizations: A 15-year review." *Journal of Management in Engineering*, *31*(1), A4014010.

Manyika, J., Ramaswamy, S., Khanna, S., Sarrazin, H., Pinkus, G., Sethupathy, G., and Yaffe, A. 2015. Digital America: A tale of the haves and have-mores. *McKinsey Global Institute*, 1-120.

Martínez-Rojas, M., Marín, N., and Vila, M. A. 2016. "The role of information technologies to address data handling in construction project management." *Journal of Computing in Civil Engineering*, 30(4): 04015064.

McGee, J. V., Prusak, L., and Pyburn, P. J. 1993. *Managing information strategically: Increase your company's competitiveness and efficiency by using information as a strategic tool* (Vol. 1). John Wiley and Sons.

Mitchell, V. L. 2006. "Knowledge integration and information technology project performance." Mis Quarterly, 30(4), 919-939.

Nath D., Kurmi J., and Rawat V., 2018. "A Survey on Longest Common Subsequence." International journal for research in applied science and engineering technology 6(4), 4553-4557. https://doi.org/10.22214/ijraset.2018.4746

Nitithamyong, P., and Skibniewski, M. J. 2004. "Web-based construction project management systems: how to make them successful?". *Automation in Construction*, *13*(4), 491-506.

Ng, S. T., Xu, F. J., Yang, Y., and Lu, M. 2017. "A master data management solution to unlock the value of big infrastructure data for smart, sustainable and resilient city planning." *Procedia engineering*, *196*, 939-947.

Olatunji, O. A. 2016. "Constructing dispute scenarios in building information modeling." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 8(1): C4515001. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000165.

Pereira, E., Ali, M., Wu, L., and Abourizk, S. 2020. "Distributed simulation–based analytics approach for enhancing safety management systems in industrial construction." *Journal of Construction Engineering and Management*, 146(1): 04019091. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001732

Penttila, H. 2006. "Describing the changes in architectural information technology to understand design complexity and free-form architectural expression." *Journal of Information Technology in Construction (ITcon)*, 11(29), 395-408. https://www.itcon.org/2006/29

Preidel, C., Daum, S., and Borrmann, A. 2017. "Data retrieval from building information models based on visual programming." *Visualization in Engineering*, 5(1), 18.

Rezgui, Y., Boddy, S., Wetherill, M., and Cooper, G. 2011. "Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction?" *Computer-Aided Design*, *43*(5), 502-515.

Ripley B. and Lapsley M. 2020. *RODBC: ODBC Database Access*. R package version 1.3-17. Accessed May 12, 2021, https://CRAN.R-project.org/package=RODBC

Santos, R., Costa, A. A., and Grilo, A. 2017. "Bibliometric analysis and review of Building Information Modelling literature published between 2005 and 2015." *Automation in Construction*, 80, 118-136.

Sanvido, V. E., and Medeiros, D. J. 1990. "Applying computer-integrated manufacturing concepts to construction." *Journal of Construction Engineering and Management*, 116(2), 365-379.

Saraf, N., Langdon, C. S., and Gosain, S. 2007. "IS application capabilities and relational value in interfirm partnerships." *Information Systems Research*, 18(3), 320-339.

Sardroud, J. M. 2015. "Perceptions of automated data collection technology use in the construction industry." *Journal of Civil Engineering and Management*, *21*(1), 54-66.

Soibelman, L., and Kim, H. 2002. "Data preparation process for construction knowledge generation through knowledge discovery in databases." *Journal of Computing in Civil Engineering*, *16*(1), 39-48.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I., and Lin, K. Y. 2008. "Management and analysis of unstructured construction data types." *Advanced Engineering Informatics*, *22*(1), 15-27.

Solihin, W., and Eastman, C. 2015. "Classification of rules for automated BIM rule checking development." *Automation in Construction*: 53(1), 69-82. http://doi.org/10.1016/j.autcon.2015.03.003

Solihin, W., Dimyadi, J., Lee, Y. C., Eastman, C., and Amor, R. 2017. "The critical role of accessible data for BIM-based automated rule checking systems." *Proceedings of the Joint Conference on Computing in Construction,* (1), pp. 53-60. https://doi.org/10.24928/JC3-2017/0161

Shi, J. J., and Halpin, D. W. 2003. "Enterprise resource planning for construction business management." *Journal of Construction Engineering and Management*, 129(2), 214-221.

Tatari, O., Castro-Lacouture, D., and Skibniewski, M. J. 2007. "Current state of construction enterprise information systems: Survey research." *Constr. Innovation*, 74, 310–319.

Tatari, O., Castro-Lacouture, D., and Skibniewski, M. J. 2008. "Performance evaluation of construction enterprise resource planning systems." *Journal of Management in Engineering*, 24(4), 198-206.

Tatari, O., Ryoo, B. Y., and Skibniewski, M. J. 2004. "Modeling of ERP system solutions for the construction industry." In *Proc., 5th European Conf. on Product and Process Modeling in AEC Industry* (pp. 393-398). Istanbul, Turkey: Istanbul Technical Univ..

Teicholz, P., and Fischer, M. 1994. "Strategy for computer integrated construction technology." *Journal of Construction Engineering and Management*, 120(1), 117-131.

Thompson, G. I. 1996. "Need for an Enterprise Resource Management Measurement/Forecasting Infrastructure." In *The 1996 22nd International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems*, CMG. Part 1(of 2) (pp. 467-478).

Tinham, B. 1999. "Advancing on planning and scheduling?" *Manufacturing Computer Solutions*, 5(3), 24-5.

Umble, E. J., Haft, R. R., and Umble, M. M. 2003. "Enterprise resource planning: Implementation procedures and critical success factors." *European Journal of Operational Research*, 146(2), 241-257.

Viljamaa, E., and Peltomaa, I. 2014. "Intensified construction process control using information integration." *Automation in Construction*, *39*, 126-133.

Voordijk, H., Van Leuven, A., and Laan, A. 2003. "Enterprise resource planning in a large construction firm: implementation analysis." *Construction Management and Economics*, 21(5), 511-521.

Wickham, H. 2014. "Tidy data." *Journal of Statistical Software*, 59(10), 1-23.

Wickham, H., François, R., Henry L., and Müller K. 2020. dplyr: A Grammar of Data Manipulation. R package version0.8.4. https://CRAN.R-project.org/package=dplyr

Wu, L., and AbouRizk, S., 2020. XiaomoLing/Detect3DRelation: First Release of the Detect 3D Relation function (Version v1.0.0). Zenodo, http://doi.org/10.5281/zenodo.4058576

Wu, L., Li, Z., and AbouRizk, S. 2020. "Automation in Extraction and Sharing Information between BIM and Project Management Databases." *Proceedings of the International Conference on Construction and Real Estate Management (ICCREM)* 37-46, Stockholm, Sweden. https://doi.org/10.1061/9780784483237.005

Yuan, X., Chen, Y. W., Fan, H. B., He, W. H., and Ming, X. G. 2019, December. "Collaborative Construction Industry Integrated Management Service System Framework Based on Big Data." In *2019 International Conference on Industrial Engineering and Engineering Management,* (pp. 1521-1525). IEEE.

Zhang, Z., Lee, M. K., Huang, P., Zhang, L., and Huang, X. 2005. "A framework of ERP systems implementation success in China: An empirical study." *International Journal of Production Eco*

3. CHAPTER 3: BAYESIAN INFERENCE WITH MARKOV CHAIN MONTE CARLO-BASED NUMERICAL APPROACH FOR INPUT MODEL UPDATING

3.1. INTRODUCTION

As a tool, modeling has been widely used in engineering disciplines to design, analyze, communicate, test, and commission industrial, commercial, and residential facilities (AbouRizk et al. 2016). Simulation models are becoming increasingly used to support critical decision-making in construction engineering. Of the myriad of simulation techniques available (Akhavian and Behzadan 2013), discrete-event simulation is most often applied in industrial and infrastructure construction decision-making processes due to its ability to simulate resource interactions and operation logistics, especially for large and complex construction projects.

The success of a simulation model is highly dependent on accurately modeling the inputs, particularly in construction where a considerable number of inputs (each imbued with a wide variety of uncertainties) all relate to the underlying random process of various activities and tasks. The more accurate the model of the random input process, the more closely the simulation model mimics real-life behavior. To account for input variability, researchers have advocated for the modeling of inputs as probability distributions in a process known as stochastic or Monte Carlo simulation. Because of their ability to incorporate the randomness and various uncertainties inherent to construction activities, stochastic simulation models have been widely studied and used in the construction industry to enhance simulation-based decision-support systems.

Despite such advancements, the application of stochastic, discrete-event simulation models has traditionally been limited to the planning phase of construction. The industry continues to face notable challenges when it comes to adopting, upgrading, and using simulation models for decision support during the execution stage, as inputs (e.g., a given distribution from

historical data or experts' judgments) are often rigid, with no reliable or effective solution for fusing actual performance with the original input distribution to achieve real-time updating of the simulation model (Akhavian and Behzadan 2013). Because of these challenges, current simulation models have difficulty (1) reflecting real-time performance because of the use of static probability distributions and (2) fusing subjective judgements with objective observations, thus limiting the application of simulation-based decision-support systems during the execution-phase of a project. Although updating techniques, such as Bayesian statistics, have been proposed as a means of achieving real-time updating, many Bayesian-based methods require input data to have an analytical solution (i.e., conjugacy), limiting the application of these techniques in practice.

This study aims to address the limitation of real-time updating through the coupling of Bayesian inference with a Markov chain Monte Carlo-based numerical approximation approach, resulting in a universal input model updating method applicable to any univariate continuous probability distribution regardless of the conjugacy (i.e., a known parametric form of the posterior distribution). Demonstrated through its application on an illustrative case study, the proposed method was found capable of (1) fusing actual performance with expert judgment, (2) integrating actual performance with historical data, and (3) processing raw data by absorbing uncertainties and randomness. By enabling efficient, dynamic updating of the rigid inputs of a simulation model with new observations or subjective expert knowledge, the proposed method is expected to considerably improve the resilience, reliability, accuracy, and practicality of stochastic simulation models during the execution phase of construction.

3.2. LITERATURE REVIEW

3.2.1. Generalized Beta Family of Distributions

Following AbouRizk and Halpin's (1992) empirical study, which demonstrated the criticalness of using a flexible distribution (e.g., generalized beta distribution) to ensure the accuracy of the input modeling, the beta distribution has been extensively used for modeling inputs of the construction process over the last two decades. Among all of the flexible distributions, the generalized beta distribution with four parameters is one of the most widely recognized distributions for modeling construction processes (Chau 1995). Many researchers have successfully employed beta distributions to model a large number of construction management parameters including, but not limited to the following: activity durations (Lu and AbouRizk 2000; Lu 2003; Poshdar et al. 2018; Zayed and Halpin 2001), construction costs (Inyim et al. 2016; Sonmez 2005; Wang et al. 2002), and quality management indicators (Ji and AbouRizk 2017).

Due to its extensive usage in construction simulation modeling, the generalized beta family of distributions is presented here and is implemented in the case study. However, it is important to note that the proposed method is not limited to the beta distribution, and can be generalized to any other parametric probability distribution functions—a key contribution of the proposed method.

Mathematically, on an interval of $[L, U]$, a generalized beta distribution can be described as follows (AbouRizk et al. 1991; Ahsanullah 2017; Johnson et al. 1994):

$$if(y; a, b, L, U) = \frac{1}{B(a,b)} \cdot \frac{(y-L)^{a-1}\,(U-y)^{b-1}}{(U-L)^{a+b-1}}, \qquad if\ L \le Y \le U$$

$$f(y; a, b, L, U) = 0, otherwise \tag{1}$$

where $B(a, b)$ is the beta function. With the transformation matrix shown below, $f(y; a, b, L, U)$ can be standardized to $f(x: a, b)$ with an interval of [0,1].

$$X = \frac{Y - L}{U - L}, \qquad if\ L\ \leq Y \leq U \tag{2}$$

Standardized beta distribution:

$$f(x: a, b) = \frac{1}{B(a, b)} \cdot\ x^{a-1}(1 - x)^{b-1}, \quad if\ 0\ \leq X\ \leq 1$$

$$f(x; a, b) = 0, otherwise \tag{3}$$

Thus, the generalized beta distribution can be treated as a standardized beta distribution with shape parameters $\{a, b\}$ scaled to the $[L, U]$ interval.

3.2.2. Bayesian Inference

The simulation models of construction processes developed in the previously-mentioned studies have modeled their inputs based on either historical data or expert knowledge with fixed parameters as inputs and rigid assumptions. Construction processes, however, are highly dependent on the specific conditions that exist at the time they are performed, rendering them prone to deviation from expected baselines (Martinez 2009): what may have been anticipated and modeled in the planning stage of construction is often not what occurs during execution. The application of many construction simulation models proposed in the literature is, consequently, limited to the planning stages of construction.

Background

The Bayesian inference approach, developed by Thomas Bayes and Richard Price in 1763 (Bayes and Price 1763), has gained popularity in the 21st century due to its ability to incorporate multiple levels of randomness, integrate data originating from different sources, and reallocate credibility across the probability distribution of the value as new observations become available. Many researchers have since studied and demonstrated the practicality and benefits of implementing Bayesian techniques for updating underlying research interests (Brandley et al. 2015; Chung et al. 2004; Ji and AbouRizk 2017; Milo et al. 2015). While the aforementioned research was limited to conjugate priors, or a specific probability distribution, they have clearly demonstrated that a Bayesian approach can improve model accuracy, credibility, and reliability by systematically updating information of interest.

In contrast to Bayesian statistics, frequentist statistics suggests that the sampling process is "random," assuming that (1) the probability of each individual in the population being included in the sample is the same and (2) separate drawings are mutually independent (Neyman 1937). It is generally agreed, "all scientific data has some degree of 'noise' in their value" (Kruschke 2014). Indeed, the underlying random processes in construction are associated with various uncertainties and conditions; however, achieving the *pure* randomness suggested by frequentists in applied statistics is impossible. Techniques used for data analysis should therefore be capable of inferring the underlying trends despite noise.

Bayesian statistics tackles the same problem from a different perspective. It systematically updates information of interest as more observations become available. Consequently, Bayesian inference is both flexible and practical due to its ability to incorporate multiple levels of randomness and to combine information from various sources while absorbing all reasonable uncertainties in the inferential summaries (Gelman et al. 2013). Derived from Bayes' theorem, the basic components of Bayesian inference include the likelihood function,

prior distribution, and joint posterior distribution. If prior distribution(s) are denoted as $p(\Theta)$ for the parameter set $\Theta = \{\theta_1, \ldots, \theta_n\}$, the likelihood function wherein all variables are related in a full probability mode denoted as $p(y|\Theta)$ and given a set of the new observation(s) of our underlying interest, $y = \{y_1, \ldots, y_n\}$,, then the joint posterior distribution $p(\Theta \mid y)$ follows the numerical relation defined by Bayes' rule:

$$p(\Theta|y) = \frac{P(\Theta)p(y|\Theta)}{p(y)} \tag{4}$$

where $p(y) = \sum_{\Theta} p(\Theta)p(y|\Theta)$ for all possible values of Θ, or $p(y) = \int p(\Theta)p(y|\Theta)d\Theta$ for continuous Θ. Factor $p(y)$ is often called the marginal distribution of y or, more informatively, the prior predictive distribution (Gelman et al. 2013). Since it does not depend on Θ, and with fixed observation set y, it is a constant. Accordingly, the posterior distribution is proportional to the prior distribution multiplied by the likelihood function, denoted as:

$$p(\Theta|y) \propto p(\Theta)p(y|\Theta) \tag{5}$$

Likelihood Function

In non-statistical parlance, one could interchange "likelihood" for "probability." Within Bayesian data analysis, however, "probability" provides us the ability to predict unobserved data; "likelihood," on the other hand, contains the available information through observed data (Statisticat 2013). Thus, the likelihood function is:

$$p(y|\Theta) = \prod_{i=1}^{n} p(y_i|\Theta) \tag{6}$$

As a result, Bayesian inference obeys the likelihood principle, which states that the same likelihood function $p(y|\Theta)$ yields the same inference for parameter(s)-Θ for a given set of observations.

Prior Distributions

In Bayesian inference, a prior probability distribution (often referred to simply as a "prior") of a parameter is a distribution that expresses uncertainty about the parameter before new observations are considered (Statisticat 2013). By applying Bayes' rule, the posterior distribution is affected by the selection of the prior distribution through multiplication. Consequently, the proper selection of the prior probability distribution strongly affects the outcome of the posterior distribution. Commonly, prior distributions are categorized into informative priors and uninformative priors, although further categorization has been suggested (Statisticat 2013). Where uninformative priors express minimal, vague, diffuse beliefs about the parameters, informative priors express specific information. If a project management team believes a current project is similar to a previous project, for example, priors that are similar to the historical data of a similar project could be defined. The model could thus consider both the historical data and current project performance.

Posterior Predictive Distributions

When making inferences about an unknown observation, the posterior predictive distribution is an indispensable component within the Bayesian data analysis of most practical problems. Given the observation data $y = \{y_1, \ldots, y_n\}$, to-be-observed data $\tilde{y}$ can be predicted using:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\Theta)p(\Theta|y)d\Theta \tag{7}$$

The first factor, $p(\tilde{y}|\Theta)$, is the probability density function of $\tilde{y}$, given the fixed parameter(s)-Θ, which does not depend on observation y. To obtain the posterior predictive distribution $p(\tilde{y}|y)$, one must first sample parameter set Θ from the joint posterior distribution, and then simulate $\tilde{y}$ using $\tilde{y}^{i} \sim p(\tilde{y}|\Theta)$.

3.2.3. Application of Bayesian Inference for Real-Time Updating of Construction Models

Though Bayes' rule specifies the mathematical solution for the posterior distribution, exact analytical solutions rely on the possibility of computing the marginal probability. Historically, Bayesian inference techniques have been restricted to models with likelihood functions paired with corresponding formulas for prior distributions, known as conjugate priors (Kruschke 2014). Readers are referred to Jen and Hsiao (2018) for a detailed list of the most commonly used conjugate probability distribution functions.

Accordingly, it has been a longstanding challenge to generate simulation models that are capable of incorporating real-time updates during the execution phase (Akhavian and Behzadan 2013) to dynamically perform data-driven analytics and to provide critical decision-making support. Although research attempts have been made to use Bayesian techniques for real-time updating purposes in construction (Brandley et al. 2015; Chung et al. 2004; Ji and AbouRizk 2017; Milo et al. 2015), the methods and solutions proposed are limited to very specific cases. While Chung (2004) proposed using a conjugate prior (i.e., normal distribution) for the probability density function (i.e., normal distribution) to achieve real-time updating of input models of a long-term repetitive tunneling project, the joint posterior distribution was assumed for the posterior predictive distribution. While both have the same mean in this case, the standard deviation differs; this results in an unrealistically

small deviation. Later, Ji and AbouRizk (2017) carefully validated the Bayesian inference methodology on a binominal case for the quality control system of pipe fabrication. While a conjugate prior (i.e., beta distribution) for the Bernoulli likelihood function (i.e., binomial case) was presented, the study did not provide a universal solution for using the Bayesian inference methodology to update any univariate continuous input models regardless of conjugacy or likelihood function. Additionally, the need for fusing real-time performance with historical data to reflect the current project condition and integrating subjective expert opinion with objective observation remains unmet.

Performing Bayesian inference for realistic applications has often been limited to very specific cases, such as the aforementioned research, where a prior distribution conjugate to the likelihood function is specified to yield an analytically solvable posterior distribution. However, with the development of random sampling algorithms (such as Markov chain Monte Carlo (MCMC)) and faster computer hardware, a broader selection of priors and likelihood functions are available for conducting Bayesian inference. With the help of MCMC and powerful computer hardware, an accurate approximation of the Bayesian posterior distribution is achievable in the absence of an exact analytical solution.

3.2.4. Markov Chain Monte Carlo

In cases where an analytical mathematical solution does not exist (i.e., where conjugacy cannot be met), a numerical approximation of the target distribution has been found to be a reliable alternative (Ji and AbouRizk 2017). The most commonly used approximation approach involves mimicking the target distribution through the random sampling of a large number of data points. Notably, in cases where the parameter space is relatively small, other approaches that systematically cover the parameter space by exhaustively computing the marginal probability can also be applied (Kruschke 2014).

Here, the Markov chain Monte Carlo (MCMC) method is used to generate an accurate approximation of the Bayesian posterior distribution, thereby providing a universal, real-time updating solution that overcomes previous limitations regarding conjugacy. The term Markov chain Monte Carlo combines two processes, namely: (1) Monte Carlo simulation, which involves the random sampling of a large number of values (Kroese et al. 2014) and (2) the Markov chain, "a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event" (Oxford Dictionaries 2019). The representativeness, accuracy, and efficiency of the MCMC method are attributable to both the algorithmic design of the method and the large number of iterations performed (Ji and AbouRizk 2017).

Of the many sampling algorithms, the Metropolis algorithm—developed in the 1950s by Metropolis and colleagues (1953), and further refined in the 1970s (Moller and Waagepetersen, 2003)—has been widely used in physics, statistics, and applied sciences to approximate distributions (Robert and Casella 2011; Hitchcock 2003). Found capable of efficiently sampling single and double parameter problems, the Metropolis algorithm is well-suited for sampling distributions commonly used to model construction processes and is, consequently, used here.

Steps of the Metropolis methods are demonstrated as follows:

Step 1: Randomly generate a proposed leap, $\Delta\Theta \sim normal(\ \mu = 0,\ \sigma)$, and denote the proposed value of the parameter as $\Theta_{proposed}\ = \Theta_{current} + \Delta\Theta$

Step 2: Calculate the probability of moving to the proposed value:

$$p_{move} = \min\left(1, \frac{p(\Theta_{proposed}|y)}{p(\Theta_{current}|y)}\right),\ p(\Theta|y)\ \propto p(\Theta)p(y|\Theta) \tag{8}$$

Step 3: Accept the proposed parameter value if a random value sampled from a [0,1] uniform distribution is less than the p_{move}; otherwise, reject the proposed parameter value, and tally the current value again.

3.3. METHODOLOGY

This research proposes a method that couples Bayesian inference with a Markov chain Monte Carlo-based numerical approximation approach for updating univariate continuous probability distributions, regardless of the conjugacy. The proposed research method is illustrated in Figure 3-1. To provide an illustrative example of the methodology, a generalized beta distribution with four parameters is outlined. It is important to note, however, that the proposed method can be applied to any univariate continuous probability distribution.

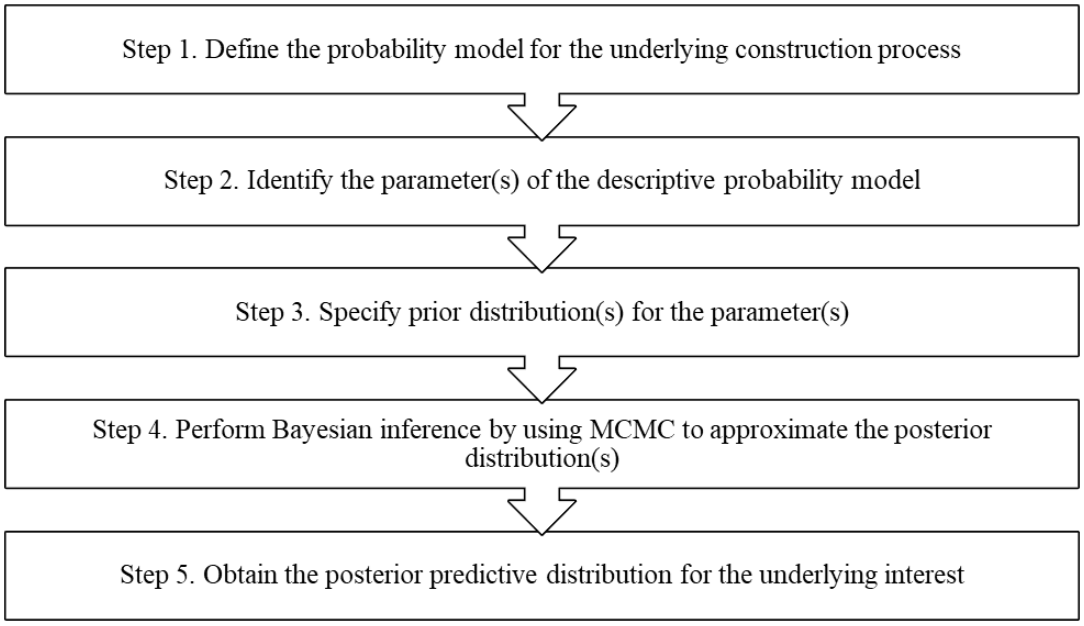


Figure 3-1 Proposed methodology

Step 1: Define the Probability Model

A descriptive probability model for all observable and unobservable quantities representing the underlying research interest (e.g., duration of a construction process, labor cost of activities, or productivity factor of a trade) is first defined. For illustrative purposes here, the underlying research interest is assumed to follow a generalized beta distribution $Y \sim Beta(y|a, b, L, U)$ with 4 parameters: a, b, L, and U.

Step 2: Identify and Understand the Parameters

The parameter(s) of the selected probability model (e.g., mean and standard deviation for a normal distribution, shape parameters for beta distribution) are then identified and understood. In the case of a generalized beta distribution, parameters L and U define the boundaries of the beta distribution. For example, if Y represents the duration of an activity, parameters L and U are the minimum and maximum durations of this activity recorded in historical data, respectively. In practice, the boundary parameters L and U are often well-established, with no further updates required; hence, they can be considered constants. In contrast, shape parameters a and b, which directly control the shape of the beta distribution, commonly differ between projects; they are, therefore, the focus of the research interest.

Step 3: Specify Prior Distribution for the Parameters

In the case of multiple parameters, the credible values of the parameters may depend on the values of other the parameters, leading to a hierarchical model. Methods for addressing this issue are beyond the scope of this research. Alternatively, the credible values of the parameter may be independent of each other. For independent parameters a and b for a generalized beta distribution, the joint prior distribution follows: $p(a, b) = p(a)p(b)$. Based on the specific situation, informative priors (e.g., normal distribution) or uninformative priors (e.g., uniform distribution) can be chosen for $p(a)$ and $p(b)$.

Step 4: Bayesian Inference with MCMC Method

As more data points are collected, a Bayesian inference is conducted using the MCMC-based numerical method to derive the posterior distribution for the parameter(s). Given $Y \sim Beta(y|a,b,L,U)$, the probability of collecting new observation(s) $y_1,\ldots,y_n$ follows the mathematical form described as:

$$p(y|a,b) = \frac{1}{B(a,b)} \cdot \frac{(y-L)^{a-1}\,(U-y)^{b-1}}{(U-L)^{a+b-1}}, if\ L\ \leq Y \leq U \tag{9}$$

Considering fixed set of observations, $y = \{y_1,\ldots,y_n\}$, $p(y|a,b)$ is the likelihood function of parameters a and b. Defined by Bayesian inference, the joint posterior distribution follows:

$$p(a,b|y) \propto p(a,b)p(y|a,b) = p(a)p(b)\prod_{i=1}^{n} p(y_i\,|a,b) \tag{10}$$

In both theory and practice, the log-likelihood is used instead of the likelihood on both the record-level and model-level. Thus:

$$\log[p(a,b|y)] \propto \log[p(a,b)p(y|a,b)] = \log[p(a)] + \log[p(b)] + \sum_{i=1}^{n} \log\,[p(y_i\,|a,b)] \tag{11}$$

To approximate the joint posterior distribution, the MCMC numerical method with the Metropolis sampling algorithm will be applied as follows:

a. The approximation simulation begins with a set of initial values of parameters (a_1, b_1);
b. At the beginning of each iteration, randomly generate $\Delta a \sim normal(\,\mu = 0,\ \sigma_1)$ and $\Delta b \sim normal(\,\mu = 0,\ \sigma_2)$. Thus, $a_{proposed} = a_i + \Delta a, b_{proposed} = b_i + \Delta b$.
c. Calculate the probability of moving to the proposed value:

$$p_{move}$$

$$= \min\left(1, \frac{p\left(a_{proposed}, b_{proposed}\,|y\right)}{p(a_i, b_i|y)}\right) \tag{12}$$

$$= \min\left(1, \frac{p\left(a_{proposed}\right)p(b_{proposed})\ \prod_{i=1}^{n} p(y_i\,|a_{proposed}, b_{proposed})}{p(a_i)p(b_i)\ \prod_{i=1}^{n} p(y_i\,|a_i, b_i)}\right)$$

d. Accept the proposed parameter values $a_{proposed}$ and $b_{proposed}$ if a random value sample from a [0,1] uniform distribution is less than p_{move}; otherwise, reject the proposed parameter values and return to Step 4.2.

Following the completion of a desired number of iterations (e.g., 100,000), a set of samples for shape parameters a and b is generated. A histogram of the data set provides the reasonable representation of the joint posterior distribution $p(a, b|y)$.

Step 5: Obtain the Posterior Predictive Distribution

Finally, the posterior predictive distribution—representing the probability distribution of the yet-to-be-recorded data given the observed data—is derived. In the case of a beta distribution, the posterior predictive distribution for a future observation $\tilde{y}$ given y can be written as:

$$p(\tilde{y}|y) = \iint p(\tilde{y}|a, b)p(a, b|y)dadb \tag{13}$$

To approximate the posterior predictive distribution, first sample a^i, b^i from the joint posterior distribution $p(a, b|y)$, then simulate $\tilde{y}^{\,i} \sim beta\left(a^i, b^i, L, U\right)$, where overtime, $\tilde{y}_1, \ldots, \tilde{y}_n$ becomes an independently and identically distributed sample from $p(\tilde{y}|y)$ (Gelman et al. 2014).

3.4. ILLUSTRATIVE CASE STUDY

3.4.1. Background

Since activity durations are one of the most studied and utilized inputs for simulation models of construction processes, a simplified simulation model of an earth-moving operation is used to demonstrate the feasibility and functionality of the proposed method. The simplified model captures a truck cycle that includes four major activities: loading, hauling, dumping, and return. The model simulates the delivery of 2,000 tons of dirt using five, 20-ton capacity trucks that are loaded by shovels, which are assumed to be an unlimited resource, as illustrated in Figure 3-2. Major activities and their durations are listed in Table 3-1. For the purposes of this case study, the duration of loading, dumping, and return are assumed to be constant, while hauling is assumed to follow a four-parameter generalized beta distribution fitted from experts' knowledge and historical observations.

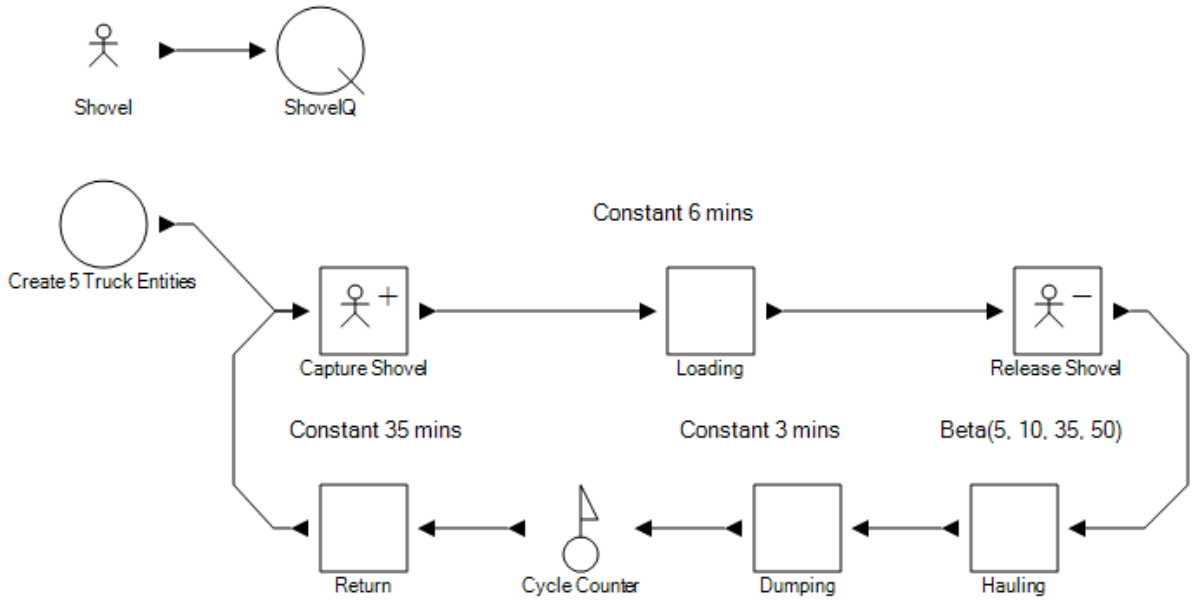


Figure 3-2 Simulation model of simplified earth-moving operation

Table 3-1 Original duration distributions of activities

Activity	Duration (min.)
Loading 20-Ton Truck	6
Truck Hauling	Beta (5, 10, 35, 50)
Truck Dumping	3
Truck Return	35

3.4.2. Bayesian Updating of Input Models

While many construction processes are repetitive, it is uncommon to collect hundreds of observations between critical reporting and decision-making periods. Thus, an input updating method capable of generating reliable results from a limited number of data points is critical to be functional in practice. To demonstrate the ability of the proposed method to perform appropriately under such conditions, only 20 new observations were generated for each of the five reporting cycles. Overall, 100 new sample observations (Table A-1) were randomly generated using the generalized beta distribution, Beta (5, 10, 35, 50). The proposed method was applied after 20 new observations were collected, and the accuracy of the proposed method was examined by comparing the input models derived using the proposed methodology (PM) with models that were directly fitted from the cumulative observations (CO), as well as the underlying distribution (UD).

3.4.3. Results

Shape parameters a and b that were obtained through direct fitting of the cumulative sampled observation actuals (i.e., Cycle 1, 20 samples were used for fitting; Cycle 2, 40 samples were used for fitting; etc.) are listed on columns "Fitted on CO" and "Difference (% True) on CO" in Table 3-2 and Table 3-3. Expectedly, the similarity of a and b to the

underlying distribution increased as the number of data points accumulated. Notably, drastic fluctuations between cycles were observed, with the results of certain cycles being similar to the underlying distribution and others deviating considerably.

Table 3-2 Shape parameters *a* fitted from cumulative observations (CO) V.S. proposed method (PM)

Cycle	True Value	Fitted on CO	Difference (% True) on CO	Fit using PM	Difference (% True) using PM
1	5	8.7314	74.63	4.6963	6.07
2	5	5.1539	3.08	4.7936	4.13
3	5	4.3740	12.52	4.6823	6.35
4	5	4.0576	18.85	4.5827	8.35
5	5	4.2786	14.43	4.6661	6.68
Average		5.3191	24.70	4.6842	6.32

Table 3-3 Shape parameters *b* obtained using cumulative observations (CO) V.S. proposed method (PM)

Cycle	True Value	Fitted on CO	Difference (% True) on CO	Fit using PM	Difference (% True) using PM
1	10	21.1659	111.66	10.3876	3.88
2	10	10.8684	8.68	9.9898	0.10
3	10	9.3608	6.39	9.9063	0.94
4	10	8.9133	10.87	9.9400	0.60
5	10	9.0413	9.59	9.8226	1.77
Average		11.8699	29.44	10.0093	1.46

While the performance of this project is assumed to be similar to historical projects that follow the generalized beta distribution, Beta (5, 10, 35, 50), the project is characterized by certain unique features and uncertainties. Accordingly, a normal distribution, Normal (5, 0.5), was defined as the prior for shape parameter *a*, and a normal distribution, Normal (10, 1), as the

prior for shape parameter b. This set of informative priors was chosen as a means of credibly considering historical data and expert opinion regarding uncertainty, where the mean value of each parameter was set at the most probable value based on the historical project data with the standard deviation representing 10% of the mean value to account for uncertainty. The posterior distribution of shape parameters a and b that was generated using the proposed method are listed on columns "Fit using PM" and "Difference (% True) using PM" in Table 3-2 and Table 3-3.

The proposed method demonstrates considerable reliability between cycles, and accuracy when compared to the underlying distribution—especially given the small set of observations. The average percentage differences between the mean value of the posterior distribution and the true value for shape parameters a and b were 6.32% and 1.46%, compared to the direct fitting on CO method, with 24.70% and 29.44%, respectively.

The histogram and trace plot of MCMC results for shape parameters a and b using PM in Cycle 1, together with the true values of the parameters, and parameters obtained through CO, are illustrated in Figure 3-3. Histograms and trace plots for Cycles 2 through 5 are illustrated in Figure A-1 through Figure A-4, respectively.

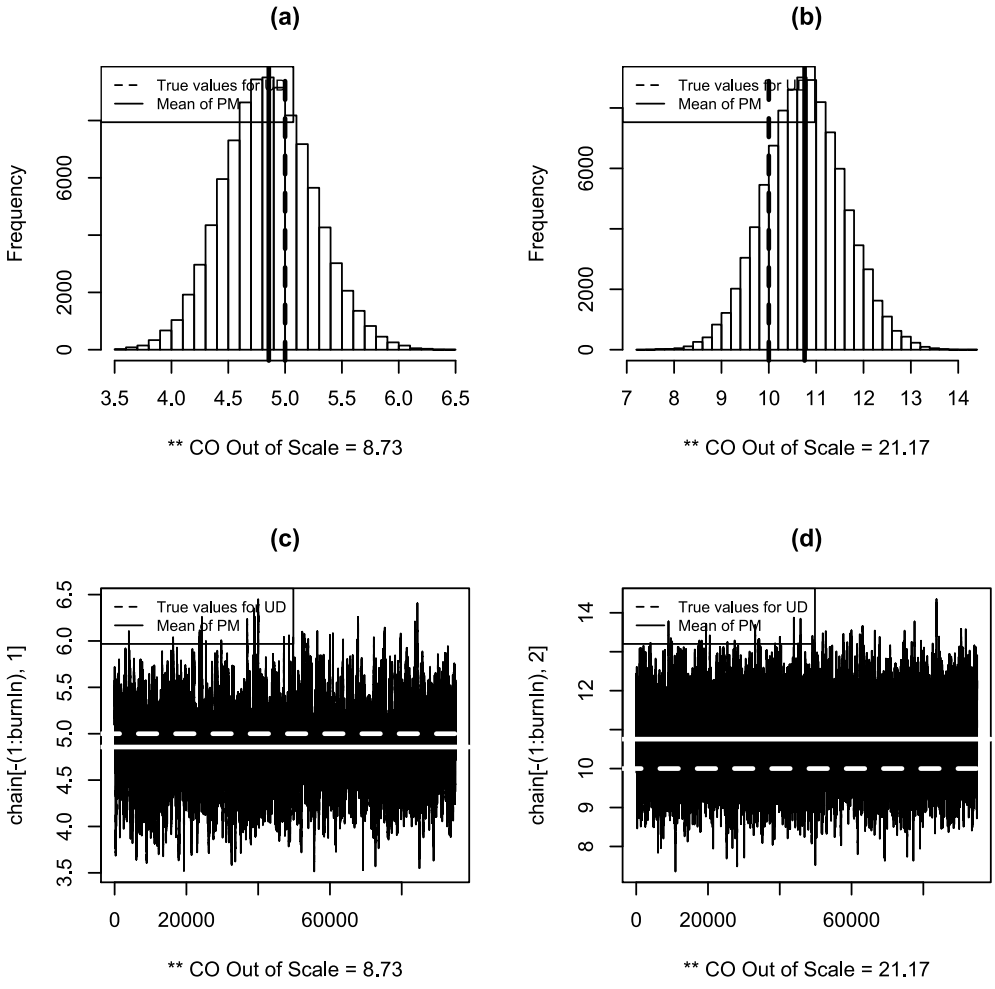


Figure 3-3 Posterior histogram ((a) and (b)) and trace plot ((c) and (d)) of parameters *a* ((a) and (c)) and *b* ((b) and (d)) for Cycle 1

The results demonstrate that (1) the mean of the MCMC posterior samples for *a* and *b* (*solid line*) were more similar to the true parameter values (*dash line*) when compared to the directly fitted from CO values (not represented in Figure 3-3; represented in Figure A-1 through Figure A-4) in all five cycles and (2) the direct fitting from CO method was associated with much larger fluctuations between cycles compared to the Bayesian inference (PM) method. Indeed, the parameter values fitted from CO were not within the presented scale for Figure 3-3. In this instance, they are not shown.

The histograms of the samples of the posterior predictive distribution in Cycle 1, together with the three input model distributions, are illustrated in Figure 3-4. Histograms for Cycles 2 through 5 are illustrated in Figure A-5 through Figure A-8. respectively. Similar to the results of shape parameters a and b, the posterior predictive distribution was consistently closer to the underlying true distribution than the input distribution fitted directly from cumulative observations. The impact of the input modeling methods on project forecasting was also examined. The project was simulated using input models either (1) directly fitted from the cumulative observations (CO), (2) derived using the proposed method (PM), or (3) the underlying distribution (UD). The forecasted project duration was determined for 1,500 runs; the results of the analysis are illustrated in Figure 3-5.

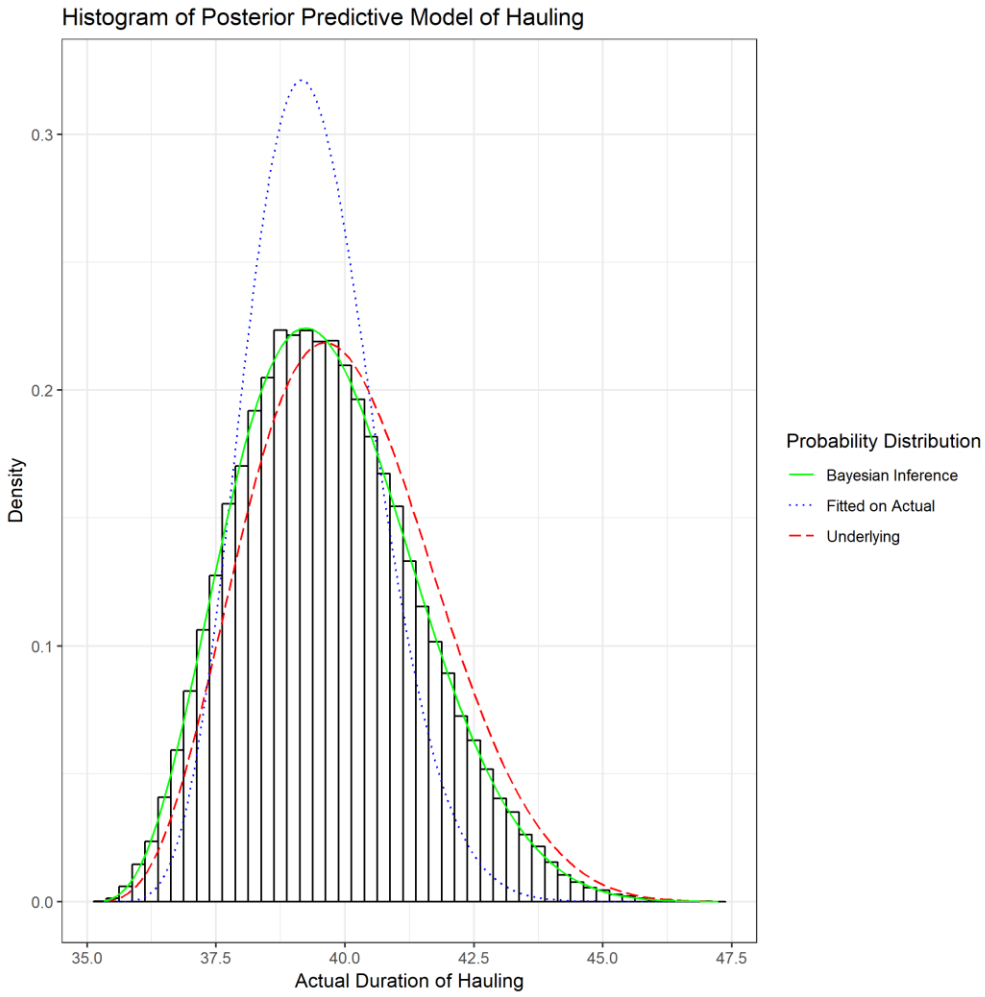


Figure 3-4 Histogram of posterior predictive hauling model for Cycle 1

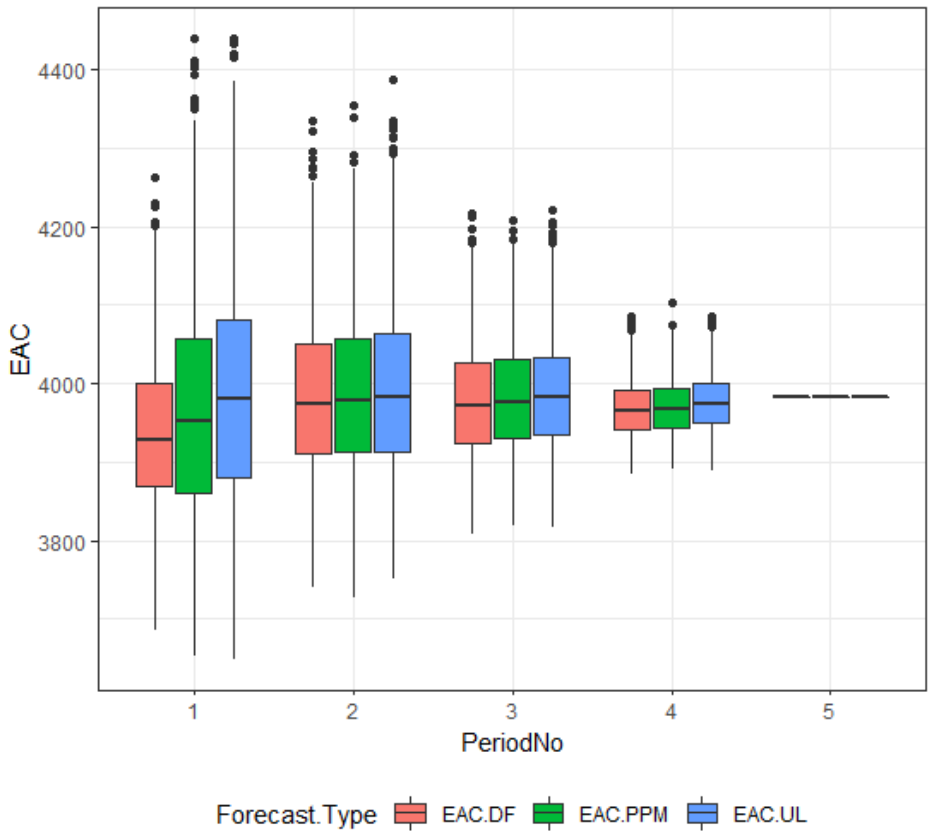


Figure 3-5 Boxplot of simulation results obtained using input models directly fitted from the cumulative observations (*CO*), derived using the proposed method (*PM*), or derived using the underlying distribution (*UD*)

Similar to the results obtained regarding the shape parameters and distributions, duration forecasts derived using the proposed input updating method were closer to the true underlying duration of the project for all five cycles when compared to the CO method. Moreover, during the first and second forecasting periods where the number of new observations was limited, the proposed method was found to more closely mimic the true underlying pattern and to more effectively incorporate various uncertainties (i.e., larger deviation window) than the direct fitting on CO method. Indeed, the narrower deviation window of the CO method may result in an over-optimistic forecast, as observed in Figure 3-5.

3.4.4. Sensitivity Analysis

To test the robustness of the proposed methodology, a sensitivity analysis designed to introduce a certain level of noise into the observation data to mimic the raw data collected from a real project site was performed. One of the most common causes of fluctuations in productivity in construction projects is the learning effect, which is known to result in significant forecasting challenges in the early stages of a project. To mimic the noise of decreased productivity resulting from the learning effect, the simulated data points from the first 3 cycles (i.e., 60 random samples) of actual hauling duration were generated using 10% of the uniform distribution, Uniform (45, 50), and 90% of generalized beta distribution, Beta (5, 10, 35, 50), placing a higher probability of sampling a lower productivity. Assuming that after 3 cycles the project had achieved optimum productivity, the simulated data points for Cycles 4 and 5 (i.e., the remaining 40 random samples) were generated using the generalized beta distribution, Beta (5, 10, 35, 50). The 100 random samples that were generated using this approach are listed in Table A-2. The shape parameters a and b fitted directly using cumulative observation samples are detailed in Table 3-4 and Table 3-5 from columns "Fitted on CO" to "Difference (% True) on CO".

Similar to the base case study scenario, the similarity of a and b to the true values from the underlying distribution, Beta (5, 10, 35, 50), increased as the number of data points accumulated. With the introduction of noise, the direct fitting using CO method took longer to approach the underlying distribution, demonstrating that this approach is sensitive to the noise in the data set. While the differences in parameters a and b from the true values settled to around 10% to 20% after a few cycles in the base scenario, (column "Difference (% True) on CO" in Table 3-2 and Table 3-3), the addition of noise resulted in a difference of around 20%

for both parameters for all five cycles, (column "Difference (% True) on CO" in Table 3-4 and Table 3-5).

Table 3-4 Shape parameters *a* fitted from cumulative observations (CO) V.S. proposed method (PM)

Cycle	True Value	Fitted on CO	Difference (% True) on CO	Fit using PM	Difference (% True) using PM
1	5	7.0902	41.80	5.1683	3.37
2	5	4.2651	14.70	5.0146	0.29
3	5	3.4740	30.52	4.8798	2.40
4	5	3.7698	24.60	4.9439	1.12
5	5	4.3418	13.16	4.9255	1.49
Averag e		4.5882	24.96	4.9864	1.74

Table 3-5 Shape parameters *b* obtained using cumulative observations (CO) V.S. proposed method (PM)

Cycle	True Value	Fitted on CO	Difference (% True) on CO	Fit using PM	Difference (% True) using PM
1	10	11.5305	15.31	9.7324	2.68
2	10	7.7220	22.78	9.6055	3.94
3	10	6.8282	31.72	9.7348	2.65
4	10	7.1323	28.68	9.5155	4.85
5	10	8.2430	17.57	9.5851	4.15
Average		8.2912	23.21	9.6347	3.65

Taking into consideration the learning effect and the uncertainties associated with recorded data, the project was anticipated to follow the generalized beta distribution, Beta (5, 10, 35, 50). Again, informative priors were chosen with normal distributions, Normal (5, 0.25) and Normal (10, 0.5), as priors for shape parameters a and b, respectively. Since posterior distribution is influenced by both new observations and the prior distributions, a proper

selection of the priors can affect the posterior given the same set of observations. If a set of uninformative priors is chosen, the posterior will show no influence from the priors but let the data speak for itself. To express firm belief in the subjective judgment of experts, the productivity fluctuation is caused by the learning effect, and the expected future observation will follow Beta (5, 10, 35, 50). The standard deviation was set as 5% of the value of the mean for both priors of parameters a and b. Corresponding posterior shape parameters are listed in Table 3-4 and Table 3-5 from columns "Fit using PM" to "Difference (% True) using PM".

As with the base scenario, the proposed method generated results that were more accurate and representative of the underlying probability distribution compared to the direct fitting using CO method. The average percentage difference between the mean value of posterior distribution and the true value for shape parameters a and b were 1.74% and 3.65%, compared to 24.96% and 23.21% fit directly from CO, respectively. The proposed method was also found to be comparatively insensitive to noise, with the average percentage difference similar for both the base scenario (column "Difference (% True) using PM" in Table 3-2 and Table 3-3) and following the addition of noise (column "Difference (% True) using PM" in Table 3-4 and Table 3-5). To conclude, the proposed method demonstrated (1) robustness when the noise was introduced and (2) desired representativeness and accuracy of both subjective opinion and objective observations.

The histogram and trace plot of MCMC results for shape parameters a and b using proposed method, together with the true values of the parameters, and parameters obtained using the other aforementioned method, are illustrated in Figure A-9 through Figure A-13. The histograms of the samples of the posterior predictive distribution, together with the three input model distributions, for Cycles 1 through 5 are illustrated in Figure A-14 through

Figure A-18. A comparison of the simulation results of project estimate at completion for each of the three input modeling methods is illustrated in Figure 3-6.

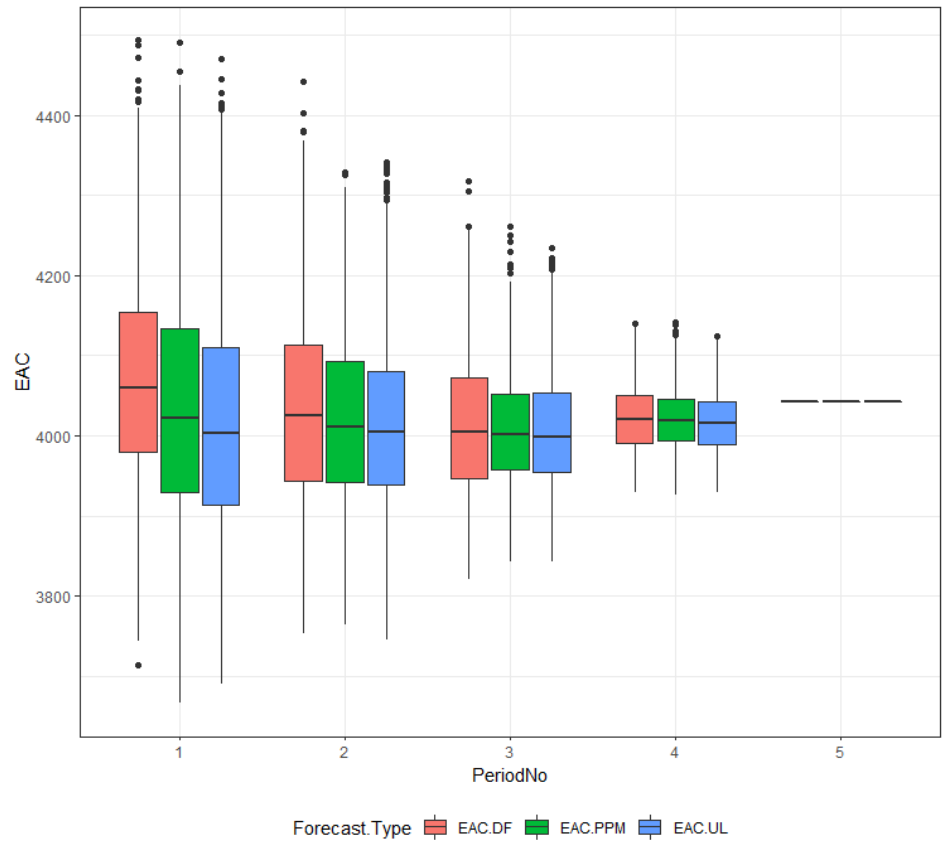


Figure 3-6 Boxplot of simulation results obtained using input models directly fitted from the cumulative observations (*CO*), derived using the proposed method (*PM*), or derived using the underlying distribution (*UD*)

As similarly, observed in the base scenario (Figure 3-5), the simulation results obtained using the proposed input updating method were associated with less fluctuation in the presence of noise and generated more reliable duration forecasts compared to the direct fitting on cumulative observation method. This was particularly evident during Cycles 1 and 2, where the robustness of the proposed method and its ability to deal with limited and noisy data were most apparent.

3.4.5. Potential Applications

The implementation of the proposed methodology facilitates the dynamic real-time integration of data into the simulation models, thus enhancing the original model's accuracy and predictability. The traditional DES benefits from the real-time auto-calibration of the input models by effectively assisting the decision-making process throughout both the project planning and execution phases of construction. This occurs in alignment with the dynamic data driven application system's philosophy (DDDAS) (Darema 2004), which has also been referred to as simulation-based analytics (AbouRizk 2018; Ji and AbouRizk 2018b), and dynamic data-driven simulation (Ji and AbouRizk 2018a).

Potential realistic applications in construction engineering and management fields include but are not limited to the following: production planning, earned value management, cost forecast, and risk management. Specifically, collected real-time performance data (such as production rate, productivity factor, actual cost, and so on) will be processed with the proposed methodology. The auto-calibrated input models will then be utilized in simulation-based decision support systems to reflect the dynamic project performance, deriving more accurate and meaningful decision-support output for practitioners. The proposed methodology can benefit any DDDAS, simulation-based analytics, or dynamic data-driven simulation developed for various engineering and applied science fields. For example, this method can be used to effectively process live sensor-generated data for real-time severe weather prediction, hazardous contamination production, traffic flow simulation, and so on.

3.5. CONCLUSIONS

Bayesian inference has been successfully implemented across many scientific and engineering disciplines to address the needs of multiple specific practical problems. However, many of the implementation methods, particularly in the area of construction engineering and management, are not generalizable due to their dependency on the availability of conjugate priors. Accordingly, many decision-support systems used in the construction industry remain unable to appropriately incorporate real-time information as it is generated.

This chapter proposes a universal, Bayesian inference-based method for systematically updating any given univariate continuous probability distribution input model of simulations as new observations become available, and implements an MCMC-based numerical approximation approach to provide solutions regardless of conjugacy. An illustrative case study is used to demonstrate the generalizability, feasibility, and functionality of the proposed Bayesian inference with MCMC-based numerical method for updating simulation input models. The proposed method has been found capable of (1) effectively and efficiently updating input models as new observations become available, (2) accurately approximating the underlying probability distribution, (3) reliably fusing information from diverse sources, including subjective judgment and objective observations, (4) exhibiting robustness and resilience in situations where data were noisy and imbued with uncertainties, and (5) being generalized and applied to any given univariate continuous probability distribution. By applying the proposed method, input models of stochastic simulations can be effectively and efficiently updated in real time throughout the execution of a construction project.

The contributions of this research should be considered in light of the several limitations. Due to the nature of the illustrative case study where the random observations were

generated based on a known underlying distribution, the fit of the model was not evaluated. In practice where the underlying distribution is unknown, however, assessing the fit of the model to the data and to the subjective knowledge of experts after obtaining the posterior predictive distribution is essential (Gelman et al., 2014). Additionally, the selection of the prior distribution is a complex problem that requires consideration of historical data, professional experience, and regard for current project conditions. Proper prior distribution selection is of the utmost importance for ensuring the accuracy of the posterior distribution. Finally, while the proposed method provides a philosophical approach for integrating information from various sources, incorporating multiple levels of uncertainty and randomness, and consistently providing accurate, reliable results, the method itself does not represent a complete, decision-support system.

Laying the foundation for further dynamic, data-driven, simulation-based, and analytics-focused research in construction, future work building upon the proposed methodology is expected to result in a new generation of quantitatively-driven, analytically-based decision-support systems capable of providing real-time analytics, fusing various information sources, and incorporating randomness to enhance the efficiency and automation in construction management.

3.6. ACKNOWLEDGEMENTS

This research is funded by an NSERC Collaborative Research and Development Grant (CRDPJ 492657). The authors would like to thank Stephen Hague for sharing his knowledge and expertise of Bayesian inference.

3.7. SUPPLEMENTAL DATA

Table A-1 and Table A-2 and Figure A-1 to Figure A-18 are provided in Appendix A.

3.8. REFERENCES

AbouRizk, S.M. (2018). "Simulation-based analytics: Advancing decision support in construction." Responsible Design and Delivery of the Constructed Project. Presented at the Proceedings of the Second European and Mediterranean Structural Engineering and Construction Conference. Beirut, Lebanon, July 23-28, 2018. ISEC Press: Fargo, ND.

AbouRizk, S. M., Hague, S. A., and Ekyalimpa, R. (2016). Construction simulation: An introduction using Simphony. University of Alberta, Edmonton, Canada.

AbouRizk, S. M., and Halpin, D. W. (1992). "Statistical properties of construction duration data." Journal of Construction Engineering and Management, 118(3), 525-544.

AbouRizk, S. M., Halpin, D. W., and Wilson, J. R. (1991). "Visual interactive fitting of beta distributions." *Journal of Construction Engineering and Management*, 117(4), 589-605.

Ahsanullah, M. (2017). *Characterizations of univariate continuous distributions*. Atlantis Press, Paris, France.

Akhavian, R., and Behzadan, A. H. (2013). "Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining." *Journal of Construction Engineering and Management*, 139(11), 04013021.

Bayes, T., and Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. *Philosophical Transactions*, 53, 370-418.

Brandley, R. L., Bergman, J. J., Noble, J. S., and McGarvey, R. G. (2015). "Evaluating a Bayesian approach to demand forecasting with simulation." *Proceedings of the 2015 Winter Simulation Conference*, IEEE, Piscataway, NJ, 1868-1879.

Chau, K. W. (1995). "Monte Carlo simulation of construction costs using subjective data." *Construction Management and Economics*, 13(5), 369-383.

Chung, T. H., Mohamed, Y., and AbouRizk, S. (2004). "Simulation input updating using Bayesian techniques." In *Proceedings of the 2004 Winter Simulation Conference*, IEEE, Piscataway, NJ, 1238-1243.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*, CRC Press, Boca Raton, FL.

Hitchcock, D. B. (2003). "A history of the Metropolis–Hastings algorithm." *The American Statistician*, 57(4), 254-257.

Inyim, P., Zhu, Y., and Orabi, W. (2016). "Analysis of time, cost, and environmental impact relationships at the building-material level." *Journal of Management in Engineering,* 32(4), 04016005.

Jen, H., and Hsiao, C. (2018). "Using Bayesian inference modeling in estimating important production parameters used in the simulation-based production planning." *Proceedings of IEEE International Conference on Applied System Innovation 2018*, IEEE, Piscataway, NJ, 1038-1041.

Ji, W., and AbouRizk, S. M. (2017). "Credible interval estimation for fraction nonconforming: Analytical and numerical solutions." *Automation in Construction*, 83, 56-67.

Ji, W., and AbouRizk, S. M. (2018a). "Data-Driven Simulation Model for Quality-Induced Rework Cost Estimation and Control Using Absorbing Markov Chains." *Journal of Construction Engineering and Management*, *144*(8), 04018078.

Ji, W., and AbouRizk, S. M. (2018b). "Simulation-based analytics for quality control decision support: A pipe welding case study." *J. Comput. Civ. Eng*, *32*(3), 05018002.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous univariate distributions,* 2nd ed. John Wiley and Sons, New York, NY.

Kroese, D. P., Brereton, T., Taimre, T., and Botev, Z. I. (2014). "Why the Monte Carlo method is so important today." *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6), 386-392.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, London, UK.

Lu, M. (2003). "Simplified discrete-event simulation approach for construction simulation." *Journal of Construction Engineering and Management*, 129(5), 537-546.

Lu, M., and AbouRizk, S. M. (2000). "Simplified CPM/PERT simulation model." *Journal of Construction Engineering and Management*, 126(3), 219-226.

Martinez, J. C. (2009). "Methodology for conducting discrete-event simulation studies in construction engineering and management." *Journal of Construction Engineering and Management*, 136(1), 3-16.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics*, 21(6), 1087-1092.

Milo, M. W., Roan, M., and Harris, B. (2015). "A new statistical approach to automated quality control in manufacturing processes." *Journal of Manufacturing Systems*, 36, 159-167.

Moller, J., and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press, Boca Raton.

Neyman, J. (1937). "X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Phil. Trans. R. Soc. Lond. A*, 236(767), 333-380.

Oxford Dictionaries. (2019). "Definition of Markov chain in US English." < https://en.oxforddictionaries.com/definition/us/markov_chain> (January 21, 2019).

Poshdar, M., González, V. A., Raftery, G. M., Orozco, F., and Cabrera-Guerrero, G. G. (2018). "A multi-objective probabilistic-based method to determine optimum allocation of time buffer in construction schedules." *Automation in Construction*, 92, 46-58.

Robert, C., and Casella, G. (2011). "A short history of MCMC: Subjective recollections from incomplete data." *Statistical Science,* 26(1), 102-115.

Sonmez, R. (2005). "Review of conceptual cost modeling techniques." *AACE International Transactions*, ES71.

Statisticat. (2013). "Bayesian inference." <chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/BayesianInference.pdf> (January 31, 2019).

Wang, L., Shen, W., Xie, H., Neelamkavil, J., and Pardasani, A. (2002). "Collaborative conceptual design—state of the art and future trends." *Computer-Aided Design*, 34(13), 981-996.

Zayed, T. M., and Halpin, D. (2001). "Simulation of concrete batch plant production." *Journal of Construction Engineering and Management*, 127(2), 132-141.

4. CHAPTER 4: A NUMERICAL-BASED APPROACH FOR UPDATING SIMULATION INPUT IN REAL-TIME

4.1. INTRODUCTION

Construction processes are subject to a wide range of uncertainties and randomness, rendering them stochastic rather than deterministic. To account for these uncertainties, stochastic simulation models have been adopted to model construction processes and related systems using Monte Carlo (MC) methods (e.g. simulation of scheduling networks, production processes, risk management, and range estimates) (Altaf et al. 2018, Song and Eldin 2012, Liu et al. 2015). Given that the simulation model captures internal construction processes and their corresponding logics, the implementation of stochastic models aggregates these uncertainties and evaluate choices that will enhance the decision-making process (Hubbard 2009, Rao et al. 2008).

Construction projects are unique, dynamic and complex: Despite the internally identical or similar mechanisms and logics (Behzadan et al. 2015), the seemly repetitive operations in construction can drastically differ due to uncertainties and external factors, such as location, weather, labor skills, morale, and utilization of technology (Seresht and Robinson Fayek 2018). Despite the rapid development of information technologies, most of the important project data—capturing these external factors—are stored in unstructured text documents or exchanged verbally among the involved professionals, making them difficult to use (Martínez-Rojas et al. 2016, Caldas et al. 2002, Al Qady and Kandil 2013). Thus, simply relying on a single data source, such as real-time observations, might lead to unrealistic predictions.

Traditionally construction simulation models are built on rigid structures and static assumptions (Akhavian 2015, Hammad and Zhang 2011). The resulting models, unfortunately, can rarely be implemented for a different project, as they are not capable of

representing deviations caused by external factors. In fact, the simulation model is often used only in the planning stage—reportedly around 60% in Leite et al.'s (2016) study—as these models are not equipped to process the real-time observational data, other project information, and expert knowledge as it becomes available. The discrepancies from the simulation results and the actual performance expand, failing to provide reliable decision support over the entire project life span.

Developing a simulation model for a construction process often starts with investigating business practices/processes/systems through the historical project(s) (AbouRizk et al. 2016a). Domain experts abstract the construction processes into a conceptual model (Abdelmegid et al. 2017, Chwif et al. 2013), then develop a simulation model to represent the required construction entities, logics, and state variables (AbouRizk et al. 2016b, AbouRizk 2010). After testing and validation with the existing project data, the resulting model will be implemented in a future project (Sargent 2010). The planning phase of a new construction project lacks actual data. Thus, project conditions are often judged based on experts' experience, and the inputs for the simulation model are often static parametric distributions that are fit to historical project data (Akhavian and Behzadan 2013). Upon project commencement, more observational data are generated and collected such as productivity, cost, weather, safety, quality, etc. Additionally, more experts (e.g. superintendents, engineers, managers, technicians, etc.) are involved, further increasing subjective information on current project conditions (e.g. congestion level, moral, crew skill level, attrition rate) and knowledge in predicting future conditions. Consequently, the construction project needs a reliable method to fuse information gathered from various origins to properly reflect current and future project conditions and provide reliable decision-support.

The barriers and grand challenges of modeling the construction process have been studied and documented by various scholars (Leite et al. 2016, Martínez-Rojas et al. 2016, Abdelmegid et al. 2020). This study focuses on addressing the following two aspects of these challenges: 1) properly reflecting the dynamics of the construction ecosystem; 2) effectively processing real-time information generated from various origins—both observational and subjective—to update the existing model thus better reflecting the current project conditions.

Regardless of the simulation techniques, the input modeling process is identical, and it is one of the most important factors to the success of the stochastic model (Gong and Caldas 2010). The appropriate selection of probability distributions ensures the representation of the underlying random input process and, ultimately, generates a meaningful result for the simulation model (as illustrated in Figure 4-1). If the inputs of a simulation model can properly fuse and represent the actual project data ranging from real-time observations, expert knowledge, and other information gathered from different sources, then the credibility, resilience, and application of the simulation models will be enhanced throughout the project. Through the better representation of the input models, to a certain degree, the simulation model captures and reflects the external mechanism in a construction system.

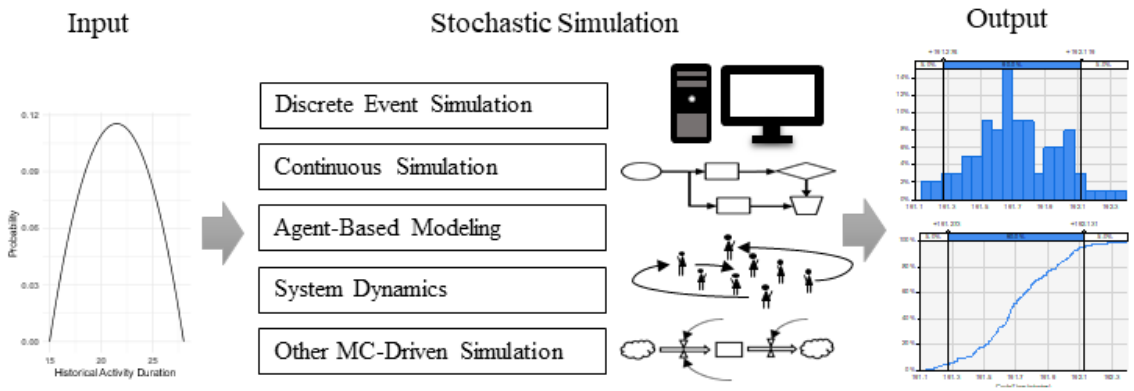


Figure 4-1 Conceptual Model of Stochastic Simulation Model

Through developing a universal methodology in updating simulation input models, this research aims to capture and represent the dynamic observational data and subjective information collected throughout the project. Specifically, this research proposes a dynamic input model updating method that couples a Metropolis-Hastings based Markov chain Monte Carlo (MCMC) process with a weighted geometric average (GA). This novel approach provides direct solutions for fusing information generated or collected through various means (e.g. historical data, real-time performance, and expert knowledge), resulting in a universal solution for updating univariate, continuous, parametric probability input models, thus, improving the robustness and application of simulation models.

4.2. MOTIVATION

In this section, a simplified discrete event simulation model—takes in one input and provides the project total duration as a distribution—is used to demonstrate the four extremely common project monitoring and controlling scenarios. The configuration of the input model (i.e. the probabilistic distribution) is elaborated through the following four scenarios. Nevertheless, it is important to note that the proposed method is not limited to discrete event simulation nor forecast function, but any MC-driven model for any construction-related process as depicted in Figure 4-1.

Scenario 1 – Planning stage with historical data

Before the start of the project, the input model is often assumed to resemble historical project data, for instance, a four-parameter generalized beta distribution, such as beta (2, 2, 15, 28).

Scenario 2 – Construction stage with real-time observations

Actual observations are generated and collected throughout project execution, and a new project's conditions—as reflected in the observations—are unlikely to be identical to any historical projects. To reflect the new external factors, the new data necessitate updating existing input models with real-time observations from the field. Especially for highly-repetitive construction activities, where the next observation(s) likely share the same external factors as current ones, such as tunneling activity with tunnel boring machine (e.g. the penetrating rate is mostly influenced by soil types) and welding at fabrication facility (highly-controlled environment). Applying Bayesian inference has proven effective for incorporating real-time data and fine-tuning the input model in such cases (Chung et al. 2004; Ji and AbouRizk 2017, Wu et al. 2020), and an updated input beta (5, 3, 15, 28) can be achieved.

Scenario 3 – Construction stage with subjective opinions

Various external factors (such as congestion level, safety, crew morale, and sustainability) can cause performance deviations; yet reasoning about performance deviations is still an emerging science (Skibniewski and Golparvar-Fard 2016). These types of project information, especially the subjective analysis of external factors, are commonly shared among involved parties verbally (Martínez-Rojas et al. 2016). Simply relying on project actual data could lead to unrealistic decision-making matrices due to ignoring foreseeable events and factors. Unlike the shop weld that takes place in a controlled environment, the performance of the field weld—extremely common yet problematic and inefficient construction activity in industrial projects—is influenced by various factors (CII 2015), such as weather (mostly precipitation), location (height and accessibility), and congestion level (whether sharing workspace with other crews, or tight workspace due to obstruction from other structures).

For instance, the current welding performance (actual data) was measured by performing activities at ground level, without the severe obstruction of other structure. With the project progressing, future welding activities are at various elevations, with potential shared workspace and obstruction from structures. Field supervisors expect the performance to decline, and three superintendents put in their three point estimates. As a result, the project manager often forecasts the project performance based on experts' judgments of project future conditions. With axiom-based aggregation approaches (e.g. weighted average), the triangular distribution (15, 28, 19) could provide an aggregated input model based on the experts' opinions.

Scenario 4 – Construction stage with various information sources

Methodologies have been developed and researched for forecasting a single type of data source, as in *Scenarios 2* and *3*. These two scenarios, however, are rather extreme cases, while most of the project management situations lay somewhere in between. No methodology, however, currently exists to accommodate the decision-maker (e.g. project manager) who would like to generate his decision based on both subjective (i.e. the triangular distribution (15, 28, 19)) and objective information sources (beta (5, 3, 15, 28)). The inability to combine and fuse relevant project information effectively as one input diminishes the simulation model's purpose and remains its Achilles heel. One input model that reflects and fuses all relevant project data is needed, as illustrated in Figure 4-2.

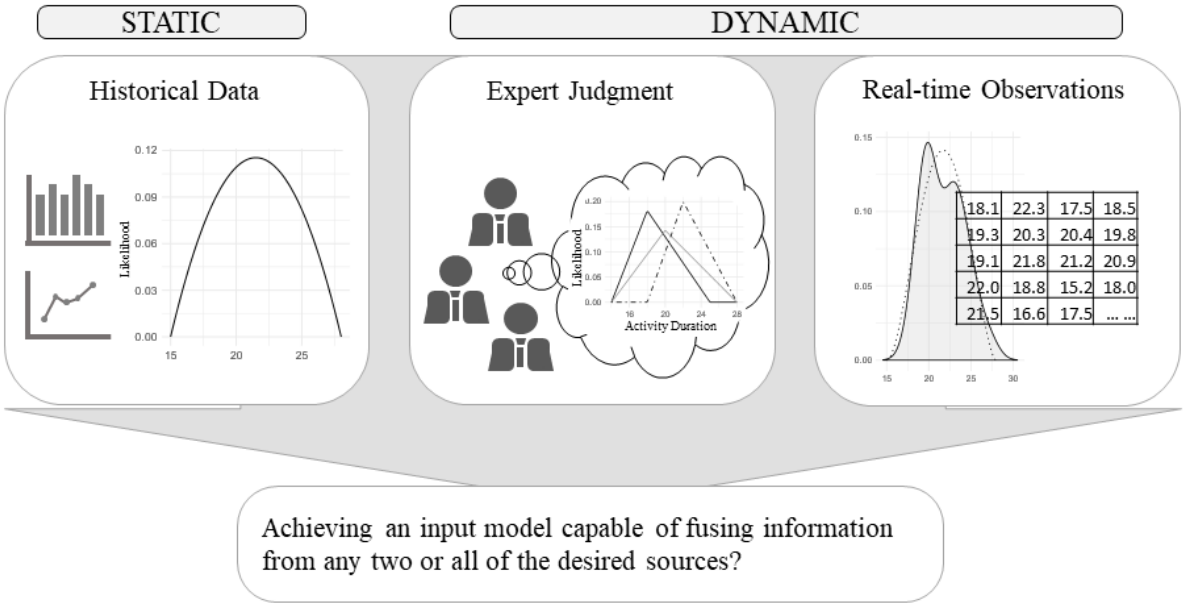


Figure 4-2 Various information sources for the input model

This chapter addresses the practical challenges illustrated in *Scenario 4* and proposes to couple the MCMC-based numerical method with a weighted GA algorithm, resulting in a universal input model updating algorithm for fusing project data collected through objective and subjective sources. The remainder of the chapter is organized into a detailed literature review on input-model updating techniques, their applications, and limitations. The hypothesis first introduces the two pillars of the proposed methodology—axiom-based aggregation methods and MCMC-based numerical approach–then proposes a hypothesis. The methodology section outlines the proposed input model updating approach, which is then tested through a Monte Carlo study as a "Proof of Concept." Lastly, an illustrative case study is presented to demonstrate the practicality, functionality, and feasibility of the proposed method in a typical industrial construction project setting.

4.3. LITERATURE REVIEW

Many studies have provided practitioners with guidance on how to fit univariate input models based on observational data (Abourizk and Halpin 1994, Nelson and Yamnitsky 1998, Biller and Nelson 2002, Kuhl et al. 2006) or expert knowledge (DeBrota 1989, AbouRizk and Halpin 1991). However, most construction simulation inputs demand a combination of data origins. Indeed, the need for a solid methodology to extract useful information from heterogeneous data has been highlighted by several researchers in the construction and civil engineering domains (Pradhan and Akinci 2012, Soibelman and Kim 2002, Chen, et al. 2005).

In the past few decades, construction industries have benefited from the rapid development of information technology. The implementation of various sensor technologies (e.g. Radio Frequency Identification, GPS, laser, and vision-based detection) in construction has drastically improved the efficiency of the data collection process in the industry (Zhang et al. 2017). Several researchers have demonstrated incorporating real-time (or near real-time) data into simulation models to facilitate project management—such as progress monitoring for earthmoving operations (Vahdatikhaki and Hammad 2014, Louis and Dunston 2017), productivity prediction (Seresht and Robinson Fayek 2018), equipment management (Li and Liu 2012, Akhavian and Behzadan 2013, ElNimr et al. 2016, Liu et al. 2020), quality management (Akinci et al. 2006, Ji and AbouRizk 2018), production planning and controlling (Altaf et al. 2018), and scheduling (Song and Eldin 2012). Demonstrated techniques for processing messy real-time data for simulation inputs include rule-based systems (knowledgebases that encapsulate expert rules to convert the noisy data) (Vahdatikhaki and Hammad 2014, Akhavian and Behzadan 2013), designing models to only respond to standardized signal inputs(Li and Liu 2012), finite-state machines (Louis and Dunston 2017), random sample consensus algorithms (extracting a subset from the noisy data using random

sample consensus algorithms) (Altaf et al. 2018), and tuning the simulation parameters upfront to take in data without further preprocessing. In these aforementioned studies, the input modeling process in regards to real-time field data and subjective information, have been 1) specific to its application or tasks and not generic for any stochastic simulation; 2) static for subjective information, such as experts' knowledge (e.g. a rule-based reasoning process, and a heuristic knowledgebase); 3) unable to address the fusing of data and information collected from various origins, including observational data and subjective data in real-time.

The Bayesian approach has been also studied for real-time simulation input (Brandley et al. 2015; Chung et al. 2004; Ji and AbouRizk 2017; Milo et al. 2015; Zhang et al. 2015), but the methods proposed in these studies are limited to specific cases due to conjugacy. Later, Wu et al. (2020) demonstrated a MCMC-based numeric Bayesian approach as a universal, real-time, input model updating method for any stochastic simulation model. However, Bayesian inference fuses information by reallocating credibility across the possibility of the parameter(s) values (the parameter(s) of the probability input model)—an indirect way of updating the probability distribution—which more than often are not intuitively meaningful (Kruschke, 2014). For instance, to update a beta distribution beta (*L, U, a, b*) with four parameters, the domain expert needs to 1) identify the parameter(s) that requires updating, followed by 2) properly select prior distribution(s) for each of the selected parameter(s), then 3) perform the numeric-based Bayesian inference to achieve posterior distribution(s) for all selected parameter(s), lastly 4) simulate the updated input model (the posterior predictive distribution) (Wu et al. 2020). Within this process, the choice of the priors heavily affects the resulting input model derived from Bayesian inference. However, studies in addressing this challenging process of defining priors are in its infancy for construction domain: Li et al (2019)

demonstrated a special case of the binomial distribution, with prior determined by only learning factor. As the case presented in *Scenario 4* of the Motivation section, using Bayesian inference to fuse expert's judgment (expressed as a triangular distribution), historical data (fitted as a beta distribution), and real-time observations, is extremely difficult, particularly in defining prior(s).

None of the advancements from the aforementioned research provides a direct, effective, and generic solution to fuse heterogeneous data—including historical project data, dynamic subjective information, and actual observational data—expressed as parametric distributions in real-time. Nevertheless, the existing research has inspired the authors to explore the possibility of an axiom-based numerical approach for aggregating information.

4.4. HYPOTHESIS

4.4.1. Axiom-based aggregation methods

Axiom-based aggregation approaches, such as weighted averages, have been studied for combining parametric distributions (Stone 1961, Genest and Zidek 1986, Clemen and Winkler 1999, Ayyub 2001). The weighted arithmetic average has been extensively used in risk management for aggregating diverse opinions due to the approach's simplicity and flexibility (Schmucker 1982, Sharfman and Fernando 2008, Yager and Kacprzyk 2012, Liu et al. 2015). The disadvantages of the weighted arithmetic average have also been studied and reported. For example, the aggregated distribution is typically multi-modal, making it difficult for use as an input model in simulations (Genest and Zidek 1986). Additionally, Winkler and Cummings (1972) have pointed out that the results of the weighted arithmetic averages are relatively insensitive to the selection of weights. Many of the above-listed

shortcomings can be overcome by using the weighted geometric average (Genest and Zidek 1986), defined as:

$$p(x) = k \prod_{i=1}^{n} p_i(x)^{w_i} \quad (1)$$

where, $k = \int \prod_{i=1}^{n} p_i^w dx$ is a normalizing constant, and $\sum_{i=1}^{n} w_i = 1$.

Practically, the challenges of deriving an analytical solution to the normalizing constant k (or the absence of it) significantly limit the applications of this method (Lindley 1985). A significant amount of research exists that uses the weighted GA for aggregating expert opinions through satisfying a number of reasonable axioms or limiting the form of pooling function (Bordley 1982, Morris 1974, Genest et al. 1984). Consequently, the weighted GA has been limited to a single source of information (expert opinions) with limited forms of the distribution function.

The geometric mean does not exist if one or more data points are zero. Often these zero values indicate less than a certain limit of detection (Costa 2017). The conventional methods include substituting 0 with a pre-determined value such as half the limit of detection, the limit of detection divided by the square root of two, the detection limit itself, or some other small value (Kayhanian et al. 2002). Sometimes the value 1 (or another constant) is added to all values to eliminate zeros or negative values, or in the case of frequency data, by adding 0.5 to all values (McDonald 2009).

4.4.2. MCMC-based numerical approach

As both computer hardware and random sampling algorithms continue to improve, many of the traditional challenges in deriving analytical solutions are being solved by developing accurate numerical approximations. One example is the Metropolis-Hastings (MH) algorithm,

which belongs to a larger class of sampling algorithms known as MCMC algorithms and has been recognized as one of the ten most influential algorithms in the development and practice of science and engineering in the 20th century (Beichl and Sullivan 2000). The MH algorithm is the most popular MCMC method (Hastings 1970, Metropolis et al. 1953); many other MCMC algorithms can be treated as a special case or extension of MH (Andrieu et al. 2003). The MH-based MCMC approach effectively solves integration and optimization problems and has been extensively applied to calculate the normalizing factor—such as the case in deriving normalizing factor for equation (1)—when an analytical solution is absent (Andrieu et al. 2003).

A handful of the above-mentioned studies have explored the MCMC-based numerical approach in developing decision-support applications in construction (Wu et al. 2020, Ji and AbouRizk 2017). Ji and AbouRizk (2017) demonstrated and compared the capability of a MH-based MCMC approach in achieving an accurate approximation to the true target distribution. Wu et al. (2020) further deployed a MCMC-based numerical approach in achieving an updated 4-parameter beta distribution for simulation input where no analytical solution exists.

4.4.3. Hypothesis Statement

Given that 1) the weighted GA effectively fuses information expressed in form of parametric distributions but has seen limited application due to the absence of an analytical solution for the normalizing constant k; 2) the MH-based MCMC numerical method can approximate target distributions when analytical solutions are absent; and 3) compared with Bayesian inference (Wu et al. 2020), the weighted GA aggregates information expressed as distributions directly—without the four steps as elaborated in the Literature Review section. This research is the first to propose coupling the MH-based MCMC numeric approach with a

weighted GA to achieve an input model updating method that fuses data collected through various means.

4.5. METHODOLOGY

The methodology of this study is illustrated in Figure 4-3. If Y denotes the underlying construction activity and $p(y)$ presents the probability density function (PDF) of Y, then $p_i(y)$ represents PDF generated from each of the various information sources, where $i = 1, \dots, n$ is the number of sources. To fuse the information generated from all sources, the proposed method updates the input model $p(y)$ through the weighted GA method. Thus, the probability of any point within the boundary reflects the information of each PDF originating from various sources.

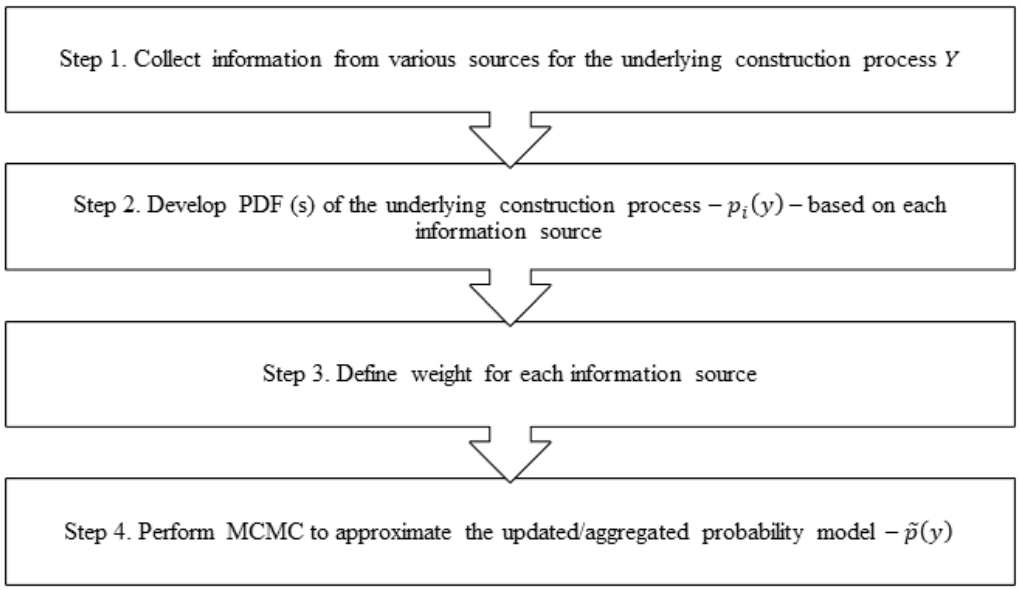


Figure 4-3 Proposed methodology

Thus, $\tilde{p}(y)$ denotes the updated/aggregated PDF of Y, mathematically as follows:

$$\tilde{p}(y) = K \prod_{i=1}^{n} p_i(y)^{w_i} \quad (2)$$

where K is the normalizing constant; the integral of $\tilde{p}(y)$ is 1; and w_i is the non-negative weight for each probability distribution and sums up to 1. Weights are herein assumed to be equal, as the choice of weight is beyond the scope of this research. With the help of the MH algorithm, $\tilde{p}(y)$ could be approximated without the analytical solution of K. Steps of the Metropolis methods are demonstrated as follows:

Step 1: Randomly generate a proposed leap, $\Delta y \sim \text{normal}(\mu = 0,\ \sigma)$, and denote the proposed value of the parameter as $y_{proposed} = y_{current} + \Delta y$

Step 2: Calculate the probability of moving to the proposed value:

$$p_{move} = \min(1, \frac{\tilde{p}(y_{proposed})}{\tilde{p}(y_{current})}) = \min\left(1, \frac{K \prod_{i=1}^{n} p_i(y_{proposed})^{\frac{1}{n}}}{K \prod_{i=1}^{n} p_i(y_{current})^{\frac{1}{n}}}\right),$$

In both theory and practice, the log-likelihood is used instead of the likelihood on both the record-level and model-level. Thus:

$$p_{move} = \min\left(\log(1), \log\left(\frac{\tilde{p}(y_{proposed})}{\tilde{p}(y_{current})}\right)\right)$$

$$= \min\left(\log(1), \frac{1}{n}\left\{\sum_{i=1}^{n} \log[p_i(y_{proposed})] - \sum_{i=1}^{n} \log[p_i(y_{current})]\right\}\right)$$

Step 3: Accept the proposed parameter value if a random value sampled from a [0,1] uniform distribution is less than the p_{move}; otherwise, reject the proposed parameter value and tally the current value again.

4.6. PROOF OF CONCEPT – MONTE CARLO STUDY

Three sets of MC experiments have been developed to test the proposed method. The three most widely used distributions for modeling construction processes are tested as listed in Table 4-1: uniform, triangular, and beta distributions. Uniform and triangular distributions are commonly used for expressing expert opinion. The uniform distribution represents a non-informative judgment for the underlying interest by placing equal probability across the possible value range, and triangular distributions effectively represent three-point estimates (Chau 1995). Generalized beta distributions are one of the most widely adopted distributions for modeling construction processes (Chau 1995) and have been successfully employed to model construction activity durations, costs, safety indicators, etc. For illustrative purposes, the experiments were performed on the interval of [0,1], assuming that boundaries are often well-established and the research interest remains in the shape of the aggregated distribution.

Table 4-1 The construction of the Monte Carlo study

Monte Carlo Study	Distribution 1 (D1)	Distribution 2 (D2)	Randomly generated parameters of D1 and D2
Experiment 1	Uniform (0, 1)	Beta (a, b)	$\{a, b\} \sim Uniform(1,10)$
Experiment 2	Triangular (0, 1, c)	Beta (a, b)	$\{a, b\} \sim Uniform(1,10)$; $c \sim Uniform(0,1)$
Experiment 3	Beta (a, b)	Beta (c, d)	$\{a, b\} \sim Uniform(1,10)$; $\{c, d\} \sim Uniform(1,15)$

As illustrated in Figure 4-4, the MC study validates the *proposed method* (PM) by examining the results with 100,000 samples generated using *mixture density* (MD): $\frac{1}{2}$ D1 + $\frac{1}{2}$ D2 . The PM results were also compared with the results from the widely-adopted *weighted arithmetic average method* (WAAM). At the start of each iteration i, parameters of the corresponding

distributions (D1 and D2) were generated randomly. Based on this set of distributions, a total of 300,000 samples are collected: 100,000 MCMC samples from the PM; 100,000 Monte Carlo samples from WAAM; and 100,000 samples generated using mixture density $\frac{1}{2}$ D1 + $\frac{1}{2}$ D2. After the sample collection, we estimated the shape parameters of three sample sets, calculated of the mean and variance, and plotted the updated distributions against the original D1 and D2. Due to the practicality and flexibility, all sample sets were fitted as a beta distribution. To examine the updated distributions from the two methods with the samples generated through the mixture density, changes of mean and variance were calculated and compared to the MD samples. The percentage differences are defined as below:

$$\%\ difference\ of\ mean = \frac{|mean\ of\ the\ MD\ samples - mean\ of\ the\ updated\ distribution|}{mean\ of\ the\ MD\ samples} \times 100\% \quad (3)$$

$$\%\ difference\ of\ variance = \frac{(variance\ of\ the\ MD\ samples - variance\ of\ the\ updated\ distribution)}{variance\ of the\ MD\ samples} \times 100\% \quad (4)$$

where absolute value has been calculated for the percentage difference of the mean to represent the deviation irrespective of direction. Potential negative values have been kept for the percentage difference to identify the shrinkage (positive value) or expansion (negative value) of the variance.

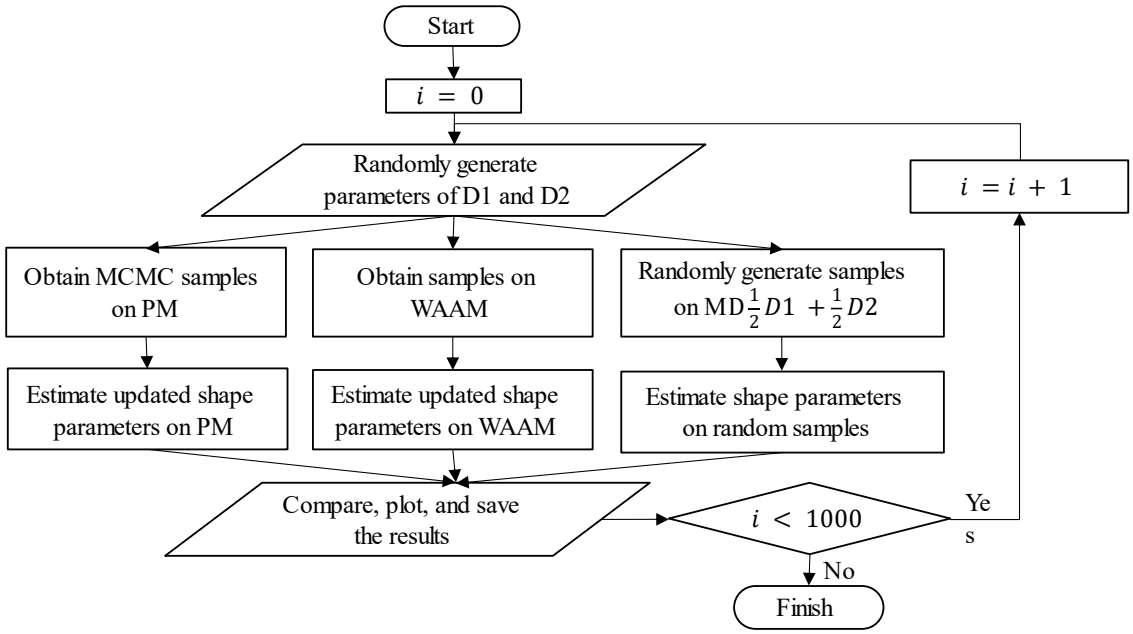


Figure 4-4 Flow chart of the Monte Carlo study

4.6.1. Monte Carlo Experiment 1 – Uniform and Beta

A sample plot (randomly picked) from the 40th experiment run is presented in Figure 4-5. The solid line represents the randomly generated beta distribution, the thin dashed line represents the uniform distribution, the dash-dotted line represents the updated distribution from PM, the thick long-dashed line represents the updated distribution from WAAM, and the dotted line represents the updated distribution from MD. Both WAAM (thick long-dashed) and PM (dash-dotted) present a strong ability to mimic the shape of the given beta distribution, regardless of the uniform distribution. The PM exhibits an expansion of the variance under the influence of a uniform distribution similar to the MD (dotted), while WAAM (thick long-dashed) presents a decreased variance. To better present the results of the 1,000 experiment runs, Table 4-2 presents summary statistics of the percentage differences. Boxplots and jitter point plots of the same data are shown in Figure 4-6. The PM.mean and PM.var respectively represent the percentage difference of mean and variance

for the PM. Similarly, WAAM.mean and WAAM.var respectively represent the percentage difference of the mean and variance for the weighted arithmetic average method.

Table 4-2 Summary statistics of percentage difference in Monte Carlo Experiment 1

	PM.mean	WAAM.mean	PM.var	WAAM.var
Min.	0.001%	0.000%	15.814%	49.680%
1st Quartile	2.881%	0.047%	42.189%	50.893%
Median	6.248%	0.103%	50.725%	53.911%
Mean	7.823%	0.121%	49.251%	56.159%
3rd Quartile	10.658%	0.173%	56.515%	60.084%
Max.	37.649%	0.537%	77.682%	73.981%

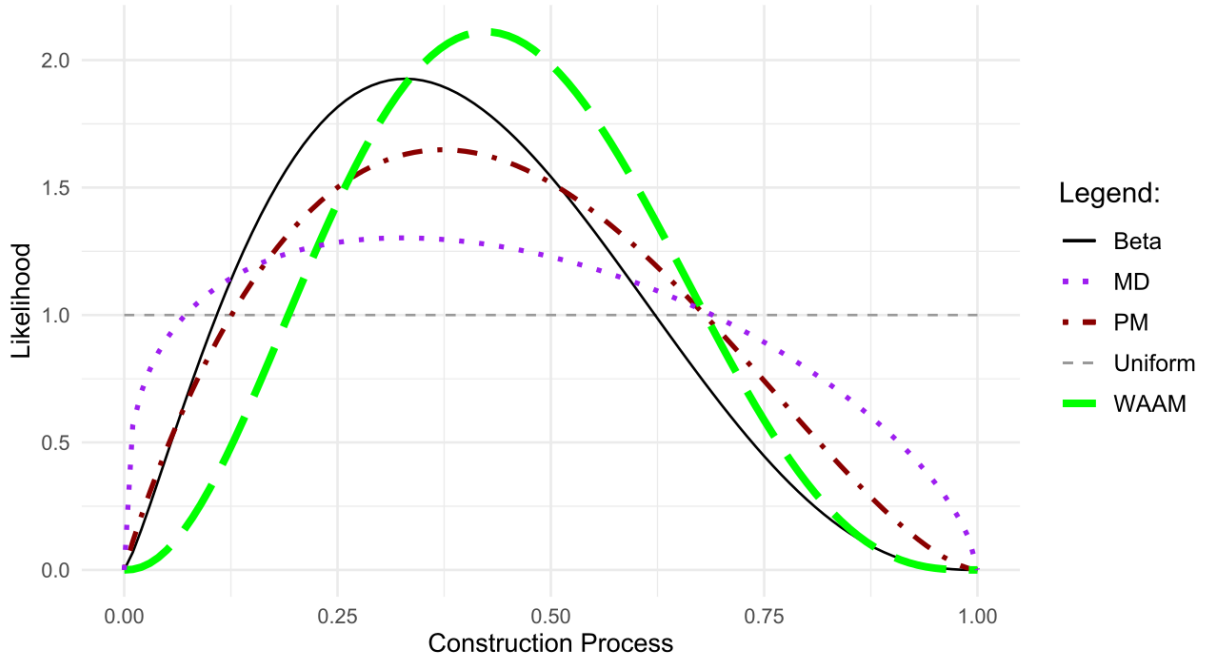


Figure 4-5 Input model plot (run #40)

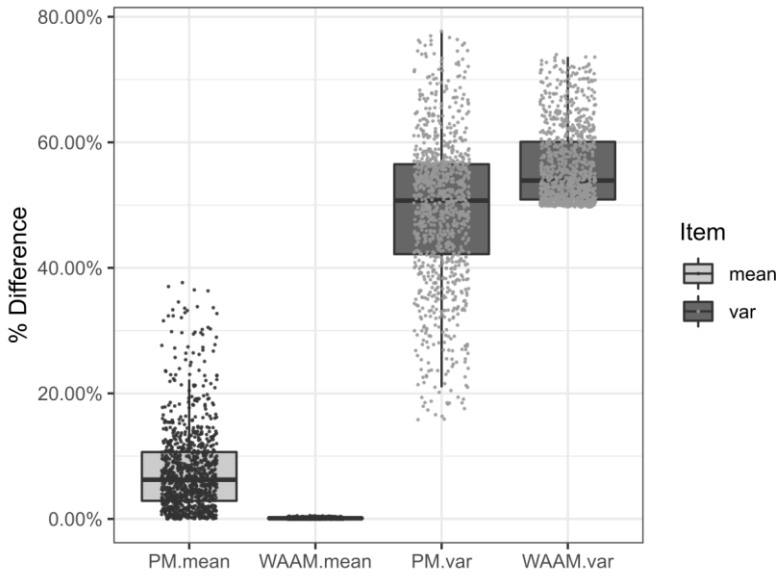


Figure 4-6 Boxplot with jitter points percentage difference for mean and variance in Monte Carlo Experiment 1

Compared to the PM, WAAM exhibits a strong capability in approaching the MD in terms of mean. The PM demonstrates consistent and reliable performance for the majority of the run and has a < 10% deviation from the mean of the MD. In terms of the variance, however, the PM demonstrates better performance in approximating the result generated using the MD. Due to the effect of uniform distribution, the MD samples recorded a wide variance from the original beta distribution. This effect has been well captured by the PM, though not as intensely. As demonstrated in Figure 4-5, resulting input models from WAAM sometimes displayed shrinkage in variance from the original beta distribution. The decrease in variance indicates a stronger belief with fewer uncertainties, counter to the logic of combining a beta distribution with a uniform distribution.

4.6.2. Monte Carlo Experiment 2 – Triangular and Beta

A randomly selected sample plot from experiment run #301 is presented in Figure 4-7, along with summary statistics of the percentage difference (Table 4-3 and Figure 4-8).

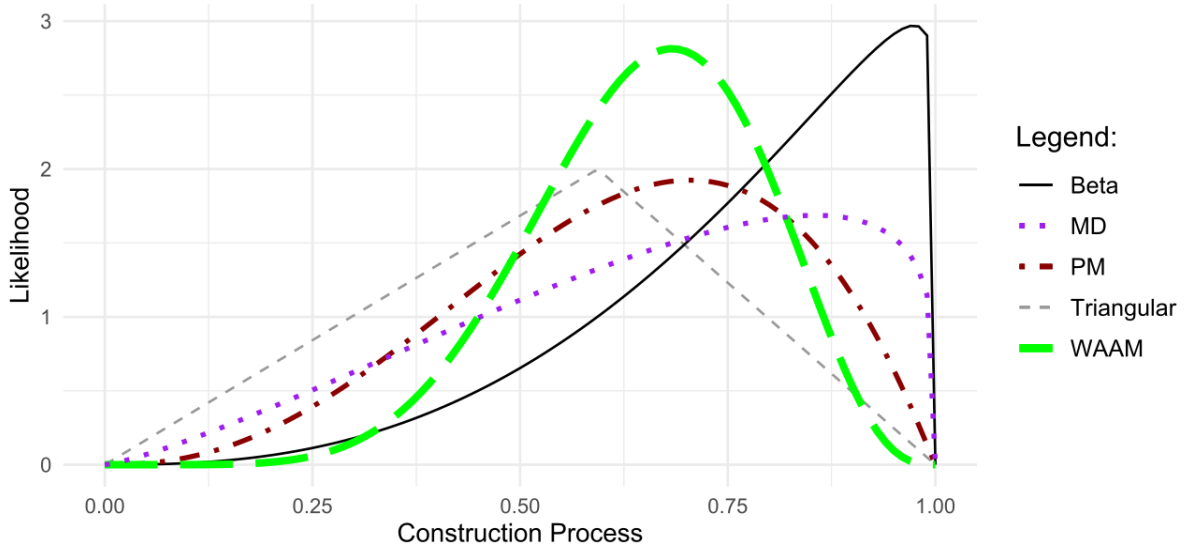


Figure 4-7 Input model plot (run #301)

Table 4-3 Summary statistics of percentage difference of Monte Carlo Experiment 2

	PM.mean	WAAM.mean	PM.var	WAAM.var
Min.	0.01%	0.00%	9.85%	49.62%
1st Quartile	1.65%	0.04%	25.63%	51.80%
Median	4.09%	0.08%	34.00%	56.94%
Mean	5.72%	0.10%	36.13%	59.52%
3rd Quartile	8.27%	0.14%	45.36%	64.95%
Max.	28.59%	0.50%	74.60%	84.85%

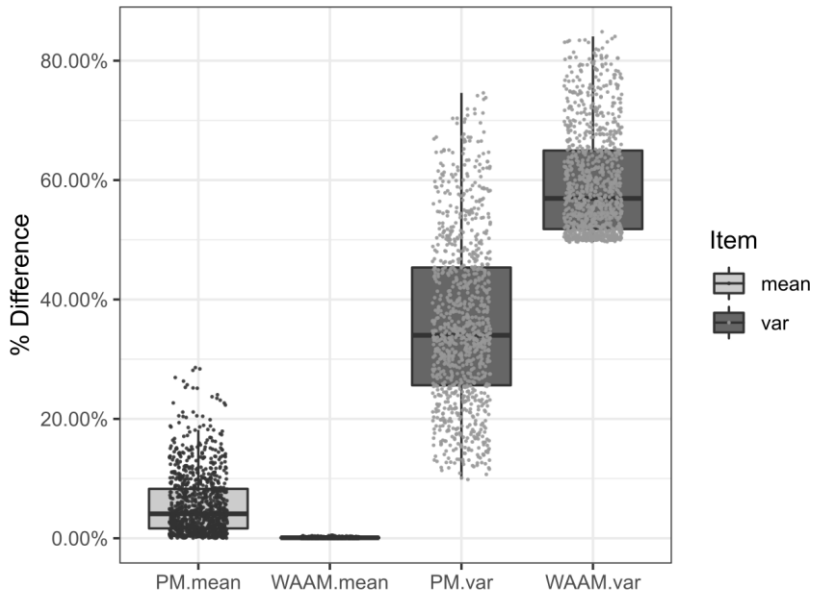


Figure 4-8 Boxplot with jittered points of percentage difference of mean and variance for Monte Carlo Experiment 2

In Experiment 2, WAAM outperformed the PM with regards to the arithmetic mean of the MD; however, the PM produces a reliable central tendency with the majority (the third quartile) of PM.mean below 10%. With regards to variance, the PM demonstrates a far better performance than WAAM: the interquartile range is located between 25% and 45% compared to WAAM's 50% to 65%. As a result, the PM produces more accurate and reliable interval estimates compared to WAAM.

4.6.3. Monte Carlo Experiment 3 – Beta and Beta

Figure 4-9and Figure 4-10 present sample-run result plots (run #77 and #80, respectively). Summary statistics are presented in Table 4-4, and Figure 4-11 presents boxplots and jitter point plots. In Experiment 3, WAAM excels in achieving the central tendency of the aggregated distribution (i.e. arithmetic mean). The PM, however, surpasses WAAM at

producing closer and more reliable results in terms of the dispersion of the aggregated distribution (i.e. variance), which is an indispensable component to interval estimate.

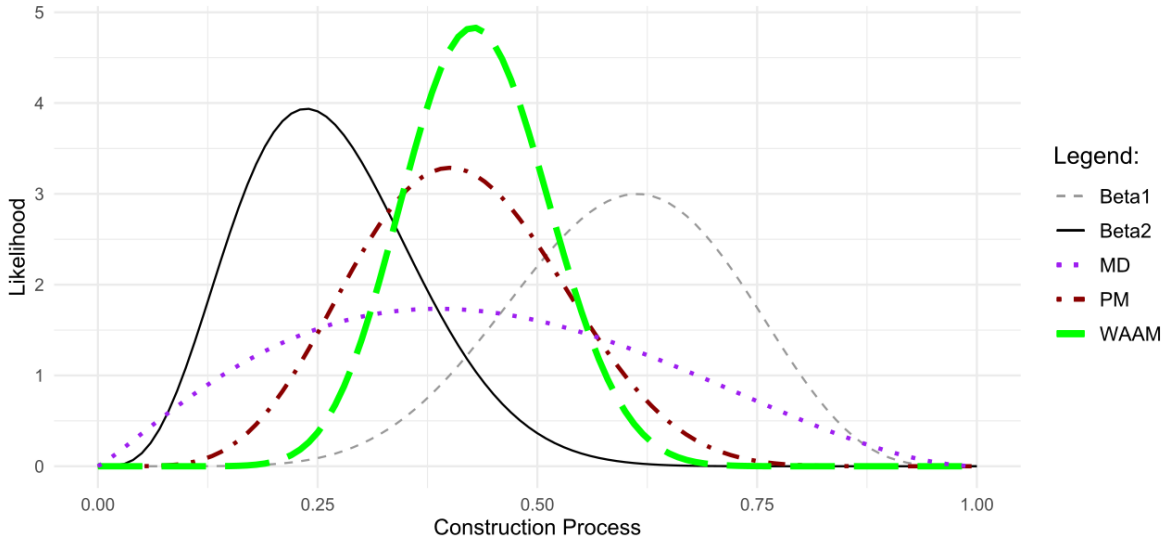


Figure 4-9 Input model plot (run #77)

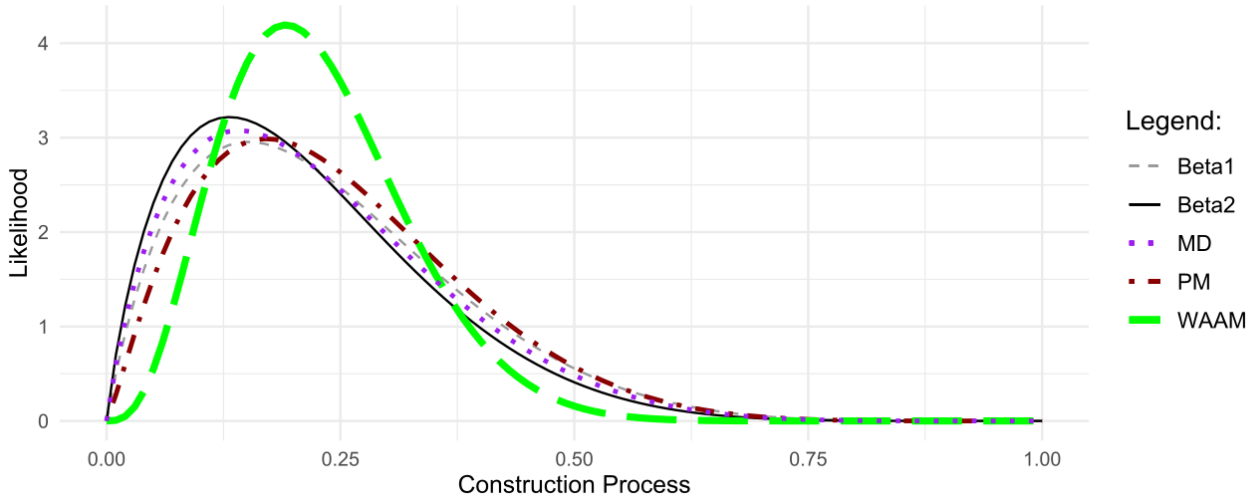


Figure 4-10 Input model plot (run #80)

Table 4-4 Summary statistics of percentage difference of Monte Carlo Experiment 3

	PM.mean	WAAM.mean	PM.var	WAAM.var
Min.	0.00%	0.00%	-8.45%	49.65%
1st Quartile	1.55%	0.03%	22.90%	55.75%
Median	4.36%	0.06%	42.26%	67.82%
Mean	6.32%	0.08%	42.47%	68.83%
3rd Quartile	9.32%	0.10%	62.30%	80.60%
Max.	38.99%	0.43%	88.66%	97.05%

Additionally, the PM.var has a much wider interquartile range than WAAM. With further analysis, a clear trend is discovered: if D1 and D2 have opposite skewness, it leads to a wider discrepancy between the PM's variance and MD's variance (e.g. run #77 in Figure 4-9). On the other hand, if D1 and D2 are same-side-skewed and their modes are very close, the PM produces very close results compared to MD (e.g, run #80 in Figure 4-10). The data points are thus filtered and then summarised based on the skewness of D1 and D2. Figure 4-12 shows the percentage difference results as boxplots. Similar behaviour was also recognized in Experiment 2 (Figure 4-13).

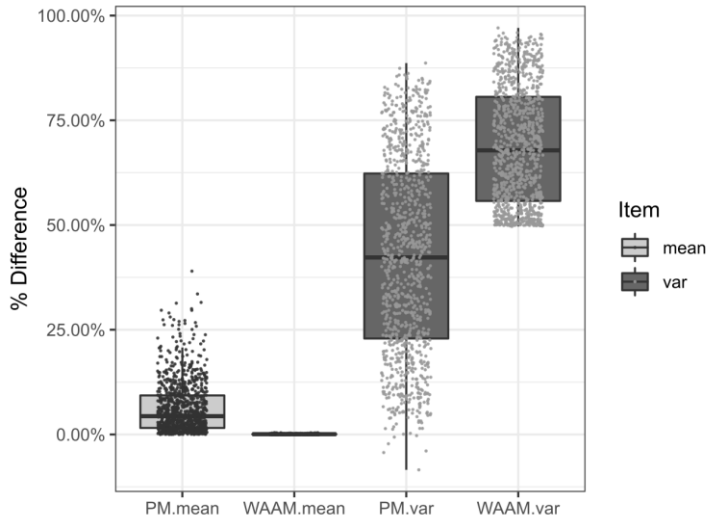


Figure 4-11 Boxplot with jittered points of percentage difference of mean and variance for Monte Carlo Experiment 3

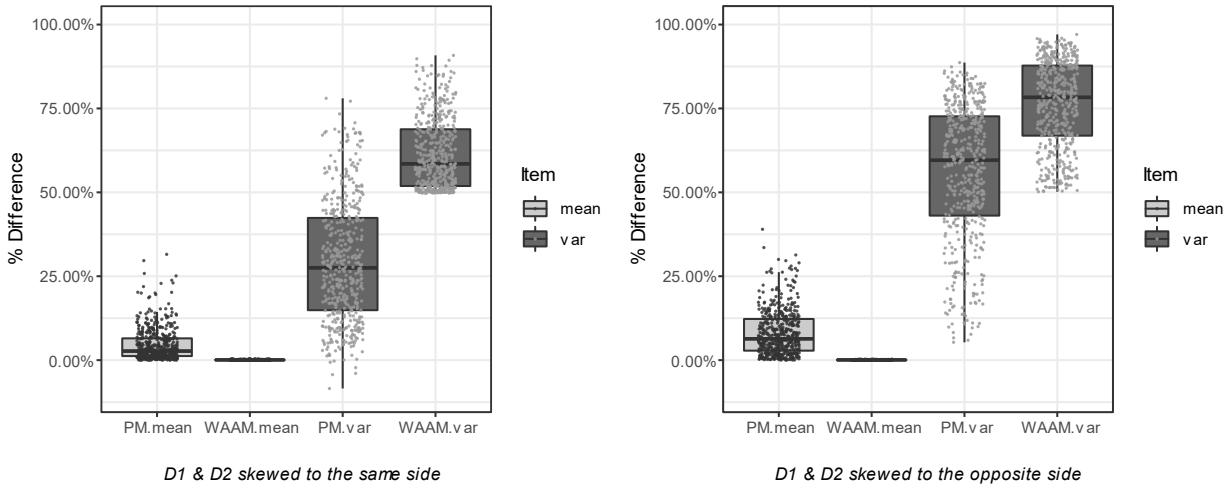


Figure 4-12 Boxplot of percentage difference of mean and variance with D1 and D2 skewed to the same side (left) and opposite side (right) for Monte Carlo Experiment 3

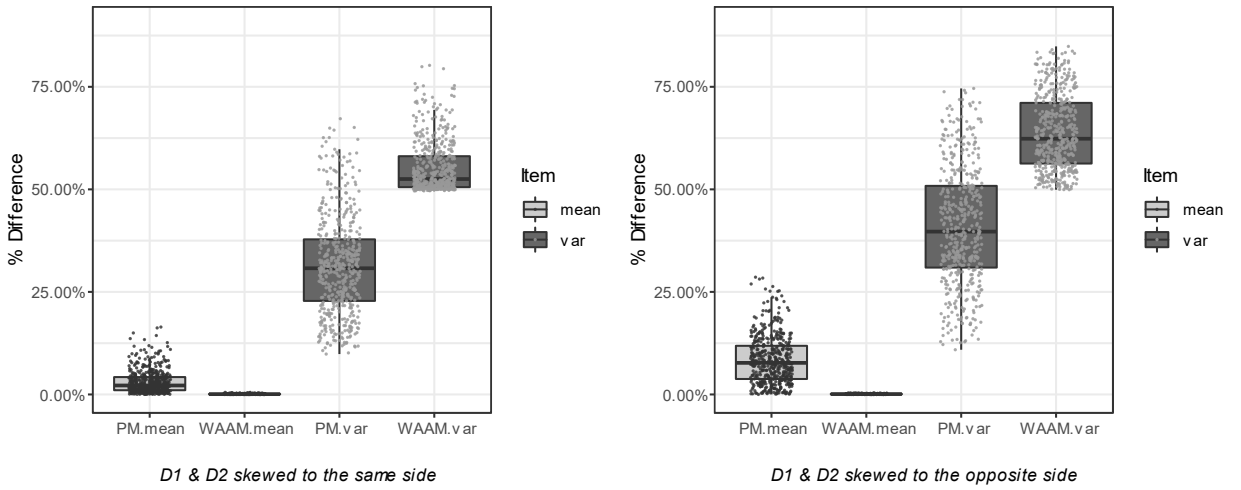


Figure 4-13 Boxplot of percentage difference of mean and variance with D1and D2 skewed to the same side (left) and opposite side (right) for Monte Carlo Experiment 2

4.6.4. Synopsis

Although this study did not exhaust all the possible combinations of parametric distributions, the three most commonly used distributions in modeling construction processes are thoroughly examined with clear trends discovered. This study demonstrates the following outcomes: 1) With regards to a point estimate (i.e. if the arithmetic mean is the interest), WAAM excels; 2) At times when the interval estimate is required (modeling construction processes imbued with uncertainties), the PM provides a more reliable and accurate result. Further, based on the three sets of experiments, it is unveiled that both the PM and WAAM tend to result in a stronger central tendency distribution compared with MD. For the PM, however, this central tendency only becomes significant when the modes of the two original distributions are set widely apart (i.e. opposite skewness). For WAAM, this central tendency is dominant and consistent regardless of the shapes of the original distribution, eliminating uncertainty and yielding an over-optimistic interval estimate.

4.7. ILLUSTRATIVE CASE STUDY

To demonstrate the feasibility and functionality of the PM, a discrete event simulation (DES) model has been developed to represent a simplified critical path method (CPM) schedule network for site welding activities in a typical industrial construction project. There are ten 20-inch standard carbon steel welds needing completion by the same crew in three modules: Mod A, Mod B, and Mod C. The precedence relationships between the activities are assumed to be "finish-to-start." As is typical for repetitive construction activities, the duration of this type of weld is well documented from previous projects and follows a generalized four-parameter beta distribution as illustrated by a dash-dotted line in Figure 4-14: beta (3, 3, 15, 28). The superintendent in charge forecasts lower productivity for the following reasons:

1. The crew consists of a higher percentage of apprentices; as such, the learning curve is expected to be steeper.
2. Mod B is stacked above Mod A, and Mod C above Mod B, resulting in difficulties accessing the working area.

The superintendent provides a three-point estimate with a most-likely duration of 26 hours, a low of 15 hours, and a high of 28 hours. The resulting triangular distribution is illustrated by the dotted line in Figure 4-14: triangular (15, 26, 28). Knowing this superintendent's estimate was overly conservative, the project manager decides to blend the historical information with the superintendent's judgment. The project manager implements the PM with equal weight on each information source. The input model of the manager's choice (i.e. the PM) is computed and plotted in Figure 4-14 as a solid line.

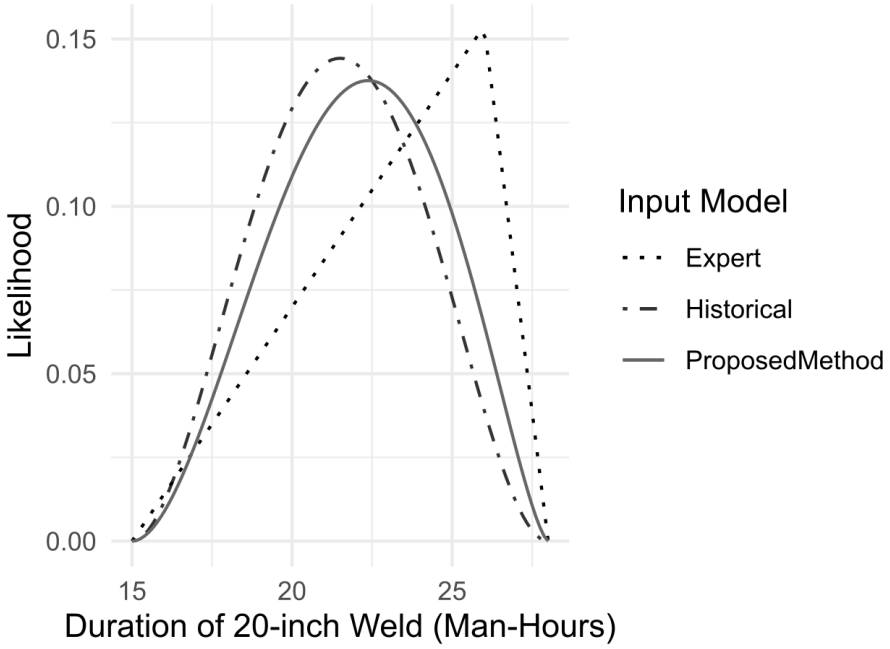


Figure 4-14 Input models of duration for a 20-inch weld

To mimic the general noise in data and decreased productivity resulting from the superintendent's evaluation (assuming the actual project conditions align with the superintendent's judgment), the simulated actual data points were generated using one third of the uniform distribution, uniform (22, 27), and two-thirds of generalized beta distribution (3, 3, 15, 28), creating a higher probability of sampling lower productivity.

The simulated project duration is generated by running the DES model 100,000 times with the three input models: the historical data, expert opinion, and PM. The results, generated with different input models, were compared and examined with the random samples generated through the MD $\frac{1}{3}\,\text{uniform}\,(22, 27) + \frac{2}{3}\,\text{beta}\,(3, 3, 15, 28)$. Specifically, a 95% high density interval (HDI) (sometimes referred to as a high density region (HDR)) and mean value were calculated and plotted as shown in Figure 4-15. HDI is an effective method of summarizing a distribution and indicates the most credible parts of a distribution (Kruschke 2014). For example, a 95% HDI is an interval wherein every value within the interval has

higher credibility than any of the points outside of the interval, and the area under the density curve between the two limits covers 95% of the area. As shown in Figure 4-15, the project duration forecast using the PM (solid line) is closer to the actual duration of the project (dash line). Further, the PM-generated HDI is located between the expert-opinion-generated HDI (dotted line) and the historical data (dotted-dash line), demonstrating the capability to aggregate information from both sources.

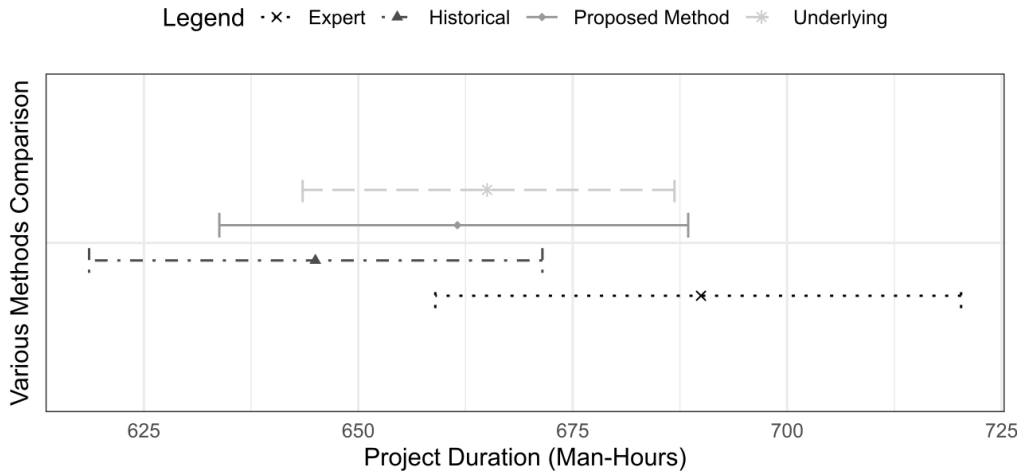


Figure 4-15 HDI plot of forecasted project duration using various inputs

The case study strongly demonstrates the functionality and practicality of implementing the PM to fuse information from historical data and expert experience to achieve an integrated forecast dynamically. The PM could be generalized to combine actual project performance, historical project data, multiple experts' opinions, or other related project data for an increasingly more effective dynamic input-model updating method.

4.8. CONCLUSIONS

This research is the first to propose a MCMC-based weighted GA method for effective fusion of information generated from diverse sources, addressing the practical challenges associated

with modeling assorted inputs. This research has developed a universal input model updating method for any given univariate parametric continuous probability distribution. Due to the effectiveness of the MH algorithm, the PM provides solutions regardless of the choice of probability distributions (i.e. in absence of an analytical solution of the normalizing factor), thus expanding potential applications in engineering and management.

A "Proof of Concept" MC study tested the PM against the WAAM, and the results have been compared with the MD samples. This study validated the capability of the PM in achieving a reliable, accurate distribution to counteract the uncertainties inherent in construction processes. Further, an illustrative case study is used to demonstrate the generalizability, feasibility, and functionality of the PM for aggregating subjective and objective information to update simulation input models in real time. The PM has been found capable of (1) effectively and efficiently updating input models given new sources of information, (2) accurately approximating the target probability distribution, (3) reliably fusing information from diverse sources, including subjective judgment and objective observations, and (4) being generalized and applied to combinations of any given univariate continuous probability distributions.

The proposed methodology facilitates the dynamic fusion of data generated from various sources and aggregates all information to serve as the simulation model input. This input modeling method enhances the model's practicality and predictability for construction project management where expert knowledge has a heavy influence on the decision-making process. The traditional simulation model benefits from effective integration of data by considering information from multiple sources and extending usage throughout project planning and execution. Potential applications in construction engineering and management fields include, but are not limited to, the following: production planning and scheduling, safety management,

cost forecast, and risk management. Actual project information such as productivity factors, safety indices, and actual costs will be aggregated with the historical data or expert judgment using the PM in a real-time manner. The aggregated input models will then be utilized in simulation-based decision-support systems to reflect dynamic project performance, deriving more accurate and meaningful decision-support output for practitioners. The proposed methodology can benefit any MC-driven analytics or dynamic, data-driven stochastic simulations developed for various engineering, management, and applied science fields.

This research should be considered in light of several limitations. The proposed methodology itself does not guarantee improved accuracy. Rather, the PM provides a solution for fusion of information from various sources. The predictability of the model ultimately relies on the accuracy of all source information. Additionally, the selection of the weights is a complex problem that has not been discussed in this paper. Proper weight selection ensures an accurate and meaningful result. The advantages and disadvantages of implementing weighted geometric average are beyond the scope of this research. Although these aspects are briefly discussed in the Monte Carlo study, further research efforts are required. Finally, while the PM provides an approach for integrating information from various sources, the method itself does not represent a complete decision-support system.

4.9. ACKNOWLEDGEMENTS

This research is funded by an NSERC Collaborative Research and Development Grant (CRDPJ 492657).

4.10. REFERENCES

Abdelmegid, M. A., González, V. A., Naraghi, A. M., O'Sullivan, M., Walker, C. G., and Poshdar, M. 2017. "Towards a conceptual modeling framework for construction simulation." In *Proceedings of 2017 Winter Simulation Conference,* 2372-2383. Piscataway, NJ: IEEE.

Abdelmegid, M. A., González, V. A., Poshdar, M., O'Sullivan, M., Walker, C. G., and Ying, F. 2020. "Barriers to adopting simulation modelling in construction industry." *Automation in Construction*, *111*, 103046.

AbouRizk, S. M., 2010. "Role of simulation in construction engineering and management." *Journal of construction engineering and management*, 136(10): 1140-1153.

AbouRizk, S. M., Hague, S. A., and Ekyalimpa, R. 2016a. *Construction simulation: An introduction using Simphony*. University of Alberta, Edmonton, Canada.

AbouRizk, S. M., Hague, S., Ekyalimpa, R., and Newstead, S. 2016b. "Simphony: A next generation simulation modelling environment for the construction domain." *Journal of Simulation*, *10*(3): 207-215.

AbouRizk, S. M., and Halpin, D. W. 1992. "Statistical properties of construction duration data." *Journal of Construction Engineering and Management*, 118(3): 525-544.

AbouRizk, S. M., Halpin, D. W., and Wilson, J. R. 1991. "Visual interactive fitting of beta distributions." *Journal of Construction Engineering and Management*. 117(4): 589-605.

AbouRizk, S. M., Halpin, D. W., and Wilson, J. R. 1994. "Fitting beta distributions based on sample data." *Journal of Construction Engineering and Management*. 120(2): 288-305.

Akinci, B., Boukamp, F., Gordon, C., Huber, D., Lyons, C., and Park, K. 2006. "A formalism for utilization of sensor systems and integrated project models for active construction quality control." *Automation in construction*. 15(2): 124-138.

Akhavian, R. 2015. "Data-driven simulation modeling of construction and infrastructure operations using process knowledge discovery." Ph.D. thesis, Orlando, Florida: University of Central Florida

Akhavian, R., and Behzadan, A. H. 2013. "Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining." *Journal of Construction Engineering and Management*. 139(11): 04013021.

Al Qady, M., and Kandil, A. 2013. "Document discourse for managing construction project documents." *Journal of computing in civil engineering*, 27(5): 466-475.

Altaf, M. S., Bouferguene, A., Liu, H., Al-Hussein, M., and Yu, H. 2018. "Integrated production planning and control system for a panelized home prefabrication facility using simulation and RFID." *Automation in construction*, 85: 369-383.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. 2003. "An introduction to MCMC for machine learning." *Machine learning*. 50(1-2): 5-43.

Ayyub, B. M. 2001. *Elicitation of expert opinions for uncertainty and risks*. CRC press.

Beichl, I., and Sullivan, F. 2000. "The metropolis algorithm." *Computing in Science & Engineering*. 2(1): 65-69.

Behzadan, A. H., Menassa, C. C., and Pradhan, A. R. 2015. "Enabling real time simulation of architecture, engineering, construction, and facility management (AEC/FM) systems: a review of formalism, model architecture, and data representation." *ITcon*. 20: 1-23.

Biller, B., and Nelson, B. L. 2002. "Answers to the top ten input modeling questions." In Vol. 1 of *Proceedings of the 2002 Winter Simulation Conference*, 35-40. Piscataway, NJ: IEEE

Brandley, R. L., Bergman, J. J., Noble, J. S., and McGarvey, R. G. 2015. "Evaluating a Bayesian approach to demand forecasting with simulation." In *Proceedings of the 2015 Winter Simulation Conference*, 1868-1879. Piscataway, NJ: IEEE

Bordley, Robert F. 1982. "A multiplicative formula for aggregating probability assessments." *Management science.* 28(10): 1137-1148.

Caldas, C. H., Soibelman, L., and Han, J. 2002. "Automated classification of construction project documents." *Journal of Computing in Civil Engineering*, 16(4): 234-243.

Chau, K. W. 1995. "Monte Carlo simulation of construction costs using subjective data." *Construction Management and Economics*. 13(5): 369-383.

Chau, K. W. 1995. "The validity of the triangular distribution assumption in Monte Carlo simulation of construction costs: empirical evidence from Hong Kong." *Construction Management and Economics*. 13(1): 15-21.

Chen, P., Buchheit, R. B., Garrett Jr, J. H., and McNeil, S. 2005. "Web-vacuum: Web-based environment for automated assessment of civil infrastructure data." *Journal of computing in civil engineering*. 19(2): 137-147.

Chung, T. H., Mohamed, Y., and AbouRizk, S. 2004. "Simulation input updating using Bayesian techniques." In *Proceedings of the 2004 Winter Simulation Conference*, 1238-1243. Piscataway, NJ: IEEE

Chwif, L., Banks, J., de Moura Filho, J. P., and Santini, B. 2013. "A framework for specifying a discrete-event simulation conceptual model". *Journal of Simulation*, *7*(1): 50-60.

CII (Construction Industry Institute). 2015. *Innovative delivery of information to the crafts*. RT-327-1. Austin, TX: CII.

Clemen, R. T., and Winkler, R. L. 1999. "Combining probability distributions from experts in risk analysis." *Risk analysis*. 19(2): 187-203.

Costa, J. 2017. *Calculating Geometric Means* Accessed July 15, 2020. https://buzzardsbay.org/special-topics/calculating-geometric-mean/

DeBrota, D. J., Dittus, R. S., Roberts, S. D., and Wilson, J. R. 1989. "Visual interactive fitting of bounded Johnson distributions." *Simulation*. 52(5): 199-205.

ElNimr, A., Fagiar, M., and Mohamed, Y. 2016. "Two-way integration of 3D visualization and discrete event simulation for modeling mobile crane movement under dynamically changing site layout." *Automation in construction*, *68*: 235-248.

Genest, C., Weerahandi, S., and Zidek, J. V. 1984. “Aggregating opinions through logarithmic pooling.” *Theory and decision*. 17(1): 61-70.

Genest, C., and Zidek, J. V. 1986. “Combining probability distributions: A critique and an annotated bibliography.” *Statistical Science.* 1(1): 114-135.

Gong, J., and Carlos H. C. 2010. “Computer vision-based video interpretation model for automated productivity analysis of construction operations.” *Journal of Computing in Civil Engineering* 24(3): 252-263.

Hammad, A., and Zhang, C. 2011. “Towards real-time simulation of construction activities considering spatio-temporal resolution requirements for improving safety and productivity.” In *Proceedings of the 2011 Winter Simulation Conference,* 3533-3544. Piscataway, NJ: IEEE

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*. 57(1): 97-109.

Hubbard, D. W. 2009. *The failure of risk management: Why it's broken and how to fix it*. Hoboken, New Jersey, US: John Wiley & Sons.

Ji, W., and AbouRizk, S. M. 2017. “Credible interval estimation for fraction nonconforming: Analytical and numerical solutions.” *Automation in Construction*. 83: 56-67.

Ji, W., and AbouRizk, S. M. 2018. “Simulation-based analytics for quality control decision support: Pipe welding case study.” *Journal of Computing in Civil Engineering*, *32*(3): 05018002.

Kayhanian, M., Amardeep Singh, and Scott Meyer. 2002. "Impact of non-detects in water quality data on estimation of constituent mass loading." *Water Science and Technology,* 45(9): 219-225.

Kruschke, J. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London, UK: Academic Press.

Kuhl, M. E., Lada, E. K., Steiger, N. M., Wagner, M. A., and Wilson, J. R. 2006. "Introduction to modeling and generating probabilistic input processes for simulation." In *Proceedings of the 2006 Winter simulation conference,* 19-35. Piscataway, NJ: IEEE

Leite, F., Cho, Y., Behzadan, A. H., Lee, S., Choe, S., Fang, Y., and Hwang, S. 2016. "Visualization, information modeling, and simulation: Grand challenges in the construction industry." *Journal of Computing in Civil Engineering*, 30(6): 04016035.

Li, Y., Ji, W., and AbouRizk, S. M. 2019. "Enhanced Welding Operator Quality Performance Measurement: Work Experience-Integrated Bayesian Prior Determination." In *Computing in Civil Engineering 2019: Data, Sensing, and Analytics,* 606-613. Reston, VA: American Society of Civil Engineers.

ASCE International Conference on Computing in Civil Engineering 2019

Li, Y., and Liu, C. 2012. "Integrating field data and 3D simulation for tower crane activity monitoring and alarming." *Automation in Construction*. 27: 111-119.

Lindley, D. V. 1985. "Reconciliation of discrete probability distributions." *Bayesian statistics*, 2:375-390.

Liu, H. C., You, J. X., Lin, Q. L., and Li, H. 2015. "Risk assessment in system FMEA combining fuzzy weighted average with fuzzy decision-making trial and evaluation laboratory." *International Journal of Computer Integrated Manufacturing*. *28*(7): 701-714.

Liu, C., Lei, Z., Morley, D., and AbouRizk, S. M. 2020. "Dynamic, data-driven decision-support approach for construction equipment acquisition and disposal." *Journal of Computing in Civil Engineering*, 34(2): 04019053.

Louis, J., and Dunston, P. S. 2017. "Methodology for real-time monitoring of construction operations using finite state machines and discrete-event operation models." *Journal of construction engineering and management*, 143(3): 04016106.

Martínez-Rojas, M., Marín, N., and Vila, M. A. 2016. "The role of information technologies to address data handling in construction project management." *Journal of Computing in Civil Engineering*, 30(4): 04015064.

McDonald, J. H. 2009. *Handbook of biological statistics.* Vol. 2. Baltimore, MD: Sparky House Publishing.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics*. 21(6): 1087-1092.

Milo, M. W., Roan, M., and Harris, B. 2015. "A new statistical approach to automated quality control in manufacturing processes." *Journal of Manufacturing Systems*. 36: 159-167.

Morris, P. A. 1974. "Decision analysis expert use." *Management Science.* 20(9): 1233-1241.

Nelson, B. L., and Yamnitsky, M. 1998. "Input modeling tools for complex problems." In Vol. 1 of *Proceedings of the 1998 Winter Simulation Conference.* 105-112. Piscataway, NJ: IEEE.

Pradhan, A., and Akinci, B. 2012. "A taxonomy of reasoning mechanisms and data synchronization framework for road excavation productivity monitoring." *Advanced Engineering Informatics*. 26(3): 563-573.

Rao, U. S., Kestur, S., and Pradhan, C. 2008. "Stochastic optimization modeling and quantitative project management." *IEEE software*, 25(3): 29-36.

Sargent, R. G. 2010. "Verification and validation of simulation models." In *Proceedings of the 2010 winter simulation conference,* 166-183 Piscataway, NJ: IEEE.

Schmucker, K. J. 1982. *Fuzzy sets: Natural language computations and risk analysis*. Rockville, MD, US: Computer Science Press.

Seresht, N. G., and Robinson Fayek, A. 2018. "Dynamic modeling of multifactor construction productivity for equipment-intensive activities." *Journal of Construction Engineering and Management*, *144*(9): 04018091.

Sharfman, M. P., and Fernando, C. S. 2008. "Environmental risk management and the cost of capital." *Strategic management journal*. 29(6): 569-592.

Skibniewski, M., and Golparvar-Fard, M. 2016. "Toward a science of autonomy for physical systems: Construction." *A white paper prepared for the Computing Community Consortium committee of the Computing Research Association*. http://cra.org/ccc/resources/ccc-led-whitepapers/.

Soibelman, L., and Kim, H. 2002. "Data preparation process for construction knowledge generation through knowledge discovery in databases." *Journal of Computing in Civil Engineering*. 16(1): 39-48.

Song, L., and Eldin, N. N. 2012. "Adaptive real-time tracking and simulation of heavy construction operations for look-ahead scheduling." *Automation in Construction*. 27: 32-39.

Vahdatikhaki, F., and Hammad, A. 2014. "Framework for near real-time simulation of earthmoving projects using location tracking technologies." *Automation in Construction*. 42: 50-67.

Winkler, R. L., and Cummings, L. L. 1972. "On the choice of a consensus distribution in Bayesian analysis." *Organizational Behavior and Human Performance*. 7(1): 63-76.

Wu, L., Ji, W., and AbouRizk, S. M. 2020. "Bayesian Inference with Markov Chain Monte Carlo–Based Numerical Approach for Input Model Updating." *Journal of Computing in Civil Engineering*. 34(1): 04019043.

Yager, R. R., and Kacprzyk, J. (Eds.). 2012. *The ordered weighted averaging operators: theory and applications*. Berlin, Germany: Springer Science & Business Media.

Zhang, M., Cao, T., and Zhao, X. 2017. "Applying sensor-based technology to improve construction safety management." *Sensors*, 17(8): 1841.

Zhang, L., Ekyalimpa, R., Hague, S., Werner, M., and AbouRizk, S. 2015. "Updating geological conditions using Bayes theorem and Markov chain. In *Proceedings of the 2015 Winter Simulation Conference (WSC),* 3367-3378. Piscataway, NJ: IEEE

5. CHAPTER 5: DATA SOLUTION TO IMPROVE PRELIMINARY RESOURCE PLANNING IN INDUSTRIAL CONSTRUCTION

5.1. INTRODUCTION

Preliminary resource planning is vital for the success of a project. Most construction projects are awarded following a period of competitive tendering. The time available to a general contractor to thoroughly plan after a bid is awarded is often limited, as they must mobilize a team to meet the project demand quickly (Loosemore et al. 2003). In addition to the tight timeline, the labor market in the construction industry is highly sensitive to wider economic activity: during an upturn in the economy, general contractors are competing with each other for skilled craft and professionals; conversely, during an economic downturn, they face challenges keeping skilled craft workers and professionals due to the curtail of investments (Statistics Canada 2011).

Beyond these common challenges, industrial construction projects face additional difficulties when it comes to preliminary resource planning. First, these projects are complex and involve multiple specialized trades, which require particular efforts to plan, schedule, and execute (Wu et al. 2014). Common trades involved in the completion of heavy industrial construction projects include boilermakers, carpenters, electricians, heavy equipment operators, insulation installers, ironworkers, millwrights, pipefitters and welders. Second, the industrial construction sector bears extra pressures stemming from market conditions and global resource pricing, resulting in the broad adoption of the fast-track project delivery method (Alberta Infrastructure 2001). Although it significantly reduces total project duration, this method leads to a severe overlap of the engineering phase and the construction phase (Williams 1995), thereby limiting the engineering information available at the preliminary trade resource planning stage.

Although challenging, preliminary resource planning should not be omitted or overlooked, as it provides a reference point that serves as a basis for monitoring, controlling, identifying risks, and developing corrective actions (Rosenau and Githens 2005). With complete engineering packages, resource planning, leveling, and optimizing can be effectively achieved by loading a critical path method network (Heon and Ei-Rayes 2011; Cheng et al. 2015; Menesi and Hegazy 2014; Markou et al. 2017). Without detailed engineering packages (such as the majority of the industrial construction projects that are carried out with fast-track methods), the preliminary resource plan predominantly relies on the experience and expertise of construction professionals. Commonly, the project manager/planner assumes one labor curve for each trade based on a similar historical project. Consequently, the plan is subjective and fails to reflect the true project design. Further, due to the disconnection from the design, the project team cannot systematically update this resource plan as more information becomes available. As a result, this plan is often outdated and can neither serve as a method of identifying risks nor as a quality reference. Eventually, this plan is replaced with a resource plan generated by loading a construction schedule network as more engineering is completed. The construction may have already been initiated, and a staffing issue (either overstaffed or understaffed) may already be pressing, which often is too late.

The industry-wide adoption of information technology and modularization provides opportunities for innovative, data-driven preliminary resource planning. Benefiting from the rapid advancement of information technology, an increasing amount of data have been generated and collected during past projects (e.g. Building Information Model (BIM), budget data, progress charts). Although tremendous insight can be found in these data, very little of it has been analyzed and transformed into useful information to improve decision-making processes such as resource planning (Dean 2014). Furthermore, modular construction has

been widely applied in heavy industrial construction projects, such as the oil and gas, petroleum, and chemical industries (Burke and Miller 1998). These modules become natural physical envelopes to divide the scope of work at a high level, which could be taken advantage of for preliminary resource planning purposes.

This research proposes a semi-supervised machine learning approach capable of achieving a logical, engineering-oriented, and design-driven preliminary resource plan at the modular level using project data that are available early in the life cycle of a project. Through the integration of unsupervised and supervised machine learning techniques, this research aims to 1) unveil the underlying similarities between modules according to its design elements (i.e. design data from BIM—data quality and level of details are discussed in the next section), resulting in a set of module clusters/classes per trade; 2) summarize critical labor resource requirement indices for each module class identified; and 3) develop classifiers to predict module class for similar future projects. This research advances the state-of-the-art by being the first method capable of 1) providing decision makers with a scientific and data-driven approach for preliminary resource planning at the modular level for each trade; 2) parsing incomplete BIM data to classify module types by trade for industrial construction projects; and 3) proposing a semi-supervised learning approach for mining construction data sets that are large, raw, and non-integrated.

The remainder of the chapter is organized as follows: why and how BIM and modularization can assist in preliminary resource planning is discussed, and a review of the existing research on preliminary resource planning that identifies the research gaps and introduces the data-driven approach to address it is then presented. The research methodology section details the stages, techniques, and algorithms involved in the proposed framework. Lastly, an industrial case study is presented to demonstrate each stage of the proposed method and the results.

5.2. BIM IN PRELIMINARY RESOURCE PLANNING

Among one of the earliest types of projects to implement BIM, industrial construction projects have benefited tremendously from these models (Ali and Mohamed 2017). Relying on BIM to further improve the efficiency of industrial construction processes, however, has been challenging. Several research studies have examined, evaluated, and documented the difficulties limiting the use of BIM beyond the design/engineering stage (Guerra and Leite 2020, Leite et al. 2010, Arshad et al. 2019, Hamdi and Leite 2013, Alwash et al. 2017, Olatunji 2015, Ashcraft 2008). First, developing BIM is a costly and multi-organizational endeavor (Eastman et al. 2009, Solihin et al. 2017, Solihin and Eastman 2015). As such, information loss, inconsistency, missed interoperability, and redundancy are inevitable during the transfer and merging of data between stakeholders (Zhou et al. 2019). This is especially prevalent in industrial construction projects due to their large scale and complex processes.

Furthermore, the lack of industry-wide standards and legal status—not perceived as a contractual deliverable—causes great contractual risks and legal concerns (Arshad et al. 2019, Ashcraft 2008, Olatunji 2015). Olatunji (2015) demonstrated potential legal disputes arising from BIM use with respect to model integrity, storage security, intellectual property, and project changes. Even interoperability between different software chosen by various stakeholder can pose potential risk for disputes (Hamdi and Leite 2013). Consequently, BIM (or 3D models) have been excluded from contracts as a source of deliverables, and these models are often perceived as needless by downstream stakeholders, such as general contractors and subcontractors (Guerra and Leite, 2020)—a primary reason why general contractors are often issued incomplete, skimmed, and out-of-date models (Davies et al. 2017)

The BIM received in industrial construction is, most often, a rough 3D model of only an approximate physical representation of the structure that is missing critical construction details and other functional characteristics. The level of detail (LOD) of BIM in these industrial projects is typically approximate geometry (or LOD 200) or somewhere between conceptual (or LOD 100) and precise geometry (or LOD 300) depending on the discipline (Leite et al. 2010, Trimble Navigation Limited 2013). The maturity level of information in these 3D models is primarily for the conceptual and engineering stage and is insufficient to support construction or project management activities. As illustrated in Figure 5-1a, the properties of the indicated elbow are either missing or meaningless. Obtaining the dimensions of elements from such limited information is difficult. Solutions, such as a customized API attached to Navisworks (Han et al. 2017), to extract rough coordinates have been developed and adopted industry (Figure 5-1a). Moreover, due to the fast-track feature, BIM models for industrial projects are often developed in tandem with the construction period, and contractors are issued updated BIM models periodically, with detailed content replacing placeholders over time (Azhar 2011).

Modularization has been defined by the Construction Industry Institute (2014) as "the large-scale transfer of stick-build construction effort from the jobsite to one or more local or distant fabrication shops/yards, to exploit any strategic advantages. A module is a portion of plant fully fabricated, assembled, and tested away from the final site placement, in so far as is practical." In a modularized industrial construction project, practitioners often categorize modules based on its primary characteristics, such as a pipe rack module, electrical module, or equipment module. Figure 5-1b captures a typical BIM representation of a pipe rack module from an oil sands industrial project.

This subjective categorization captures the design information at a macro level, which reveals the design feature for the major trade involved in this specific type of module. To plan preliminary resources for each trade, however, the design features must be parsed at a micro-level. For instance, the scope of work for ironworkers may be similar between a pipe rack module and an equipment module. Despite the low level of detail, BIM naturally contains geographic information through its rough physical representation of the structure (such as location coordinates, length of a pipe, and the number of staircases). It provides opportunities for machine learning algorithms to discover the design characteristics (i.e. similarities or dissimilarities) between modules for every trade at a micro-level.

(a) (b)

Figure 5-1 3D Model of Pipe Elbow (a); and a Typical Pipe Rack Module (b)

In a fast-tracked modularized industrial construction project, project information, such as the incomplete BIM, module list, and module lifting schedule, commonly become available early in the life cycle of a project, as the majority of onsite construction activities are successors to the completion of the module, including the transportation and lifting of these modules onto their foundation and supporting structures. With such data available, preliminary design information extracted from BIM can be classified at a modular level per discipline. Loading the labor resource requirements learned for a certain type of modular design to the module lifting schedule can provide a solid foundation for a preliminary resource plan.

5.3. PREVIOUS RESEARCH

A critical activity in front end planning—a preliminary labor plan—identifies the labor resource requirements for a project, analyzing the source of labor, adequacy of supply, and addressing potential resource issues (George et al. 2008). While previous studies (CII 1995; CII 2006; CII 2009; Hwang and Ho, 2011) have confirmed the benefits of preliminary resource planning, including assisting the project team to appropriately dedicate resources, the aforementioned research studied the topic from a front-end planning perspective, which involves many activities beyond preliminary resource planning. Furthermore, these studies were survey-based or qualitative, and did not yield a methodology capable of addressing the challenges associated with preliminary resource planning.

The potential for using BIM for enhancing resource planning has been explored. In a medium-sized building construction setting, Babic and collaborators (2010) demonstrated the use of BIM as a channel to connect a pre-existing enterprise resource system with design information, such as AutoCAD drawings. Although similar to residential/commercial construction sectors, the industrial construction sector has unique challenges regarding BIM adoption. As mentioned previously, incomplete and unstandardized BIM data with a low level of detail severely limits its direct use for project management purposes.

Data-driven applications developed with machine learning algorithms have been extensively used in the construction management industry to assist the discovery of hidden patterns and relationships embedded in large, raw (i.e. incomplete, messy, and non-integrated) data sets, thus providing critical decision support to practitioners (Zhang et al. 2019; Zhong et al, 2019; Sarkar et al. 2020; Zou et al. 2017; Ali and Mohamed 2017). Unsupervised learning—clustering—aims to discover unknown, yet "natural" groups of objects (Jain et al. 1999; Dy

and Brodley 2004). Supervised learning—classification—learns the structure and content of a given data set that has been partitioned into groups (also referred to as classes), and yields a model (e.g. classifier) that predicts the labels of the classes for the unseen data set with known predictor features and unknown class labels (Aggarwal 2015). The territory between supervised and unsupervised learning is called semi-supervised learning (Witten et al. 2016), which improves the accuracy of supervised learning by exploiting information in unlabeled data. As such, semi-supervised learning algorithms have been successfully applied to derive solutions to problems, where the input is a fully or partially unlabelled large dataset, with a goal of developing a predictive model (Ferraretti et al. 2012).

A semi-supervised learning approach is adopted in this research for three primary reasons. First, BIM contains a large amount of unlabelled design data, and manually labeling these data is impractical and introduces biases; moreover, the underlying data structure of each module for each trade is unclear. As such, when starting with an unsupervised learning approach, design similarity (quantity summary of key design elements) can be learned between modules for each trade at a micro-level. Second, resource requirements follow design, where a more complex design leads to higher labor-hour. With the "natural" grouping of modules, resource requirements can be summarized for each module cluster for each trade. Third, BIM data are commonly available at the design stage before the release of the complete engineering package with periodic updates. With supervised learning algorithms, classifiers can be trained to predict module classes following the latest BIM release, thereby yielding an updated resource plan.

5.4. METHODOLOGY

As illustrated in Figure 5-2, the proposed framework involves four main stages – data pre-processing, unsupervised learning, summarize resource planning indices, and supervised learning, this will produce the final product—data-driven preliminary resource plan. Each stage is further discussed in the following subsections.

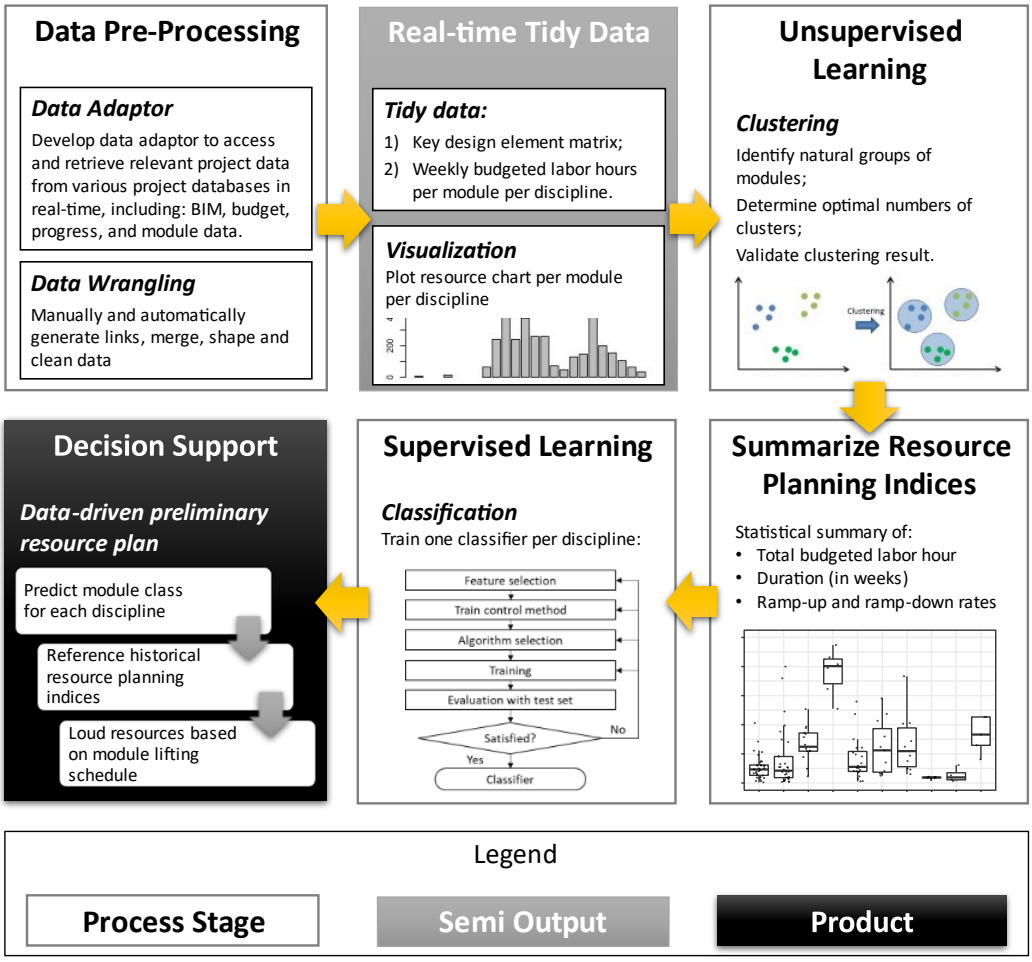


Figure 5-2 Proposed framework

5.4.1. Data Pre-Processing

As construction data are highly fragmented and often low in quality, data pre-processing is an often crucial initial step for any data-driven decision support system in construction management.

Raw data used in this research may include budget, progress, design, module, and module lifting schedules, each stored in separate databases. A construction data adaptor can be used to retrieve relevant data (tables) stored in separate locations and merge them into a central repository that can be further cleaned and analyzed without disturbing original data sets.

Although certain algorithms can be applied to noisy, inconsistent data (Han et al. 2011), data wrangling is critical for ensuring robust and reliable results from a majority of machine learning algorithms. Once the data has been merged, extensive data wrangling tasks may still be required to further aggregate and reshape the data into a tidy format (where each variable is a column and each observation is a row) (Wickham 2014)—especially for industrial construction data, where they are often developed and owned by different organizations/stakeholders, resulting in drastically different structures or naming conventions. For example, design data (i.e. BIM) are usually generated by the engineering team, module-related data are provided by module assembly facilities, while budget, progress, and module lift schedules are created by contractors. The differences in database structures and naming conventions prevent these data from linking or merging freely. A series of manual and automated data wrangling steps, including filtering, sub-sectioning, identifying common key attributes (Wu et al. 2020, Li et al. 2019, Wu and AbouRizk 2020), must be performed to create links.

5.4.2. Real-time Tidy Data

Two tidy tables are required as input into the following steps within the proposed framework. One table summarizes the key design by modules for each trade to a matrix. The other table groups weekly budgeted labor hours by module and by trade. The latter are then plotted as resource charts for visualization. During the merging and visualization process, data should be further examined, cleaned, and outliers identified. The data are then ready for input into the following machine learning stages.

5.4.3. Unsupervised Learning

Unsupervised learning, or clustering, is then performed to identify the similarities (or dissimilarities) between modules based on their key design features for each trade. A selection of methods can be used to analyze and validate clustering results: the multivariate analysis of variance (MANOVA), which tests the relevance between the cluster results and the variates; principal components analysis (PCA), which reduces the dimension of the data set thus allowing the cluster results to be plotted on a coordinate system with the first two principal components (PCs); use of external knowledge from experts (Rendon et al. 2011); and/or the use of internal data indices, such as Bayesian information criteria (Raftery 1986), to measure the similarity between different data points (Barbara et. al 2002; Santos and Morais 2013; Cleary and Trigg 1995).

5.4.4. Summarize Resource Planning Indices

Following clustering, a set of four resource planning indices are proposed for each discipline at the modular level, namely total budgeted labor-hours, duration, ramp-up rate, and ramp-down rate, as they capture the key elements of a resource plan. Then, a statistical summary is gathered for these four indices on each module cluster for every trade. The resulting

resource indices summary presents the typical resource requirements for each module class, thus linking the preliminary resource plan to the actual design of the structure.

5.4.5. Supervised Learning

With the confirmed clustering results (i.e. the classes), one classifier for each trade is developed. Supervised learning follows the loop described in Figure 5-2. Feature selection (such as Pearson correlation coefficient matrix and PCA), as the first step, investigates the degree of association, reduces the attribute dimension, and increases classification success (Abdi and Williams 2010).

Practitioners may choose any available supervised learning algorithms. This study recommends those listed in Table 5-1, as they are widely used for investigating construction problems and are useful for modeling complex relationships. The development of this set of classifiers bridges future project design to their typical resource requirements learned from historical projects. Together with the resource indices summary, this framework achieves the proposed data-driven preliminary resource plan as outlined in Figure 5-2.

5.4.6. Decision Support: Data-Driven Preliminary Resource Plan

The development of this set of classifiers (one classifier per trade) bridges future project design to their typical resource requirements learned from historical projects. Together with the resource indices summary, this framework achieves the proposed data-driven preliminary resource plan as outlined in Figure 5-2. For disciplines with insufficient data, estimation using the percentage method based on relevant disciplines can be used (such as to estimate hydrotest and insulation data based on piping data).

Table 5-1 Proposed classification algorithms

Classification Algorithm	Type	Tuning Parameters
Artificial Neural Networks (ANN)	Perceptron-based techniques	Size (Hidden Units); Decay (Weight Decay)
Naive Bayes	Statistical learning algorithms	Laplace (Laplace Correction); Usekernel (Distribution Type); Adjust (Bandwidth Adjustment)
K-nearest Neighbors (KNN)	Instance-based learning	K (#Neighbors)
Support Vector Machines (SVM)	Support Vector Machines	Sigma; Tau (Regularization Parameter)
Random Forest	Decision trees	NA

5.5. CASE STUDY

The functionality of the proposed research framework is demonstrated following its analysis of the historical data of an industrial construction project that lasted two years. The project is a typical oil sand secondary extraction project located in northern Alberta, Canada. This type of project accounts for a large portion of our industrial partner's business portfolio. In this case study, *R* (R Core Team 2019), a free programming software environment for statistical computing and graphics, was used to perform all functionalities of the proposed methodology.

5.5.1. Data Sources

In this case study, the data required were stored in three separate databases—BIM, module lift schedule, and a progress database. The BIM model was generated by the engineering team; the progress database, which captured the work breakdown structure and project progress (budget hours earned per week), was developed in-house by the general contractor (our industrial partner) as a relational database; and the module lift schedule was co-developed between the general contractor and the module assembly facility. The practical

challenge of summarizing labor-hours by modules in this case study was caused by the disconnection between the geographic information of the project and budget information, as they were stored in two different databases with no direct link. The BIM contained geographic information for all design elements, making it relatively easy to group design elements into a modular level based on the boundaries defined through the module lift schedule. However, the progress databases followed a traditional work breakdown structure, aimed at most effectively performing and tracking the onsite construction activities, with each work package usually containing one to two weeks of labor hours for one crew. This tree-like structure usually broke down the scope of work into construction areas, disciplines, work packages, activities, and steps and did not follow the physical breakdown of modules.

Data adaptor was developed in *R* through package *RODBC* (Ripley and Lapsley 2020) as a repository to retrieve useful information from all databases to a central location for further data wrangling. Notably, this data adaptor can be used in future projects to connect to the desired databases in a real-time manner. The weekly progress table queried from the progress database (Figure 5-3) contained the following information: discipline, work package, activity, and weekly records of labor-hours earned for each activity. The component list table extracted from the BIM database (Figure 5-3) contained the coordinates of each component (the smallest modeling item at geometry level); component name, which is a long string that contains most of the design-related information (such as "SUPPORT 12 of BRANCH/BFW-41039-01-00M56/B1"); and discipline. Additionally, a list of modules with the module name, project ID, and its geographic coordinates was extracted from the module lift schedule database (Figure 5-3).

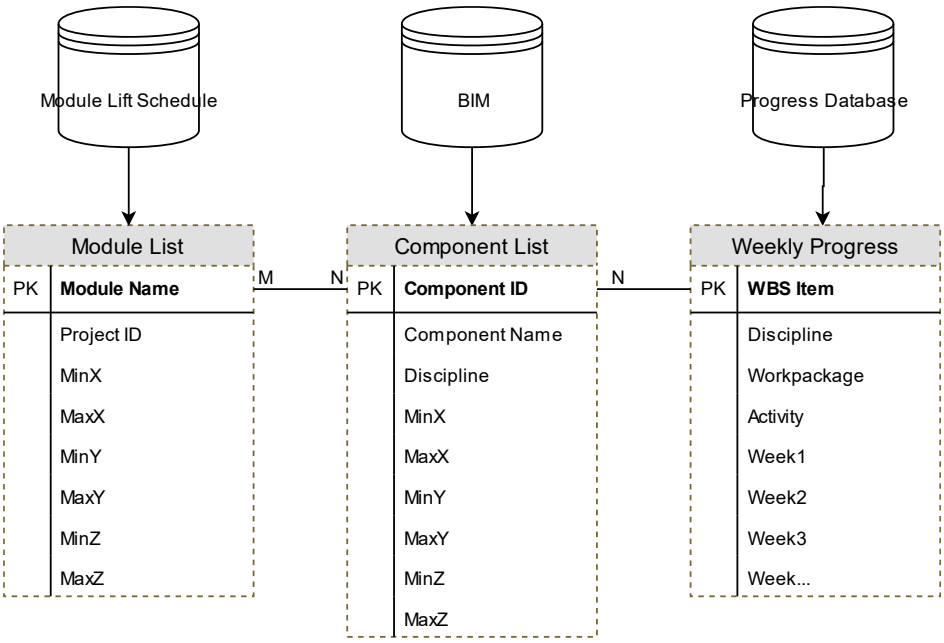


Figure 5-3 Entity-relationship diagram of module list table, component table, and progress table

Although data from different databases were gathered to a central location through the data adaptor, further efforts were required to link the data (as shown in Figure 5-3) and aggregate information into a tidy format for further analysis. The many-to-many relationship between module list and the component table was generated by comparing each component's coordinates with those of the module. The generation of a one-to-many relationship between the progress table and component list began with the manual parsing of data and patterns on both tables. For instance, a majority (3,709 out of 4,766 entries) of the piping progress activities began with "ISO-MCP-##-." The pattern was directly followed by a serial identification code combined with letters, numbers, and symbols such as "PW-40005-01-01A03." In the meantime, the piping serial identification code (e.g. "PW-40005-01-01A03") prevailingly existed (105,503 out of 105,716) in the component names column. The data pre-processing step was a joint exercise involving both the research team and our industrial partner and consisted of multiple rounds of manual verification for ensuring an accurate and

optimal result. Any progress activities that could not find a link to BIM elements were removed, as without location information, these activities could not be summarized into at a modular level.

Table 5-2 lists the results of the data aggregation for the major disciplines. For illustrative purposes, as well as the limitation of the data set, only the piping discipline is presented here. However, the proposed framework can be generalized for any discipline given the availability of the data. This is a key contribution of this research, as one module could belong to different clusters/classes given different disciplines. As a result, the proposed methodology tailors the resource requirement of each discipline for its specific design. For disciplines with insufficient data, estimation using the percentage method based on relevant disciplines can be used (such as to estimate hydrotest and insulation data based on piping data).

Table 5-2 Result of weekly progress table linked to the modular level

Discipline	Number of activities allocated to modules / total number of activities	Labor-hours of activities allocated to modules / total labor-hours
Electrical	2,272 / 11,765	11,765 / 268,808 labor-hour
Piping	3,629 / 4,766	307,727 / 368,517 labor -hour
Mechanical	167 / 358	61,207 / 99,781 labor -hour

5.5.2. Hierarchical clustering

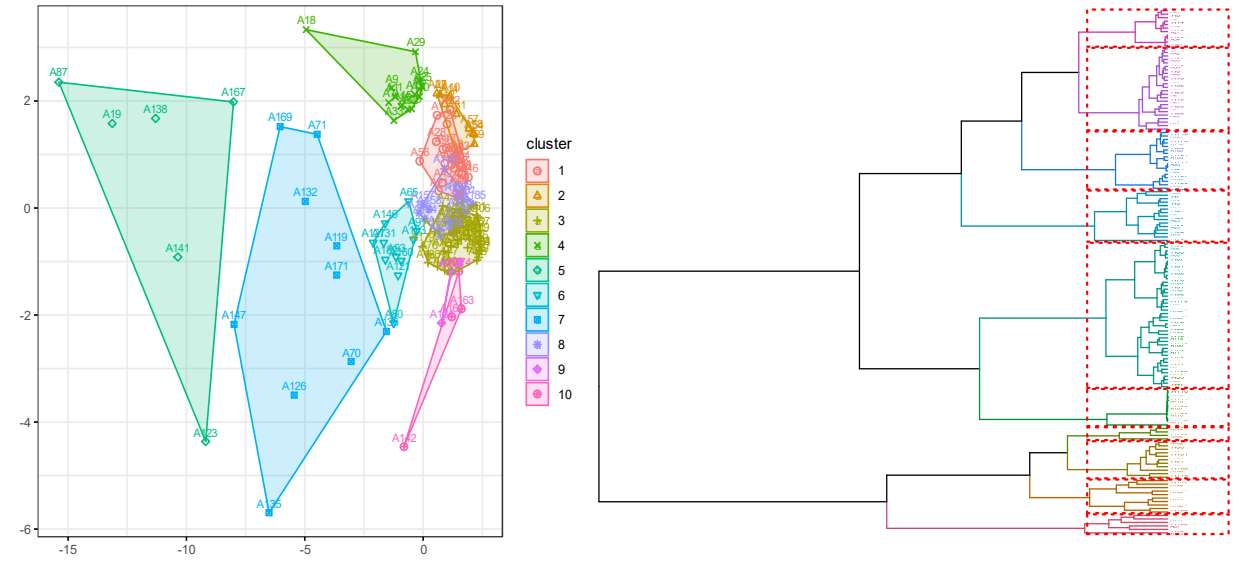
In preparation for the clustering experiment, a list of key design elements (cap, closure, coupling, elbow, blind, flange, gasket, instrument, reducer, support, tee, tube, valve, weld) were extracted from the component name column in the component list table. The count of each design element in each module was summarized into a matrix. Together with the elevation of the module, this matrix served as the input for the clustering process.

Followed by standardization and calculation of the Euclidean distance, the hierarchical clustering algorithm was applied to the input matrix. Hierarchical clustering, as an alternative approach to k-means clustering, was selected in the case study because: 1) when compared to k-means clustering, it does not require a pre-determined number of clusters; 2) it yields a tree-structured representation of the data, which is referred to as a dendrogram; and 3) from the dendrogram, a partition can be defined through a horizontal cut at a specified level that is defined by the user. The similarity between clusters is calculated using the ward method, as it is less susceptible to noise or outliers. The clustering result is presented as a dendrogram (Figure 5-4b)—a tree structure structured diagram that records the sequence of the merge.

To assist in choosing the optimal number of clusters, the elbow method, silhouette method, and gap statistic method were used to provide further insights into the data set regarding natural separations. As shown in Figure 5-5a, with the maximum number of clusters set as 12: no obvious "elbow" (location of bend, which is an indicator of optimal number) is observed from 1 to 12 clusters; the silhouette method (Figure 5-5b) suggests 2 clusters; and the gap statistic method suggests 12 clusters. Both gap statistics and elbow methods agree that more clusters result in a better description of the data. Finally, experts were invited to confirm a manageable number of clusters for practice, and 10 clusters were chosen in this case.

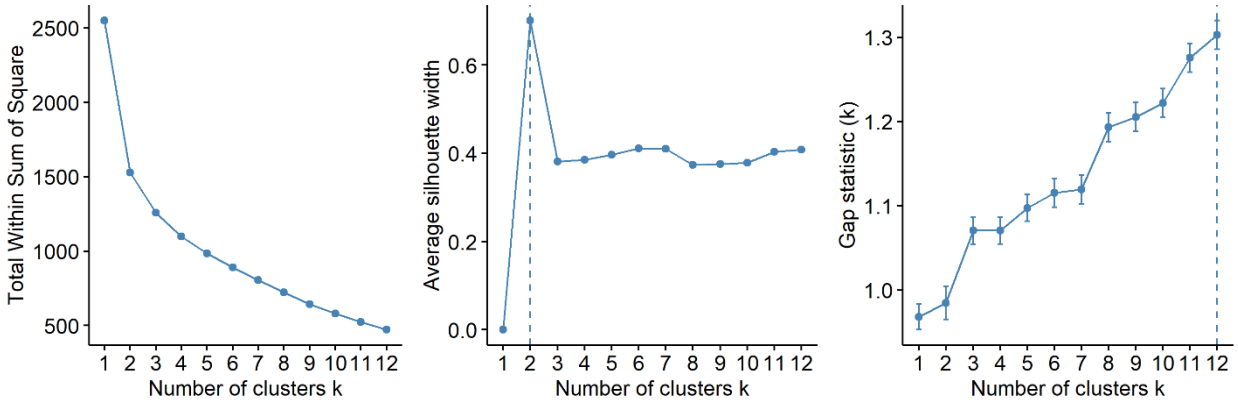
With the chosen 10 clusters, goodness of the cluster results was tested using MANOVA, visually examined via plotting with PCA components (Figure 5-4a), and validated by experts. The MANOVA result presented a strong correlation between the variables and the cluster results, as a very small (2.2e-16) p-value was observed. The result of PCA (Figure 5-6b) indicated that PC1 explained half of the variability in the data set and, together with PC2, explained about 60% of the variability in the dataset. With the reduced dimensions, the

clustering result was able to plot in a coordinate system by the first two principal components (Figure 5-4a). Lastly, the clustering result was reviewed and validated by two experts (as demonstrated in Table 5-3), who have extensive site experience, and both working on this project.



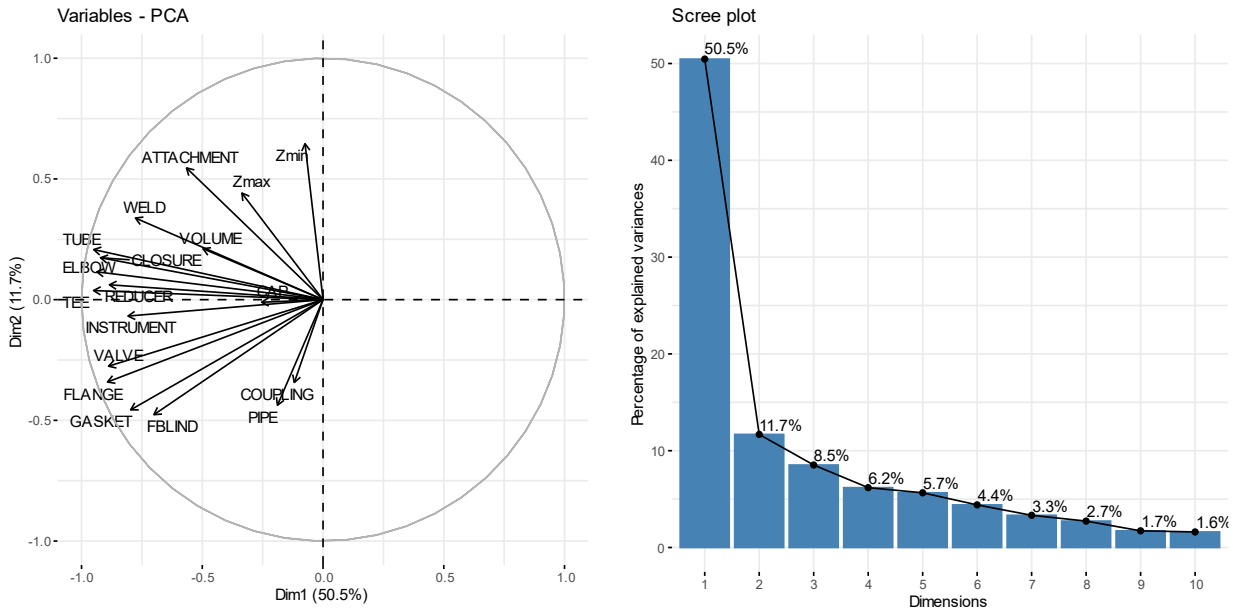
(a) **(b)**

Figure 5-4 Cluster result: (a) scatter plot on the first two PCs; (b) the dendrogram



(a) **(b)** **(c)**

Figure 5-5 Optimal number of clusters analysis: (a) elbow method, (b) silhouette method, and (c) gap statistic



(a) **(b)**

Figure 5-6 PCA result: (a) graph of variables, (b) scree plot

Table 5-3 Expert validation of clusters

Clusters	Expert Validation
C1	Module – Primarily cable tray, little pipe
C2	Piperack Module with – higher density piping
C3	Piperack Module and/or its substructure, minor piping, ground level
C4	Larger areas or Buildings with minimal piping
C5	Equipment areas with minimal piping
C6	Areas with large tanks
C7	Pre-packaged skids
C8	Equipment modules
C9	Equipment areas with piping
C10	Fin Fan Coolers

5.5.3. Summary of the resource planning indices

The statistical summary was gathered for the following four indices: total budgeted labor-hours, duration, ramp-up rate, and ramp-down rate for each cluster. As shown in **Error! R eference source not found.**—a randomly selected resource chart—the resource chart can be replicated with these four resource indices. To present the ramp-up and ramp-down rate, each resource chart of a module was fitted using the four-parameter generalized beta distribution. As a flexible distribution, the generalized beta distribution has been widely used in construction for representing resources, duration, cost, schedule, risk and more (Lu and AbouRizk 2000; Lu 2003; Poshdar et al. 2018; Zayed and Halpin 2001; Inyim et al. 2016; Sonmez 2005; Wang et al. 2002). The shape parameters of a generalized beta distribution represent the shape of the resource chart. Together with Total Labor Hour and Duration, these four resource indices are summarized for each module cluster in Table 5-4 and Figure 5-8.

With a closer inspection of the resource summary, trends were revealed. First, certain module cluster/class had a higher priority in demanding the resources. For instance, cluster 4 (C4) had more labor hours than the other clusters. The average total labor hour of C4 was over twice of the average for any other cluster, while the average duration (in weeks) of most of the clusters was between 35 weeks to 45 weeks, including C4. Second, trends, in terms of ramp-up and ramp-down, varied between clusters. For instance, C1, C2, and C10 had similar ramp-up and ramp-down rates (similar values of shape parameters a and b), which indicated that resource charts had a fairly symmetrical shape; for C4, shape parameter a was larger than shape parameter b, indicating a slower ramp-up but sharper ramp-down rate; whereas in C3, C7, C8, and C9, a quicker ramp-up, yet slower ramp-down, rate was observed. To

summarize, different design leads to different resource requirements, and critical resource trends can be learned from historical project data.

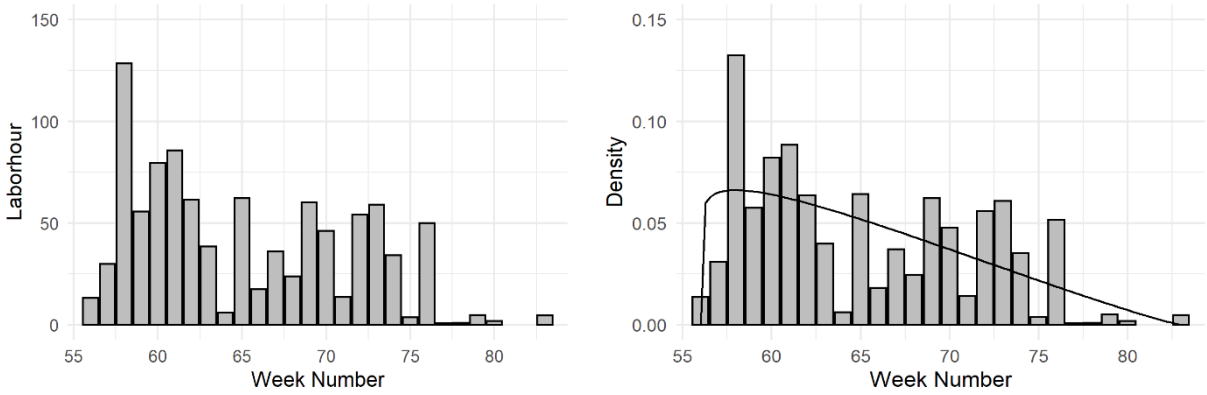


Figure 5-7 Historical resource chart for a randomly selected module

Table 5-4 Statistical summary of resource indices for each cluster

Resource Indices	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Total Labor Hour										
Min	19	37	228	2,550	55	250	303	101	65	811
Q1	262	182	1,091	3,589	396	371	558	179	106	1,293
Med	465	423	1,245	4,079	554	1,120	1,094	189	191	1,663
Q3	616	910	1,716	4,528	1,084	1,867	1,891	195	350	2,267
Max	2,058	3,999	2,536	10,196	2,796	2,909	3,665	217	612	2,275
Duration in week										
Min	17	10	30	24	16	12	10	15	25	22
Q1	30.5	21	38.75	36.5	23.5	24.5	25.5	17	31	30
Med	42.5	33	42	51	35.5	33	34.5	20	34	42
Q3	48.75	48.5	68	62.5	41.75	42	44.25	20	35	46
Max	70	68	73	72	50	55	58	20	35	64
Shape Parameter a										
Min	0.07	0.40	0.64	0.55	0.14	0.32	0.50	0.17	0.08	0.76
Q1	0.71	0.97	1.15	2.22	0.39	0.64	0.84	0.50	0.45	1.21
Med	0.99	1.25	1.46	3.20	1.31	1.15	2.14	2.09	0.67	2.12
Q3	1.81	1.98	2.17	4.37	5.13	3.83	3.06	3.09	0.80	2.88
Max	5.41	5.47	5.90	10.00	9.17	11.16	6.47	4.05	0.92	6.00
Shape Parameter b										
Min	0.05	0.22	1.24	1.17	0.04	0.67	0.94	0.30	0.39	1.16
Q1	0.89	1.24	2.23	1.97	0.97	1.49	1.92	2.22	0.61	1.46
Med	1.30	1.70	2.38	2.71	2.26	2.11	2.60	3.87	2.32	1.82
Q3	2.19	3.10	2.85	2.88	3.42	3.76	3.44	6.07	4.05	2.27
Max	6.85	11.94	4.63	8.49	5.41	6.58	5.62	7.35	4.31	4.01

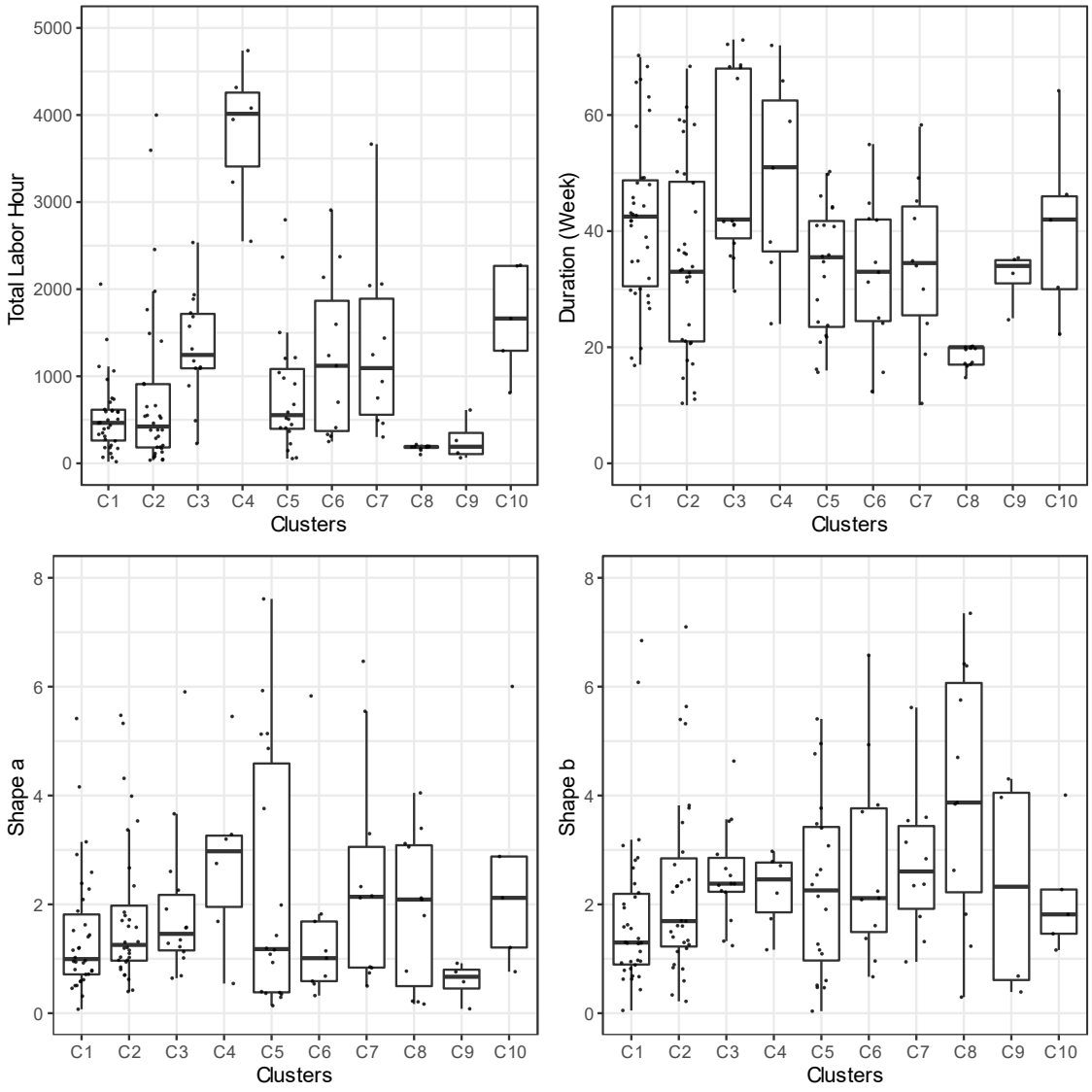


Figure 5-8 Boxplot with jittered point plot of the four resource indices

5.5.4. Classification

With resource plan indices summarized and trends discovered based on the confirmed cluster result, a classifier for piping discipline was built to predict the module type for future projects. With this classifier, the planner can effectively identify the type of the module based on the design elements, thus reference the historical resource indices (Table 5-4) for resource planning purposes.

Feature selection significantly reduced the dimensions, decreasing the intensity of computing efforts by removing irrelevant/redundant attributes to the target concept (Dash and Liu, 1997). Additionally, it explains the structure of the data regarding the target result and helps practitioners gain better insight into the data set.

In this case study, the Boruta algorithm, the automated feature selection (*fscaret* (Szlęk et al. 2013)) based on the *caret R* package (Kuhn 2008) modeling method, and the Pearson correlation coefficient were applied. After 15 iterations, the Boruta algorithm confirmed all variables as important (Figure 5-9). These results were consistent with those obtained using *fscaret* (Figure 5-10), where variables were ranked for models trained using each of the algorithms listed in Table 5-1. The 5-ranked variable list showed no consistent most or least important variables. To visualize the results, sequential indexes (with 1 as the most important variable and 18 as the least important) were assigned to the 5-ranked variable list. The sum of the sequential indexes for every variable were plotted as a pie chart. Notably, if one variable is consistently ranked as the most important among all 5 lists, its sequential index sum will be less than 0.1%. Conversely, if one variable is consistently ranked as the least important among all 5 lists, it will have a sequential index sum greater than 10%.

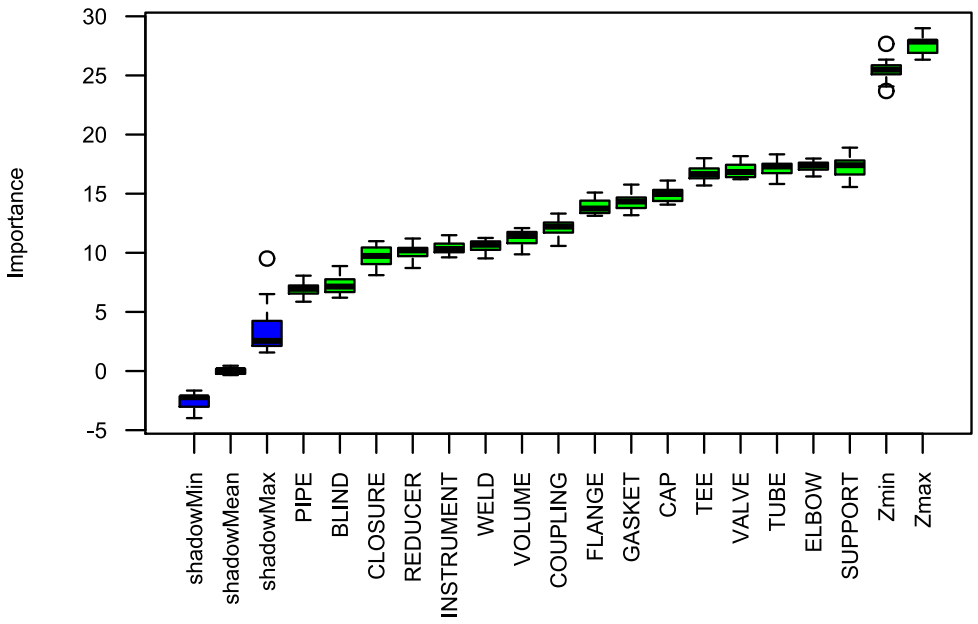


Figure 5-9 Boruta feature selection result

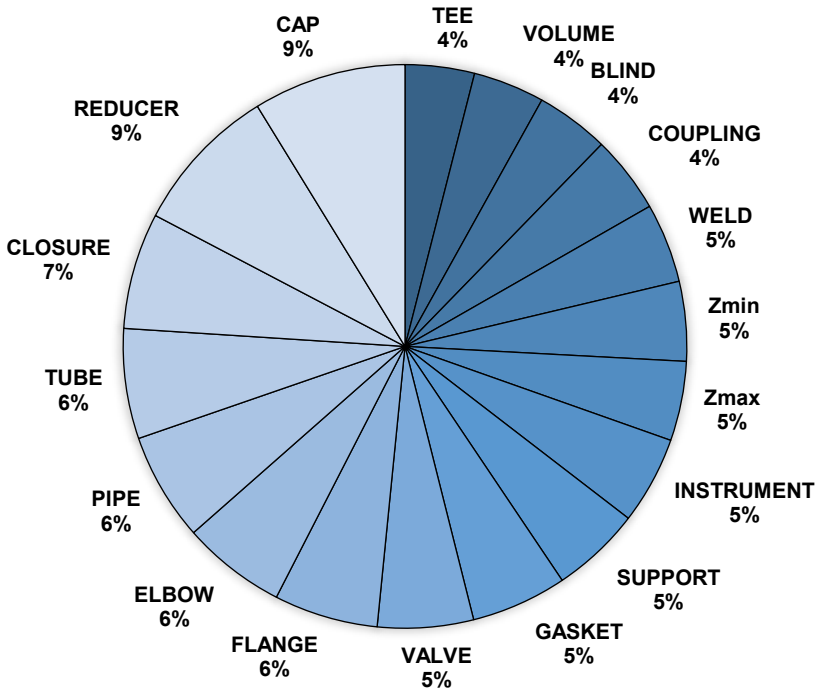


Figure 5-10 *fscaret* feature selection result

With all attributes and class labels (i.e. expert validated clusters), classifiers were developed using 10-fold cross-validation method. The average accuracy was shown in Table 5-5 for each classification algorithm. KNN model was chosen in this case study with the most stable performance, and the confusion matrix for the selected model was presented in Figure 5-11**Error! Reference source not found.**.

Table 5-5 Average accuracy for each classification algorithm

Classification Algorithm	Accuracy (average)
Artificial Neural Networks (ANN)	0.8941
Naive Bayes	0.9073
K-nearest neighbours (KNN)	0.9603
Support Vector Machines (SVM)	0.9139
Random Forest	0.8694

```
            Reference
Prediction    1    2    3    4    5    6    7    8    9   10
        1  23.2  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
        2   0.0 21.9  0.0  0.0  0.7  0.7  0.0  0.0  0.7  0.0
        3   0.0  0.0  9.3  1.3  0.0  0.0  0.0  0.7  0.0  0.0
        4   0.0  0.0  0.0  3.3  0.0  0.0  0.0  0.0  0.0  0.0
        5   0.0  0.0  0.0  0.0 13.2  0.0  0.0  0.0  0.0  0.0
        6   0.0  0.0  0.0  0.0  0.0  6.6  0.0  0.0  0.0  0.0
        7   0.0  0.0  0.0  0.0  0.0  0.0  6.6  0.0  0.0  0.0
        8   0.0  0.0  0.0  0.0  0.0  0.0  0.0  6.6  0.0  0.0
        9   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  2.0  0.0
        10  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  3.3

 Accuracy (average) : 0.9603
```

Figure 5-11 Confusion matrix for the chosen KNN classifier

5.5.5. Validation

The proposed methodology and the summarized preliminary resource indices were validated through a second project that is currently in the pre-construction planning phase. As this

point, no detailed work breakdown structure or schedule are available. The only data available for validation is the total budgeted labor-hours.

Of the 70 modules in the second project, 42 had a sufficient number of piping labor-hours required for validation purposes (i.e., more than 20 labor-hours). The chosen KNN classifier was applied to these 42 modules. Once classes were predicted for each module, total budgeted labor-hour predictions (Table 5-4), obtained from the analysis of the historical data previously performed, were assigned to each module. A comparison of the prediction obtained using the proposed approach versus the actual total labor-hours is presented in Figure 5-12. The scattered points represent actual budgeted total labor hour for each module, and the boxplot represents the statistical summary gathered from historical project data (same as top left chart of Figure 5-8).

The proposed approach predicted labor-hours of the second, unseen project dataset with reasonable accuracy. Over 90% of the actual data points fell within the predicted range, with only 4 points falling outside of the predicted range (circled in red in Figure 5-12). Notably, the actual total piping labor-hour was also within the prediction range.

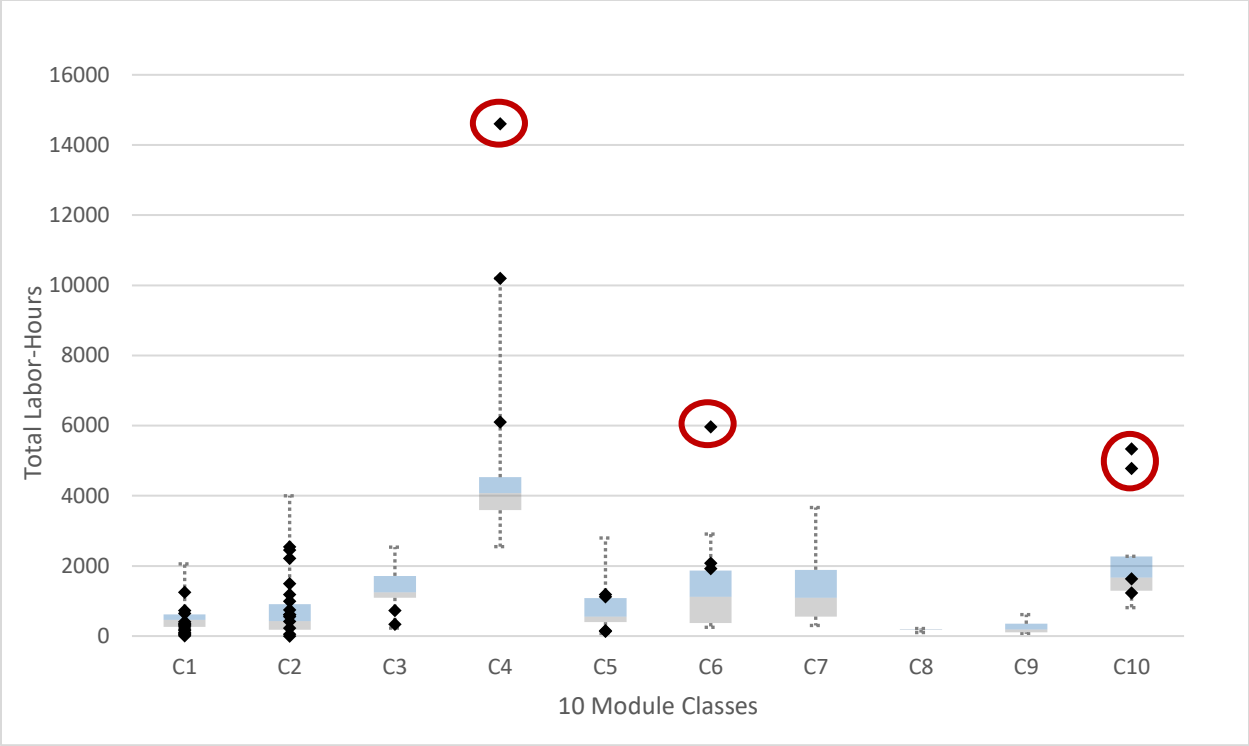


Figure 5-12 Plot of actual (scattered points) and predicted (box plots) total labor-hours

5.5.6. Synopsis

This case study demonstrated the practicality and effectiveness of the proposed methodology in parsing, understanding, and learning critical resource planning information from a large, unlabeled, and non-integrated historical construction project dataset. Specifically, multi-dimensional data sources were accessed through a data adaptor, resulting in a tidy table for input into various machine learning algorithms. This was followed by unsupervised learning and investigation of parameter space, which revealed the natural segregation of the dataset based on key module design elements. According to expert validated clustering result, resource requirements were summarized statistically into four indices. Lastly, a classifier was constructed for future projects in the prediction of the module type.

The classifier, together with the resource summary, can serve as an important decision-support tool for future projects. Upon the release of a BIM model for a new project, design elements can be quickly summarized and supplied to the classifier. Once module classes are predicted, associated resource requirements can be used for planning purposes. The proposed framework also developed a solid foundation to incorporate potential preliminary indirect labor hour planning (Wu et al. 2014). In addition, and equally as important, the outputs of the proposed approach can be used as inputs for other data-driven planning tools, such as optimization and simulation (Pereira et al. 2020, Li et al. 2019, Wu et al. 2018), for advanced decision support.

5.6. CONTRIBUTIONS AND FUTURE WORK

This research proposes a framework to effectively use incomplete BIM and other early-on available project data for preliminary resource planning purposes. The proposed method enables the practitioners a powerful insight into historical data through unsupervised learning. Then, through supervised learning, the learned information from unsupervised learning can be trained as a prediction model for future planning purposes. As demonstrated through the application of the framework to a case study of data from real industrial construction projects, the summarized resource planning indices, together with the classifiers, can provide practitioners with critical decision support for future project for high-level resource planning. Additionally, this research demonstrated how semi-supervised learning can effectively parse a large amount of unlabelled, raw construction data while, at the same time, yielding meaningful classifiers and alleviating the need for the manual labeling of raw data.

Through the case study, this research demonstrated that 1) regardless of the data origin, various types construction data can be accessed easily and in real-time through a data adaptor; 2) valuable information can be learned from incomplete BIM data through various machine learning algorithms to provide critical decision support for preliminary resource planning; 3) unsupervised learning can effectively process a large amount of construction data and provide critical insight, such as natural separation, based on design elements; 4) based on the results of unsupervised learning, supervised learning can further enhance the data mining process, providing a decision support system for future projects; and 5) semi-supervised learning significantly reduces labor-intensive processes associated with manual data labeling, thus increasing the efficiency with which a large amount of unlabeled data can be processed.

Although the benefits of the proposed method were demonstrated using data from actual projects, the method itself is not limited to any specific organization or data structure. The implementation of the data adaptor, unsupervised learning, and supervised learning will largely reduce the manual effort involved in machine learning for general contractors, in turn encouraging usage of historical construction data for future decision support.

In light of the contributions, this research should consider the following limitations. First, feature selection was conducted only at the supervised learning stage to reduce redundant or irrelevant features towards the class. Feature selection was not conducted at the unsupervised learning stage to provide as much relevant information as possible. Second, in the case study, experts were invited to the validation of unsupervised learning. The usage of expert knowledge to validate the unsupervised learning increases the explainability of the machine learning result, while unavoidably introduces a certain degree of subjective bias.

Third, the availability of the data limits the implementation of the proposed methodology to the data set of two projects.

5.7. ACKNOWLEDGMENTS

This work was generously supported by PCL Industrial Management Inc. and was funded by an NSERC Collaborative Research and Development Grant (CRDPJ 492657).

5.8. REFERENCES

Aggarwal, C. C. 2015. *Data mining: the textbook*. USA: Springer.

Abdi, H., and Williams, L. J. 2010. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics*. 2(4): 433-459.

Alberta Infrastructure. 2001. "Construction management an owner's guide to using the 'construction management' project delivery system on Alberta infrastructure funded building projects." Accessed June 15, 2020. https://www.alberta.ca/assets/documents/tr/tr-constmgmt.pdf

Ali, M., and Mohamed, Y. 2017. "A method for clustering unlabeled BIM objects using entropy and TF-IDF with RDF encoding." *Advanced Engineering Informatics*. 33: 154-163.

Azhar, S. 2011. "Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry." *Leadership and management in engineering*. 11(3): 241-252.

Babic, N. C., Podbreznik, P., and Rebolj, D. 2010. "Integrating resource production and construction using BIM." *Automation in Construction*. *19*(5): 539-543.

Barbara, D., Li, Y., and Couto, J. 2002, COOLCAT: an entropy-based algorithm for categorical clustering. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 582-589). ACM.

Burke, G. P., and Miller, R. C. 1998. "Modularization speeds construction." *Power engineering*. 102(1): 20-23.

Cheng, M. Y., Prayogo, D., and Tran, D. H. 2015. "Optimizing multiple-resources leveling in multiple projects using discrete symbiotic organisms search." *Journal of Computing in Civil Engineering*. *30*(3): 04015036

CII (Construction Industry Institute). 2006. *Data analysis in support of front end planning implementation*. RR213-11. Austin, TX: CII

CII (Construction Industry Institute). 1995. *Pre-project planning handbook*. SP39-2. Austin, TX: CII

CII (Construction Industry Institute). 2006. *Optimizing construction input in front end planning*. RR241-11. Austin, TX: CII

CII (Construction Industry Institute). 2014. *Industrial Modularization: Five Solution Elements*. RR283-2. Austin, TX: CII

CII (Construction Industry Institute). 2015. Innovative delivery of information to the crafts. RT-327-1. Austin, TX: CII.

Cleary, J. G., and Trigg, L. E. 1995. "An instance-based learner using an entropic distance measure." In Machine Learning Proceedings 1995(pp. 108-114). Morgan Kaufmann.

Dash, M., and Liu, H. 1997. "Feature selection for classification." *Intelligent data analysis*. 1(3): 131-156.

Davies, K., McMeel, D. J., and Wilkinson, S. 2017. "Making friends with Frankenstein: hybrid practice in BIM." *Engineering, construction and architectural management*. 24(1): 78-93.

Dean, J. (2014). *Big data, data mining, and machine learning: Value creation for business leaders and practitioners*. Hoboken, NJ: Wiley.

Dy, J. G., and Brodley, C. E. 2004. "Feature selection for unsupervised learning." *Journal of machine learning research*. 5(Aug): 845-889.

Eastman, C., Lee, J. M., Jeong, Y. S., and Lee, J. K. 2009. "Automatic rule-based checking of building designs." *Automation in construction*, *18*(8), 1011-1033.

Fan, H., and Li, H. 2013. "Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques." *Automation in construction*. 34: 85-91.

Ferraretti, D., Gamberoni, G., and Lamma, E. 2012. "Unsupervised and supervised learning in cascade for petroleum geology." *Expert Systems with Applications*. 39(10): 9504-9514.

George, R., Bell, L. C., and Edward Back, W. 2008. "Critical activities in the front-end planning process." *Journal of Management in Engineering*. *24*(2): 66-74.

Guerra, B. C., and Leite, F. 2020. "Bridging the Gap between Engineering and Construction 3D Models in Support of Advanced Work Packaging." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *12*(3), 04520029.

Han, J., Pei, J., and Kamber, M. 2011. *Data mining: concepts and techniques*. Waltham USA: Elsevier.

Heon J. D., and El-Rayes, K. 2011. "Multiobjective optimization of resource leveling and allocation during construction scheduling." *Journal of construction engineering and management*. *137*(12): 1080-1088.

Hwang, B. G., and Ho, J. W. 2011. "Front-end planning implementation in Singapore: Status, importance, and impact." *Journal of construction engineering and management*. *138*(4): 567-573.

Inyim, P., Zhu, Y., and Orabi, W. 2016. "Analysis of time, cost, and environmental impact relationships at the building-material level." *Journal of Management in Engineering*. 32(4): 04016005.

Jain, A.K., Murty, M. N., and Flynn, P. 1999. "Data clustering: A review." *ACM Computing Surveys*. 31(3): 264–323.

Kuhn, M. 2008. "Building Predictive Models in R Using the caret Package." *Journal of Statistical Software, 28(5)*, 1 - 26. http://dx.doi.org/10.18637/jss.v028.i05

Li, Y., Xu, S., Wu, L., AbouRizk, S., Tae, K., & Lei Z. (2019) "A generic simulation model for selecting fleet size in snow plowing operations" *Proceedings of the Winter Simulation*

Conference, National Harbor, MD, US. https://doi.org/10.1109/WSC40007.2019.9004954

Li, Z., Wu, L., and AbouRizk, S. 2019. *XiaomoLing/LongestCommonSubString: First Release of Longest Common SubString R Library* (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.4057067

Liao, C. W., and Perng, Y. H. 2008. "Data mining for occupational injuries in the Taiwan construction industry." *Safety science*. 46(7): 1091-1102.

Loosemore, M., Dainty, A., and Lingard, H. 2003. *Human resource management in construction projects: strategic and operational approaches*. UK: Taylor & Francis Group.

Lu, M., and AbouRizk, S. M. 2000. "Simplified CPM/PERT simulation model." *Journal of Construction Engineering and Management*. 126(3): 219-226.

Lu, M. 2003. "Simplified discrete-event simulation approach for construction simulation." *Journal of Construction Engineering and Management*. 129(5): 537-546.

Markou, C., Koulinas, G. K., and Vavatsikos, A. P. 2017. "Project resources scheduling and leveling using Multi-Attribute Decision Models: Models implementation and case study." *Expert Systems with Applications*. *77*: 160-169.

Menesi, W., and Hegazy, T. 2014. "Multimode resource-constrained scheduling and leveling for practical-size projects." *Journal of management in engineering*. *31*(6): 04014092.

Pereira, E., Ali, M., Wu, L., and Abourizk, S. 2020. "Distributed Simulation–Based Analytics Approach for Enhancing Safety Management Systems in Industrial Construction." *Journal of construction engineering and management*, *146*(1), 04019091.

Poshdar, M., González, V. A., Raftery, G. M., Orozco, F., and Cabrera-Guerrero, G. G. 2018. "A multi-objective probabilistic-based method to determine optimum allocation of time buffer in construction schedules." *Automation in Construction*. 92: 46-58.

R Core Team. 2019. "R: A language and environment for statistical computing." *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M. 2011. "Internal versus external cluster validation indexes." *International Journal of computers and communications*, *5*(1), 27-34.

Ripley B. and Lapsley M. 2020. *RODBC: ODBC Database Access*. R package version 1.3-17. https://CRAN.R-project.org/package=RODBC

Rosenau, M. D., and Githens, G. D. 2005. *Successful project management, 4th Ed*. New York USA: Wiley.

Solihin, W., Dimyadi, J., Lee, Y. C., Eastman, C., and Amor, R. 2017. "The critical role of accessible data for BIM-based automated rule checking systems." In Vol. 1 of *Proc., the joint conference on computing in construction (JC3),* 53-60.

Solihin, W., and Eastman, C. 2015. "Classification of rules for automated BIM rule checking development." *Automation in construction*, *53*, 69-82.

Santos, J. M., and Morais, F. 2013. "Evaluating entropic based clustering algorithms on biomedical data." *In Proc., 12th Mexican International Conference on Artificial Intelligence* 194-199. IEEE.

Sarkar, S., Pramanik, A., Maiti, J., and Reniers, G. 2020. "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data." *Safety science*, *125*, 104616.

Sonmez, R. 2005. "Review of conceptual cost modeling techniques." *AACE International Transactions*. ES71

Statistics Canada. 2011 "Canada Year Book, 2011" Accessed June 15,2020. https://www150.statcan.gc.ca/n1/pub/11-402-x/2011000/chap/construction/construction-eng.htm

Szlęk, J., Pacławski, A., Lau, R., Jachowicz, R., and Mendyk, A. 2013. "Heuristic modeling of macromolecule release from PLGA microspheres." *International journal of nanomedicine*, *8*, 4601. https://doi.org/10.2147/IJN.S53364

Wickham, H. 2014. "Tidy data." *Journal of Statistical Software*. 59(10): 1-23.

Williams, G. V. 1995. "Fast track pros and cons: Considerations for industrial projects." *Journal of management in engineering*. 11(5): 24-32.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. 2016. *Data Mining: Practical machine learning tools and techniques*. Cambridge USA: Elsevier.

Wang, L., Shen, W., Xie, H., Neelamkavil, J., and Pardasani, A. 2002. "Collaborative conceptual design—state of the art and future trends." *Computer-Aided Design*, 34(13): 981-996.

Wu, L., Ali, M., Pereira, E., and AbouRizk, S. 2018 "Linear regression and agent-based modeling approach for equipment market value prediction" *Proceedings of the 17th International Conference on Modeling and Applied Simulation*, Budapest, Hungary.

Wu, L., Mohamed, Y., Taghaddos, H., and Hermann, R. 2014. "Analyzing scaffolding needs for industrial construction sites using historical data." In *Construction Research Congress 2014: Construction in a Global Network* (pp. 1596-1605). https://doi.org/10.1061/9780784413517.163

Wu, L., and AbouRizk, S. 2020, *XiaomoLing/Detect3DRelation: First Release of the Detect 3D Relation function* (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.4058576

Wu, L., Li, Z., and AbouRizk, S. 2020. "Automation in Extraction and Sharing Information between BIM and Project Management Databases." *Proceedings of the International Conference on Construction and Real Estate Management (ICCREM) 2020:* 37-46, Stockholm, Sweden. https://doi.org/10.1061/9780784483237.005

Zayed, T. M., and Halpin, D. 2001. "Simulation of concrete batch plant production." *Journal of Construction Engineering and Management*. 127(2): 132-141

Zou, Y., Kiviniemi, A., and Jones, S. W. 2017. "Retrieving similar cases for construction project risk management using Natural Language Processing techniques." *Automation in Construction*, 80: 66-76.

Zhang, F., Fleyeh, H., Wang, X., and Lu, M. 2019. "Construction site accident analysis using text mining and natural language processing techniques." *Automation in Construction*, *99*, 238-248.

Zhong, B., Xing, X., Love, P., Wang, X., and Luo, H. 2019. "Convolutional neural network: Deep learning-based classification of building quality problems." *Advanced Engineering Informatics*, *40*, 46-57.

Zhou, Y. W., Hu, Z. Z., Lin, J. R., and Zhang, J. P. 2019. "A review on 3D spatial data analytics for building information models." *Archives of Computational Methods in Engineering*, 1-15.

6. CHAPTER 6: CONCLUSION

6.1. RESEARCH SUMMARY

This research examines the challenge of exploiting big data in construction and transforming into insights that can support critical project management decisions.

Chapter 2 develops a framework to effectively process raw, fragmented construction data to achieve meaningful decision support metrics for construction. This framework serves as a cornerstone for the following research components: within the proposed framework, the data adaptor provides real-time database access; the DP-based longest common substring algorithm and the interval-based 3D objects relationship detection algorithm automate the two common data pre-process tasks for identifying relationship types between databases; SQL functions integrate information stored from various databases into a tidy format; and lastly, various data mining techniques and simulation models effectively process the tidy data and produce meaningful decision-support metrics.

Both custom functions (i.e. DP-based longest common substring algorithm and the interval-based 3D objects relationship detection algorithm) were validated through randomly-generated artificial data sets and generalized as public *R* libraries. The practicality and feasibility of the framework have been demonstrated following its application to a mega-sized industrial construction project.

Chapter 3 proposes a Bayesian inference-based method for systematically updating any given univariate continuous probability distribution input model of simulations, as new observations become available, and implements an MCMC-based numerical approximation approach to provide solutions regardless of conjugacy. An illustrative case study was used to demonstrate the generalizability, feasibility, and functionality

of the proposed Bayesian inference together with the MCMC-based numerical method for updating simulation input models.

Chapter 4 proposes a numerical-based weighted geometric average method for effective fusion of information generated from diverse sources (both observational data and expert opinions) for stochastic simulation. A Monte Carlo study, as the "Proof of Concept," was developed to test the proposed method against the weighted arithmetic average method and the mixture density samples. An illustrative case study was organized to demonstrate the generalizability, feasibility, and functionality of the proposed method for aggregating subjective and objective information to update simulation input models in real time.

Chapter 5 develops a data solution for preliminary resource planning in industrial construction projects with incomplete, fragmented—yet early available—construction data. Expanding upon the proposed framework in **Chapter 2**, the proposed data-driven application deploys semi-supervised machine learning techniques to parse data, and gain insights from a large, unlabeled, and non-integrated historical construction data. The proposed data solution is applied to a case study, with one historical project and one currently "under planning" project.

6.2. CONTRIBUTIONS

Compared with science and technology, engineering turns methods, algorithms, and ideas into something tangible that has a visible impact on society and the daily lives of people. The contributions of this research largely reflect this nature by responding to the current challenge in the construction community—how to fully exploit the value of construction data for informed decision-making. This research identified three bottlenecks that cause blockages

in information flow and limit data-driven decision-support systems in construction management. By adapting multi-disciplinary methods, designing domain specific frameworks, and developing automated functions for common tasks, this research removes some barriers, improves data usage, extends the boundary of existing knowledge, and promotes data-driven applications, thus impacting the construction industry and society. Specific academic and industrial contributions are subsequently discussed.

6.2.1. Academic contributions

The academic contributions of this work are summarized as follows:

1) Proposing a novel framework for enhanced data-driven applications built upon fragmented construction data. This framework provides a universal solution for bridging raw, segmented construction data with various data analytics.
2) Abstracting construction data pre-processing problems and adapting algorithms from computing science and applied mathematics to automate and streamline the otherwise manual data pre-processing steps. Additionally, generalizing these two custom functions into public R libraries ("Chrisfufu/LongestCommonSubString" and "XiaomoLing/Detect3DRelation") for broader audience.
3) Proposing a numerical-based Bayesian inference method for systematically updating input model of simulations as new observations become available. Coupled with a Markov chain Monte Carlo-based random sampling method, the proposed method extends Bayesian inference to any given univariate continuous probability distribution regardless of conjugacy. Additionally, the proposed method has been found capable of (1) accurately approximating the underlying probability distribution, (2) reliably fusing information from diverse sources, including subjective judgment

(through choice of priors) and objective observations, and (3) exhibiting robustness and resilience in situations where data were noisy and imbued with uncertainties.

4) Proposing a Markov chain Monte Carlo-based weighted geometric average method to effectively fuse information generated from diverse sources for stochastic simulation inputs. The proposed method has been found capable of (1) effectively and efficiently updating input models given new sources of information, (2) accurately approximating the target probability distribution, (3) reliably fusing information from diverse sources, including subjective judgment and objective observations, and (4) being generalized and applied to combinations of any given univariate continuous probability distributions. Thus, this method has potential applications to the greater engineering and management community.
5) Developing a data solution for scientifically planning project resources before the completion of engineering. The proposed data-driven framework demonstrates how the combined utilization of supervised and unsupervised machine learning algorithms can effectively parse a large amount of unlabelled, raw construction data while, at the same time, yielding meaningful classifiers.

6.2.2. Industrial contributions

The industrial contributions of this work can be summarized as follows:

1) Development of two custom functions (DP-based longest common substring algorithm and interval-based 3D objects relationship detection algorithm) to automate the common data pre-processing steps. The generalization of these two custom functions (public *R* libraries—"Chrisfufu/LongestCommonSubString" and "XiaomoLing/Detect3DRelation") frees domain experts from the periodic manual manipulation of non-integrated construction data and allows them to spend their time

understanding the data to producing meaningful matrices/indices for critical decision support.

2) Development of a framework for enhanced data-driven applications built upon fragmented construction data. This framework works with the construction industry's existing stand-alone information systems without the need for costly alterations or investments in a new system. This framework significantly reduces manual data manipulation, improves data quality, and streamlines the processing of raw, segmented data into data solutions.
3) Developing and demonstrating a numerical-based Bayesian method for continued updating of simulation inputs when new observational data become available. This method enhances the practicality of simulations and extends its potential applications to beyond the planning stage of a construction project. As demonstrated through the illustrative case study, through defining the probability model and priors, the proposed numerical-based Bayesian inference is capable of incorporating subjective opinions, and reflecting the influence of such information in the final results.
4) Developing and demonstrating a numerical-based weighted geometric average method to address the practical challenges associated with modeling assorted data origins. This method allows domain experts to effectively integrate observational data with expert opinion for continued updates of simulation models when project conditions change, thus enhancing the model's practicality and predictability, especially for construction project management, where expert knowledge has a heavy influence on the decision-making process.
5) Developing and demonstrating a data solution for preliminary resource planning in industrial construction projects with incomplete, fragmented—yet early available—construction data. The resulting resource indices and module class predictor become

critical components for preliminary resource planning. The resource indices specifically, can be supplied to a simulation model yielding project resource charts. This application greatly enhances the traditional craft-oriented project resource planning practices, and pushes the state-of-the-art one step further towards achieving full scientific and data-driven decision support in construction.

6.3. RESEARCH LIMITATIONS

The findings and contributions of this research should be applied in consideration of the following limitations:

- Although the DP-based longest common substring algorithm is simple and effective, it sacrifices running time. A generalized suffix tree algorithm could significantly reduce running time through reduced computational complexity from $O(m \times n)$ to $O(m + n)$, given string lengths m and n.
- Among many repetitive manual data pre-processing activities, this research develops generalized solutions to automate two common tasks.
- The proposed framework provides effective solutions to bridge non-integrated construction raw data with data solutions; however, it does not address the root causes nor change the status of the fragmented information systems in the construction industry.
- Due to a lack of appropriate real project data, both case studies developed to demonstrate the input model updating methods (Chapter 3 and 4) use artificial data sets. Using randomly generated observations based on a known underlying distribution eliminates the process of testing the fit of the chosen model (i.e.

distribution). In both studies, the models (such as beta distribution) are chosen assuming they are good fit for the data.

- Although advancing the methodology of input modeling in real-time, both methods (Chapter 3 and 4) do not represent a complete, decision-support system on their own.
- While both input modeling methods address cases of univariate parametric continuous probability distributions, certain real-life construction data will not belong to this category.
- The proposed numerical-based geometric average method itself does not guarantee improved accuracy. Rather, it provides a solution for fusing information from various sources. The predictability of the model ultimately relies on the accuracy of all source information and the selection of the weights—which is a complex problem that is beyond the scope of this research.
- Although a quick comparison between geometric average and arithmetic average is briefly discussed in the Monte Carlo study, it does not provide a comprehensive overview of the topic and should not be used as a guide.
- The proposed data solution for enhanced preliminary resource planning (Chapter 5) was applied to two projects (trained on one and validated on the second); reliability and accuracy can be improved with more data sets.
- The case study of the data solution for enhanced preliminary resource planning (Chapter 5) should consider the following limitations: 1) a limited number of machine learning algorithms are deployed; 2) feature selection was not conducted at the unsupervised learning stage—only at the supervised learning stage; and 3) using expert validation for the unsupervised learning result, which has inherent benefits and challenges.

6.4. FUTURE WORK

Beyond extending the research to address the above-listed limitations, future research can be carried out in the following areas:

- Adapting the proposed framework for different construction practices aiming at automating manual processes and continuing the transformation from the craft-oriented culture to a data-driven one. In this process, continue to identify specific pain points of the construction industry in general, explore innovative methods from other disciplines (e.g. applied mathematics, applied statistics, and computer science), and propose solutions for advancing both the academic field and construction practices.
- Investigating various construction management processes (such as cost forecast, safety management, stakeholder management) and applying diverse simulation techniques to model the complex and dynamic construction ecosystem to achieve more reliable and accurate real-time decision support metrics for industry, while, at the same time, enhancing the simulation environment by increasing resilience, reliability, and adaptability.
- A plurality of the current construction decision support systems follow a linear pattern, while real-world decisions follow feedback loops. Future research efforts are needed to 1) identify potential construction practice feedback loops; 2) construct dynamic, data-driven models to simulate these construction feedback loops; 3) understand the implications of changes over time, especially the side effects (unexpected factors and implications) to avoid potential pitfalls in decision making; and 4) improve mental models of all decision-makers involved to improve construction practices in general.
- Exploring artificial intelligence-powered and/or machine learning algorithm-based simulation applications for the construction domain.

BIBLIOGRAPHY

Abdelmegid, M. A., González, V. A., Poshdar, M., O'Sullivan, M., Walker, C. G., and Ying, F. 2020. "Barriers to adopting simulation modelling in construction industry." *Automation in Construction*, *111*, 103046.

Abdelmegid, M. A., González, V. A., Naraghi, A. M., O'Sullivan, M., Walker, C. G., and Poshdar, M. 2017. "Towards a conceptual modeling framework for construction simulation." In *Proceedings of 2017 Winter Simulation Conference,* 2372-2383. Piscataway, NJ: IEEE.

Abdi, H., and Williams, L. J. 2010. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics*. 2(4): 433-459.

AbouRizk, S.M. 2018. "Simulation-based analytics: Advancing decision support in construction." Responsible Design and Delivery of the Constructed Project. Presented at the *Proceedings of the Second European and Mediterranean Structural Engineering and Construction Conference*. Beirut, Lebanon, July 23-28, 2018. ISEC Press: Fargo, ND.

AbouRizk, S. M., Halpin, D. W., and Wilson, J. R. 1991. "Visual interactive fitting of beta distributions." *Journal of Construction Engineering and Management*, 117(4), 589-605.

AbouRizk, S. M., 2010. "Role of simulation in construction engineering and management." *Journal of construction engineering and management*, 136(10): 1140-1153.

AbouRizk, S. M., Hague, S. A., and Ekyalimpa, R. 2016a. *Construction simulation: An introduction using Simphony*. University of Alberta, Edmonton, Canada.

AbouRizk, S. M., Hague, S., Ekyalimpa, R., and Newstead, S. 2016b. "Simphony: A next generation simulation modelling environment for the construction domain." *Journal of Simulation*, *10*(3): 207-215.

AbouRizk, S. M., and Halpin, D. W. 1992. "Statistical properties of construction duration data." *Journal of Construction Engineering and Management*, 118(3): 525-544.

AbouRizk, S. M., Halpin, D. W., and Wilson, J. R. 1991. "Visual interactive fitting of beta distributions." *Journal of Construction Engineering and Management*. 117(4): 589-605.

AbouRizk, S. M., Halpin, D. W., and Wilson, J. R. 1994. "Fitting beta distributions based on sample data." *Journal of Construction Engineering and Management*. 120(2): 288-305.

Adriaanse, A., Voordijk, H., and Dewulf, G. 2010. "Adoption and use of interorganizational ICT in a construction project." Journal of Construction Engineering and Management, 136(9): 1003-1014.

Agarwal, R., Chandrasekaran, S., and Sridhar, M. 2016. "Imagining construction's digital future." *McKinsey and Company*.

Aggarwal, C. C. 2015. *Data mining: the textbook*. USA: Springer.

Ahsanullah, M. 2017. *Characterizations of univariate continuous distributions*. Atlantis Press, Paris, France.

Ahuja, H. N., Dozzi, S. P., and Abourizk, S. M. 1994. *Project management: techniques in planning and controlling construction projects*. John Wiley & Sons.

Akhavian, R. 2015. "Data-driven simulation modeling of construction and infrastructure operations using process knowledge discovery." Ph.D. thesis, Orlando, Florida: University of Central Florida

Akhavian, R., and Behzadan, A. H. 2013. "Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining." *Journal of Construction Engineering and Management*, 139(11), 04013021.

Akinci, B., Boukamp, F., Gordon, C., Huber, D., Lyons, C., and Park, K. 2006. "A formalism for utilization of sensor systems and integrated project models for active construction quality control." *Automation in construction*. 15(2): 124-138.

Al Qady, M., and Kandil, A. 2013. "Document discourse for managing construction project documents." *Journal of Computing in Civil Engineering*, *27*(5), 466-475.

Alberta Infrastructure. 2001. "Construction management an owner's guide to using the 'construction management' project delivery system on Alberta infrastructure funded building projects." Accessed June 15, 2020. https://www.alberta.ca/assets/documents/tr/tr-constmgmt.pdf

Ali, M., and Mohamed, Y. 2017. "A method for clustering unlabeled BIM objects using entropy and TF-IDF with RDF encoding." *Advanced Engineering Informatics*. 33: 154-163.

Altaf, M. S., Bouferguene, A., Liu, H., Al-Hussein, M., and Yu, H. 2018. "Integrated production planning and control system for a panelized home prefabrication facility using simulation and RFID." *Automation in construction*, 85: 369-383.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. 2003. "An introduction to MCMC for machine learning." *Machine learning*. 50(1-2): 5-43.

Arshad, M. F., Thaheem, M. J., Nasir, A. R., and Malik, M. S. A. 2019. "Contractual risks of building information modeling: Toward a standardized legal framework for design-bid-build projects." *Journal of Construction Engineering and Management*, 145(4): 04019010. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001617

Ashcraft, H. W. 2008. "Building information modeling: A framework for collaboration." *Construction Lawyer* 28(3): 1–14. Accessed May 12, 2021, https://heinonline.org/HOL/Page?collection=journals&handle=hein.journals/conlaw28&id=126&men_tab=srchresults

Ayyub, B. M. 2001. *Elicitation of expert opinions for uncertainty and risks*. CRC press.

Azhar, S. 2011. "Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry." *Leadership and Management in Engineering*, 11(3), 241-252.

Babic, N. Č., Podbreznik, P., & Rebolj, D. (2010). Integrating resource production and construction using BIM. *Automation in Construction*, 19(5), 539-543. https://doi.org/10.1016/j.autcon.2009.11.005

Barbara, D., Li, Y., and Couto, J. 2002, COOLCAT: an entropy-based algorithm for categorical clustering. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 582-589). ACM.

Barbosa, F., Woetzel, J., Mischke, J., Ribeirinho, M. J., Sridhar, M., Parsons, M., Bertram, N. and Brown, S. 2017. "Reinventing construction through a productivity revolution." *McKinsey Global Institute*.

Bayes, T., and Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. Philosophical Transactions, 53, 370-418.

Behzadan, A. H., Aziz, Z., Anumba, C. J., and Kamat, V. R. 2008. "Ubiquitous location tracking for context-specific information delivery on construction sites." *Automation in Construction*, *17*(6), 737-748.

Behzadan, A. H., Menassa, C. C., and Pradhan, A. R. 2015. "Enabling real time simulation of architecture, engineering, construction, and facility management (AEC/FM) systems: a review of formalism, model architecture, and data representation." *ITcon*. 20: 1-23.

Beichl, I., and Sullivan, F. 2000. "The metropolis algorithm." *Computing in Science & Engineering*. 2(1): 65-69.

Bellman, R. (1954). "The theory of dynamic programming." *Bulletin of the American Mathematical Society*, *60*(6), 503-515.

Berteaux, F., and Javernick-Will, A. 2015. "Adaptation and integration for multinational project-based organizations." *Journal of Management in Engineering*, 31(6), 04015008.

Biller, B., and Nelson, B. L. 2002. "Answers to the top ten input modeling questions." In Vol. 1 of *Proceedings of the 2002 Winter Simulation Conference*, 35-40. Piscataway, NJ: IEEE

Bordley, Robert F. 1982. "A multiplicative formula for aggregating probability assessments." *Management science.* 28(10): 1137-1148.

Bourgon, R., 2015. *intervals: Tools for Working with Points and Intervals*. R package version 0.15.1. https://CRAN.R-project.org/package=intervals

Bowden, S., Dorr, A., Thorpe, T., and Anumba, C. 2006. "Mobile ICT support for construction process improvement." *Automation in Construction*, *15*(5), 664-676.

Brandley, R. L., Bergman, J. J., Noble, J. S., and McGarvey, R. G. 2015. "Evaluating a Bayesian approach to demand forecasting with simulation." In *Proceedings of the 2015 Winter Simulation Conference*, 1868-1879. Piscataway, NJ: IEEE

Burke, G. P., and Miller, R. C. 1998. "Modularization speeds construction." *Power engineering*. 102(1): 20-23.

Caldas, C. H., and Soibelman, L. 2003. "Automating hierarchical document classification for construction management information systems." *Automation in Construction*, *12*(4), 395-406.

Caldas, C. H., Soibelman, L., and Han, J. 2002. "Automated classification of construction project documents." *Journal of Computing in Civil Engineering*, *16*(4), 234-243.

Chassiakos, A. P., and Sakellaropoulos, S. P. 2008. "A web-based system for managing construction information." *Advances in Engineering Software*, 39(11), 865-876.

Chau, K. W. 1995. "Monte Carlo simulation of construction costs using subjective data." *Construction Management and Economics*. 13(5): 369-383.

Chau, K. W. 1995. "The validity of the triangular distribution assumption in Monte Carlo simulation of construction costs: empirical evidence from Hong Kong." *Construction Management and Economics*. 13(1): 15-21.

Chen, P., Buchheit, R. B., Garrett Jr, J. H., and McNeil, S. 2005. "Web-vacuum: Web-based environment for automated assessment of civil infrastructure data." *Journal of computing in civil engineering*. 19(2): 137-147.

Cheng, M. Y., Prayogo, D., and Tran, D. H. 2015. "Optimizing multiple-resources leveling in multiple projects using discrete symbiotic organisms search." *Journal of Computing in Civil Engineering*. *30*(3): 04015036

Chung, B. Y., Skibniewski, M. J., Lucas Jr, H. C., and Kwak, Y. H. 2008. "Analyzing enterprise resource planning system implementation success factors in the engineering–construction industry." *Journal of Computing in Civil Engineering*, 22(6), 373-382. https://doi.org/10.1061/(ASCE)0887-3801(2008)22:6(373)

Chung, T. H., Mohamed, Y., and AbouRizk, S. 2004. "Simulation input updating using Bayesian techniques." In *Proceedings of the 2004 Winter Simulation Conference*, 1238-1243. Piscataway, NJ: IEEE

Chwif, L., Banks, J., de Moura Filho, J. P., and Santini, B. 2013. "A framework for specifying a discrete-event simulation conceptual model". *Journal of Simulation*, *7*(1): 50-60.

CII (Construction Industry Institute). 2006. *Data analysis in support of front end planning implementation*. RR213-11. Austin, TX: CII

CII (Construction Industry Institute). 1995. *Pre-project planning handbook*. SP39-2. Austin, TX: CII

CII (Construction Industry Institute). 2006. *Optimizing construction input in front end planning*. RR241-11. Austin, TX: CII

CII (Construction Industry Institute). 2014. *Industrial Modularization: Five Solution Elements*. RR283-2. Austin, TX: CII

CII (Construction Industry Institute). 2015. *Innovative delivery of information to the crafts*. RT-327-1. Austin, TX: CII.

Cleary, J. G., and Trigg, L. E. 1995. "An instance-based learner using an entropic distance measure." In Machine Learning Proceedings 1995(pp. 108-114). Morgan Kaufmann.

Clemen, R. T., and Winkler, R. L. 1999. "Combining probability distributions from experts in risk analysis." *Risk analysis*. 19(2): 187-203.

Costa, J. 2017. *Calculating Geometric Means* Accessed July 15, 2020. https://buzzardsbay.org/special-topics/calculating-geometric-mean/

Dash, M., and Liu, H. 1997. "Feature selection for classification." *Intelligent data analysis*. 1(3): 131-156.

Data, Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/data Retrieved December 13, 2020

Davies, K., McMeel, D. J., and Wilkinson, S. 2017. "Making friends with Frankenstein: hybrid practice in BIM." *Engineering, construction and architectural management*. 24(1): 78-93.

Dean, J. (2014). *Big data, data mining, and machine learning: Value creation for business leaders and practitioners*. Hoboken, NJ: Wiley.

DeBrota, D. J., Dittus, R. S., Roberts, S. D., and Wilson, J. R. 1989. "Visual interactive fitting of bounded Johnson distributions." *Simulation*. 52(5): 199-205.

Dy, J. G., and Brodley, C. E. 2004. "Feature selection for unsupervised learning." *Journal of machine learning research*. 5(Aug): 845-889.

Eastman, C., Lee, J. M., Jeong, Y. S., and Lee, J. K. 2009. "Automatic rule-based checking of building designs." *Automation in Construction*, 18(8), 1011-1033. https://doi.org/10.1016/j.autcon.2009.07.002

ElNimr, A., Fagiar, M., and Mohamed, Y. 2016. "Two-way integration of 3D visualization and discrete event simulation for modeling mobile crane movement under dynamically changing site layout." *Automation in construction*, *68*: 235-248.

Fan, G. G. 2018. "Customized Manufacturing Enterprise Resource Planning System for Offsite Modular Light Gauge Steel Construction."

Fan, H., and Li, H. 2013. "Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques." *Automation in construction*. 34: 85-91.

Ferraretti, D., Gamberoni, G., and Lamma, E. 2012. "Unsupervised and supervised learning in cascade for petroleum geology." *Expert Systems with Applications*. 39(10): 9504-9514.

Forcada, N., Casals, M., Roca, X., and Gangolells, M. 2007. "Adoption of web databases for document management in SMEs of the construction sector in Spain." *Automation in Construction*, *16*(4), 411-424.

"Frequently Asked Questions About the National BIM Standard-United States - National BIM Standard - United States". Nationalbimstandard.org. Archived from the original on 16 October 2014. Retrieved 17 October 2014.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian data analysis, CRC Press, Boca Raton, FL.

Genest, C., Weerahandi, S., and Zidek, J. V. 1984. "Aggregating opinions through logarithmic pooling." *Theory and decision*. 17(1): 61-70.

Genest, C., and Zidek, J. V. 1986. "Combining probability distributions: A critique and an annotated bibliography." *Statistical Science.* 1(1): 114-135.

George, R., Bell, L. C., and Edward Back, W. 2008. "Critical activities in the front-end planning process." *Journal of Management in Engineering*. *24*(2): 66-74.

Gibson, N., Holland, C. P., and Light, B. (1999, January). Enterprise resource planning: a business approach to systems development. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences*. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers (pp. 9-pp). IEEE.

Gong, J., and Carlos H. C. 2010. "Computer vision-based video interpretation model for automated productivity analysis of construction operations." *Journal of Computing in Civil Engineering* 24(3): 252-263.

Guerra, B. C., and Leite, F. 2020. "Bridging the Gap between Engineering and Construction 3D Models in Support of Advanced Work Packaging." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *12*(3), 04520029. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000419

Gusfield, D. 1997 *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511574931

Hamdi, O., and Leite, F. 2014. "Conflicting side of building information modeling implementation in the construction industry." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 6(3): 03013004. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000137

Hammad, A., and Zhang, C. 2011. "Towards real-time simulation of construction activities considering spatio-temporal resolution requirements for improving safety and productivity." In *Proceedings of the 2011 Winter Simulation Conference,* 3533-3544. Piscataway, NJ: IEEE

Han, J., Pei, J., and Kamber, M. 2011. *Data mining: concepts and techniques*. Waltham USA: Elsevier.

Hard data. (n.d.) McGraw-Hill Dictionary of Scientific & Technical Terms, 6E. 2003. https://encyclopedia2.thefreedictionary.com/hard+data Retrieved December 13, 2020

Hard data, Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/hard-data Retrieved December 13, 2020

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*. 57(1): 97-109.

Heon J. D., and El-Rayes, K. 2011. "Multiobjective optimization of resource leveling and allocation during construction scheduling." *Journal of construction engineering and management*. *137*(12): 1080-1088.

Hitchcock, D. B. (2003). "A history of the Metropolis–Hastings algorithm." The American Statistician, 57(4), 254-257.

Hovnanian, G., Kroll, K., and Sjödin, E. 2019. How analytics can drive smarter engineering and construction decisions. *McKinsey & Company Capital Projects & Infrastructure*.

Hu, W. 2008, November. Information lifecycle modeling framework for construction project lifecycle management. In *2008 International Seminar on Future Information Technology and Management Engineering* (pp. 372-375). IEEE.

Hubbard, D. W. 2009. *The failure of risk management: Why it's broken and how to fix it*. Hoboken, New Jersey, US: John Wiley & Sons.

Hwang, B. G., and Ho, J. W. 2011. "Front-end planning implementation in Singapore: Status, importance, and impact." *Journal of construction engineering and management*. *138*(4): 567-573.

Inyim, P., Zhu, Y., and Orabi, W. 2016. “Analysis of time, cost, and environmental impact relationships at the building-material level.” *Journal of Management in Engineering*. 32(4): 04016005.

Jain, A.K., Murty, M. N., and Flynn, P. 1999. “Data clustering: A review.” *ACM Computing Surveys*. 31(3): 264–323.

Jen, H., and Hsiao, C. 2018. "Using Bayesian inference modeling in estimating important production parameters used in the simulation-based production planning." *Proceedings of IEEE International Conference on Applied System Innovation* 2018, IEEE, Piscataway, NJ, 1038-1041.

Ji, W., and AbouRizk, S. M. 2017. “Credible interval estimation for fraction nonconforming: Analytical and numerical solutions.” *Automation in Construction*, 83, 56-67.

Ji, W., and AbouRizk, S. M. 2018a. “Data-Driven Simulation Model for Quality-Induced Rework Cost Estimation and Control Using Absorbing Markov Chains.” *Journal of Construction Engineering and Management*, 144(8), 04018078.

Ji, W., and AbouRizk, S. M. 2018b. “Simulation-based analytics for quality control decision support: A pipe welding case study.” *Journal of Computing in Civil Engineering*, 32(3): 05018002.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). Continuous univariate distributions, 2nd ed. John Wiley and Sons, New York, NY.

Jung, Y., and Gibson, G. E. 1999. Planning for computer integrated construction. *Journal of computing in civil engineering*, 13(4): 217-225

Kaner, I., Sacks, R., Kassian, W., and Quitt, T. 2008. “Case studies of BIM adoption for precast concrete design by mid-sized structural engineering firms.” *Journal of Information Technology in Construction (ITcon)*, 13(21), 303-323.

Kayhanian, M., Amardeep Singh, and Scott Meyer. 2002. “Impact of non-detects in water quality data on estimation of constituent mass loading.” *Water Science and Technology,* 45(9): 219-225.

Kimmance, A. G. 2002. “An integrated product and process information modelling system for on-site construction” (Doctoral dissertation, © Andrew George Kimmance).

Koeleman, J., Ribeirinho, M. J., Rockhill, D., Sjödin, E., and Strube, G. 2019. “Decoding digital transformation in construction.” *McKinsey and Company: Chicago, IL, USA*.

Krause, J., Perer, A., and Ng, K. 2016. “Interacting with predictions: Visual inspection of black-box machine learning models.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* 5686-5697.

Kroese, D. P., Brereton, T., Taimre, T., and Botev, Z. I. (2014). "Why the Monte Carlo method is so important today." Wiley Interdisciplinary Reviews: Computational Statistics, 6(6), 386-392.

Kruschke, J. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London, UK: Academic Press.

Kuhl, M. E., Lada, E. K., Steiger, N. M., Wagner, M. A., and Wilson, J. R. 2006. “Introduction to modeling and generating probabilistic input processes for simulation.” In *Proceedings of the 2006 Winter simulation conference,* 19-35. Piscataway, NJ: IEEE

Kuhn, M. 2008. "Building Predictive Models in R Using the caret Package." *Journal of Statistical Software, 28(5)*, 1 - 26. http://dx.doi.org/10.18637/jss.v028.i05

Leite, F., Cho, Y., Behzadan, A. H., Lee, S., Choe, S., Fang, Y., and Hwang, S. 2016. "Visualization, information modeling, and simulation: Grand challenges in the construction industry." *Journal of Computing in Civil Engineering*, 30(6): 04016035.

Li, Z., Wu, L., and AbouRizk, S. 2019. XiaomoLing/LongestCommonSubString: First Release of Longest Common SubString R Library (Version v1.0.0). Zenodo, http://doi.org/10.5281/zenodo.4057067

Li, Y., Ji, W., and AbouRizk, S. M. 2019. "Enhanced Welding Operator Quality Performance Measurement: Work Experience-Integrated Bayesian Prior Determination." In *Computing in Civil Engineering 2019: Data, Sensing, and Analytics,* 606-613. Reston, VA: American Society of Civil Engineers. ASCE International Conference on Computing in Civil Engineering 2019

Li, Y., and Liu, C. 2012. "Integrating field data and 3D simulation for tower crane activity monitoring and alarming." *Automation in Construction*. 27: 111-119.

Li, Y., Xu, S., Wu, L., AbouRizk, S., Tae, K., & Lei Z. 2019. "A generic simulation model for selecting fleet size in snow plowing operations" *Proceedings of the Winter Simulation Conference*, National Harbor, MD, US. https://doi.org/10.1109/WSC40007.2019.9004954

Liao, C. W., and Perng, Y. H. 2008. "Data mining for occupational injuries in the Taiwan construction industry." *Safety science*. 46(7): 1091-1102.

Lindley, D. V. 1985. "Reconciliation of discrete probability distributions." *Bayesian statistics*, 2:375-390.

Liu, C., Lei, Z., Morley, D., and AbouRizk, S. M. 2020. "Dynamic, data-driven decision-support approach for construction equipment acquisition and disposal." *Journal of Computing in Civil Engineering*, 34(2): 04019053.

Liu, H. C., You, J. X., Lin, Q. L., and Li, H. 2015. "Risk assessment in system FMEA combining fuzzy weighted average with fuzzy decision-making trial and evaluation laboratory." *International Journal of Computer Integrated Manufacturing*. *28*(7): 701-714.

Loosemore, M., Dainty, A., and Lingard, H. 2003. *Human resource management in construction projects: strategic and operational approaches*. UK: Taylor & Francis Group.

Louis, J., and Dunston, P. S. 2017. "Methodology for real-time monitoring of construction operations using finite state machines and discrete-event operation models." *Journal of construction engineering and management*, 143(3): 04016106.

Lu, M., and AbouRizk, S. M. 2000. "Simplified CPM/PERT simulation model." *Journal of Construction Engineering and Management*. 126(3): 219-226.

Lu, M. 2003. "Simplified discrete-event simulation approach for construction simulation." *Journal of Construction Engineering and Management*. 129(5): 537-546.

Lu, Y., Li, Y., Skibniewski, M., Wu, Z., Wang, R., and Le, Y. 2015. "Information and communication technology applications in architecture, engineering, and construction

organizations: A 15-year review." *Journal of Management in Engineering*, *31*(1), A4014010.

Manyika, J., Ramaswamy, S., Khanna, S., Sarrazin, H., Pinkus, G., Sethupathy, G., and Yaffe, A. 2015. "Digital America: A tale of the haves and have-mores" *McKinsey Global Institute*.

Markou, C., Koulinas, G. K., and Vavatsikos, A. P. 2017. "Project resources scheduling and leveling using Multi-Attribute Decision Models: Models implementation and case study." *Expert Systems with Applications*. *77*: 160-169.

Martinez, J. C. 2009. "Methodology for conducting discrete-event simulation studies in construction engineering and management." *Journal of Construction Engineering and Management*, 136(1), 3-16.

Martínez-Rojas, M., Marín, N., and Vila, M. A. 2016. "The role of information technologies to address data handling in construction project management." *Journal of Computing in Civil Engineering*, 30(4): 04015064.

McDonald, J. H. 2009. *Handbook of biological statistics*. Vol. 2. Baltimore, MD: Sparky House Publishing.

McGee, J. V., Prusak, L., and Pyburn, P. J. 1993. *Managing information strategically: Increase your company's competitiveness and efficiency by using information as a strategic tool* (Vol. 1). John Wiley and Sons.

Menesi, W., and Hegazy, T. 2014. "Multimode resource-constrained scheduling and leveling for practical-size projects." *Journal of management in engineering*. *31*(6): 04014092.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics*. 21(6): 1087-1092.

Milo, M. W., Roan, M., and Harris, B. 2015. "A new statistical approach to automated quality control in manufacturing processes." *Journal of Manufacturing Systems*. 36: 159-167.

Mitchell, V. L. 2006. "Knowledge integration and information technology project performance." Mis Quarterly, 30(4), 919-939.

Moller, J., and Waagepetersen, R. P. (2003). Statistical inference and simulation for spatial point processes. CRC Press, Boca Raton.

Morris, P. A. 1974. "Decision analysis expert use." *Management Science.* 20(9): 1233-1241.

Nath D., Kurmi J., and Rawat V., 2018. "A Survey on Longest Common Subsequence." International journal for research in applied science and engineering technology 6(4), 4553-4557. https://doi.org/10.22214/ijraset.2018.4746

Neelamkavil, J. 2009. "Automation in the prefab and modular construction industry." In *26th Symposium on Construction Robotics ISARC*.

Nelson, B. L., and Yamnitsky, M. 1998. "Input modeling tools for complex problems." In Vol. 1 of *Proceedings of the 1998 Winter Simulation Conference.* 105-112. Piscataway, NJ: IEEE.

Neyman, J. (1937). "X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." Phil. Trans. R. Soc. Lond. A, 236(767), 333-380.

Ng, S. T., Xu, F. J., Yang, Y., and Lu, M. 2017. "A master data management solution to unlock the value of big infrastructure data for smart, sustainable and resilient city planning." *Procedia Engineering*, *196*, 939-947.

Nitithamyong, P., and Skibniewski, M. J. 2004. "Web-based construction project management systems: how to make them successful?". *Automation in Construction*, *13*(4), 491-506.

Olatunji, O. A. 2016. "Constructing dispute scenarios in building information modeling." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 8(1): C4515001. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000165.

Oxford Dictionaries. (2019). "Definition of Markov chain in US English." < https://en.oxforddictionaries.com/definition/us/markov_chain> (January 21, 2019).

Penttila, H. 2006. "Describing the changes in architectural information technology to understand design complexity and free-form architectural expression." *Journal of Information Technology in Construction (ITcon)*, 11(29), 395-408. https://www.itcon.org/2006/29

Pereira, E., Ali, M., Wu, L., and Abourizk, S. 2020. "Distributed simulation–based analytics approach for enhancing safety management systems in industrial construction." *Journal of Construction Engineering and Management*, 146(1): 04019091. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001732

Poshdar, M., González, V. A., Raftery, G. M., Orozco, F., and Cabrera-Guerrero, G. G. 2018. "A multi-objective probabilistic-based method to determine optimum allocation of time buffer in construction schedules." *Automation in Construction*. 92: 46-58.

Pradhan, A., and Akinci, B. 2012. "A taxonomy of reasoning mechanisms and data synchronization framework for road excavation productivity monitoring." *Advanced Engineering Informatics*. 26(3): 563-573.

Preidel, C., Daum, S., and Borrmann, A. 2017. "Data retrieval from building information models based on visual programming." *Visualization in Engineering*, 5(1), 18.

R Core Team. 2019. "R: A language and environment for statistical computing." *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rao, U. S., Kestur, S., and Pradhan, C. 2008. "Stochastic optimization modeling and quantitative project management." *IEEE software*, 25(3): 29-36.

Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M. 2011. "Internal versus external cluster validation indexes." *International Journal of computers and communications*, *5*(1), 27-34.

Rezgui, Y., Boddy, S., Wetherill, M., and Cooper, G. 2011. "Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction?" *Computer-Aided Design*, *43*(5), 502-515.

Ripley B. and Lapsley M. 2020. *RODBC: ODBC Database Access*. R package version 1.3-17. Accessed May 12, 2021, https://CRAN.R-project.org/package=RODBC

Robert, C., and Casella, G. (2011). "A short history of MCMC: Subjective recollections from incomplete data." Statistical Science, 26(1), 102-115.

Rosenau, M. D., and Githens, G. D. 2005. *Successful project management, 4th Ed*. New York USA: Wiley.

Rudin, C. 2019 "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nat Mach Intell,* 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Santos, J. M., and Morais, F. 2013. "Evaluating entropic based clustering algorithms on biomedical data." *In Proc., 12th Mexican International Conference on Artificial Intelligence* 194-199. IEEE.

Santos, R., Costa, A. A., and Grilo, A. 2017. "Bibliometric analysis and review of Building Information Modelling literature published between 2005 and 2015." *Automation in Construction*, 80, 118-136.

Sanvido, V. E., and Medeiros, D. J. 1990. "Applying computer-integrated manufacturing concepts to construction." *Journal of Construction Engineering and Management*, 116(2), 365-379.

Saraf, N., Langdon, C. S., and Gosain, S. 2007. "IS application capabilities and relational value in interfirm partnerships." *Information Systems Research*, 18(3), 320-339.

Sardroud, J. M. 2015. "Perceptions of automated data collection technology use in the construction industry." *Journal of Civil Engineering and Management*, *21*(1), 54-66.

Sargent, R. G. 2010. "Verification and validation of simulation models." In *Proceedings of the 2010 winter simulation conference,* 166-183 Piscataway, NJ: IEEE.

Sarkar, S., Pramanik, A., Maiti, J., and Reniers, G. 2020. "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data." *Safety science*, *125*, 104616.

Schmucker, K. J. 1982. *Fuzzy sets: Natural language computations and risk analysis*. Rockville, MD, US: Computer Science Press.

Seresht, N. G., and Robinson Fayek, A. 2018. "Dynamic modeling of multifactor construction productivity for equipment-intensive activities." *Journal of Construction Engineering and Management*, *144*(9): 04018091.

Sharfman, M. P., and Fernando, C. S. 2008. "Environmental risk management and the cost of capital." *Strategic management journal*. 29(6): 569-592.

Shi, J. J., and Halpin, D. W. 2003. "Enterprise resource planning for construction business management." *Journal of Construction Engineering and Management*, 129(2), 214-221.

Sidawi, B., and Alsudairi, A. 2014. "The potentials of and barriers to the utilization of advanced computer systems in remote construction projects: case of the Kingdom of Saudi Arabia." *Visualization in Engineering*, *2*(1), 3.

Skibniewski, M., and Golparvar-Fard, M. 2016. "Toward a science of autonomy for physical systems: Construction." *A white paper prepared for the Computing Community Consortium committee of the Computing Research Association*. http://cra.org/ccc/resources/ccc-led-whitepapers/.

Soft data, Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/soft-data Retrieved December 13, 2020

Soibelman, L., and Kim, H. 2002. "Data preparation process for construction knowledge generation through knowledge discovery in databases." *Journal of Computing in Civil Engineering*, *16*(1), 39-48.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I., and Lin, K. Y. 2008. "Management and analysis of unstructured construction data types." *Advanced Engineering Informatics*, *22*(1), 15-27.

Solihin, W., and Eastman, C. 2015. "Classification of rules for automated BIM rule checking development." *Automation in Construction*: 53(1), 69-82. http://doi.org/10.1016/j.autcon.2015.03.003

Solihin, W., Dimyadi, J., Lee, Y. C., Eastman, C., and Amor, R. 2017. "The critical role of accessible data for BIM-based automated rule checking systems." *Proceedings of the Joint Conference on Computing in Construction,* (1), pp. 53-60. https://doi.org/10.24928/JC3-2017/0161

Song, L., and Eldin, N. N. 2012. "Adaptive real-time tracking and simulation of heavy construction operations for look-ahead scheduling." *Automation in Construction*. 27: 32-39.

Sonmez, R. 2005. "Review of conceptual cost modeling techniques." *AACE International Transactions*. ES71

Statisticat. (2013). "Bayesian inference." <chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/BayesianInference.pdf> (January 31, 2019).

Statistics Canada. 2011 "Canada Year Book, 2011" Accessed June 15,2020. https://www150.statcan.gc.ca/n1/pub/11-402-x/2011000/chap/construction/construction-eng.htm

Szlęk, J., Pacławski, A., Lau, R., Jachowicz, R., and Mendyk, A. 2013. "Heuristic modeling of macromolecule release from PLGA microspheres." *International journal of nanomedicine*, *8*, 4601. https://doi.org/10.2147/IJN.S53364

Tatari, O., Castro-Lacouture, D., and Skibniewski, M. J. 2007. "Current state of construction enterprise information systems: Survey research." *Constr. Innovation*, 74, 310–319.

Tatari, O., Castro-Lacouture, D., and Skibniewski, M. J. 2008. "Performance evaluation of construction enterprise resource planning systems." *Journal of Management in Engineering*, 24(4), 198-206.

Tatari, O., Ryoo, B. Y., and Skibniewski, M. J. 2004. "Modeling of ERP system solutions for the construction industry." In *Proc., 5th European Conf. on Product and Process Modeling in AEC Industry* (pp. 393-398). Istanbul, Turkey: Istanbul Technical Univ..

Teicholz, P., and Fischer, M. 1994. "Strategy for computer integrated construction technology." *Journal of Construction Engineering and Management*, 120(1), 117-131.

Thompson, G. I. 1996. "Need for an Enterprise Resource Management Measurement/Forecasting Infrastructure." In *The 1996 22nd International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems*, CMG. Part 1(of 2) (pp. 467-478).

Tinham, B. 1999. "Advancing on planning and scheduling?" *Manufacturing Computer Solutions*, 5(3), 24-5.

Umble, E. J., Haft, R. R., and Umble, M. M. 2003. "Enterprise resource planning: Implementation procedures and critical success factors." *European Journal of Operational Research*, 146(2), 241-257.

Vahdatikhaki, F., and Hammad, A. 2014. "Framework for near real-time simulation of earthmoving projects using location tracking technologies." *Automation in Construction*. 42: 50-67.

Viljamaa, E., and Peltomaa, I. 2014. "Intensified construction process control using information integration." *Automation in Construction*, *39*, 126-133.

Voordijk, H., Van Leuven, A., and Laan, A. 2003. "Enterprise resource planning in a large construction firm: implementation analysis." *Construction Management and Economics*, 21(5), 511-521.

Wang, L., Shen, W., Xie, H., Neelamkavil, J., and Pardasani, A. 2002. "Collaborative conceptual design—state of the art and future trends." *Computer-Aided Design*, 34(13): 981-996.

Wickham, H. 2014. "Tidy data." *Journal of Statistical Software*. 59(10): 1-23. http://theta.edu.pl/wp-content/uploads/2012/10/v59i10.pdf, last accessed Nov 02, 2020.

Wickham, H., François, R., Henry L., and Müller K. 2020. dplyr: A Grammar of Data Manipulation. R package version0.8.4. https://CRAN.R-project.org/package=dplyr

Williams, G. V. 1995. "Fast track pros and cons: Considerations for industrial projects." *Journal of management in engineering*. 11(5): 24-32.

Winkler, R. L., and Cummings, L. L. 1972. "On the choice of a consensus distribution in Bayesian analysis." *Organizational Behavior and Human Performance*. 7(1): 63-76.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. 2016. *Data Mining: Practical machine learning tools and techniques*. Cambridge USA: Elsevier.

Wu, L., Ali, M., Pereira, E., and AbouRizk, S. 2018 "Linear regression and agent-based modeling approach for equipment market value prediction" *Proceedings of the 17th International Conference on Modeling and Applied Simulation*, Budapest, Hungary.

Wu, L., and AbouRizk, S. 2020, *XiaomoLing/Detect3DRelation: First Release of the Detect 3D Relation function* (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.4058576

Wu, L., and AbouRizk, S. 2021a. "Towards construction's digital future: a roadmap for enhancing data value" *9th CSCE International Construction Specialty Conference,* Niagara Falls, Canada [In Press]

Wu, L., and AbouRizk, S. 2021b. "Numerical-Based Approach for Updating Simulation Input in Real Time." *Journal of Computing in Civil Engineering*, *35*(2), 04020067.

Wu, L., Li, Z., and AbouRizk, S., 2020a. "Automation in extraction and sharing information between BIM and project management databases" *Proceedings of the International Conference on Construction and Real Estate Management* (ICCREM), Stockholm, Sweden. https://doi.org/10.1061/9780784483237.005

Wu, L., Ji, W., and AbouRizk, S. M. 2020b. "Bayesian inference with Markov chain Monte Carlo–based numerical approach for input model updating." *Journal of Computing in Civil Engineering*, *34*(1), 04019043.

Wu, L., Mohamed, Y., Taghaddos, H., and Hermann, R. 2014. "Analyzing scaffolding needs for industrial construction sites using historical data." In *Construction Research Congress 2014: Construction in a Global Network* (pp. 1596-1605). https://doi.org/10.1061/9780784413517.163

Yager, R. R., and Kacprzyk, J. (Eds.). 2012. *The ordered weighted averaging operators: theory and applications*. Berlin, Germany: Springer Science & Business Media.

Yuan, X., Chen, Y. W., Fan, H. B., He, W. H., and Ming, X. G. 2019, December. "Collaborative Construction Industry Integrated Management Service System Framework Based on Big Data." In *2019 International Conference on Industrial Engineering and Engineering Management,* (pp. 1521-1525). IEEE.

Zayed, T. M., and Halpin, D. 2001. "Simulation of concrete batch plant production." *Journal of Construction Engineering and Management*. 127(2): 132-141

Zhang, F., Fleyeh, H., Wang, X., and Lu, M. 2019. "Construction site accident analysis using text mining and natural language processing techniques." *Automation in Construction*, *99*, 238-248.

Zhang, L., Ekyalimpa, R., Hague, S., Werner, M., and AbouRizk, S. 2015. "Updating geological conditions using Bayes theorem and Markov chain. In *Proceedings of the 2015 Winter Simulation Conference (WSC),* 3367-3378. Piscataway, NJ: IEEE

Zhang, M., Cao, T., and Zhao, X. 2017. "Applying sensor-based technology to improve construction safety management." *Sensors*, 17(8): 1841.

Zhang, Z., Lee, M. K., Huang, P., Zhang, L., and Huang, X. 2005. "A framework of ERP systems implementation success in China: An empirical study." *International Journal of Production Eco*

Zhong, B., Xing, X., Love, P., Wang, X., and Luo, H. 2019. "Convolutional neural network: Deep learning-based classification of building quality problems." *Advanced Engineering Informatics*, *40*, 46-57.

Zhou, Y. W., Hu, Z. Z., Lin, J. R., and Zhang, J. P. 2019. "A review on 3D spatial data analytics for building information models." *Archives of Computational Methods in Engineering*, 1-15.

Zou, Y., Kiviniemi, A., and Jones, S. W. 2017. "Retrieving similar cases for construction project risk management using Natural Language Processing techniques." *Automation in Construction*, 80: 66-76.

APPENDIX A: SUPPLYMENTRY DATA OF CHAPTER 3

Table A-1 Random sample observations

39.1117	38.1296	38.3570	39.6073	43.0557	38.4125	38.5379	42.2452	39.7328	39.4371
37.7469	38.9056	39.2908	40.4102	41.4452	37.0880	37.1596	38.5262	41.7413	41.5174
40.0859	40.2558	42.0594	42.5692	41.0155	40.6565	39.5214	41.7210	39.0587	43.0453
41.4322	39.7516	42.1266	37.8114	41.7545	43.7355	36.1974	41.7388	39.2034	40.7984
38.4779	37.5317	37.7961	40.4053	39.5397	40.0600	41.4969	37.0703	42.4494	37.9855
39.2276	40.3603	36.6281	39.5792	39.2934	38.8898	42.8492	40.7404	38.0009	37.9231
39.0543	40.3280	40.4806	40.9962	38.7289	40.9901	37.6877	37.6485	39.9221	39.6215
41.6288	40.0520	42.8261	36.8715	37.6524	41.3409	40.0653	39.2155	43.0775	43.3543
39.9616	40.4537	41.9212	40.9779	36.9697	36.6020	37.2736	41.4926	40.7506	41.3381
37.4715	37.7115	43.3153	41.6636	36.7722	39.8993	39.6618	38.2491	39.1360	38.4436

Table A-2 Random sample observations with noise

39.3190	41.1568	40.5044	39.6062	39.5307	37.8915	38.2635	38.3171	40.8220	39.4031
37.5139	40.2163	41.4583	40.5882	41.3203	38.7499	39.0247	41.3504	41.4560	38.2089
39.1404	43.0789	42.2437	41.8150	37.9378	40.5490	41.0662	38.3499	40.7783	39.6539
39.3147	41.8099	42.1317	36.6428	38.2513	38.8294	44.2813	38.0468	42.6481	40.9333
43.2237	42.2768	37.3082	38.2448	38.0313	41.8849	40.4088	41.9140	40.1424	41.0127
38.7793	41.5439	36.0635	39.3326	42.4370	42.7737	42.3768	41.9963	40.3113	39.8324
41.5404	40.0366	42.0074	37.8014	40.7394	37.5615	39.3699	42.4111	41.3478	37.9244
42.5197	38.9267	38.8029	40.1320	36.8774	42.5452	41.5832	38.5655	38.5604	39.8962
38.6484	42.6908	40.0156	44.7790	40.5194	35.4282	40.8148	44.2883	39.4529	38.7011
40.8539	41.8416	41.7084	39.0253	38.1635	41.4857	39.0028	40.0311	41.5266	39.9560

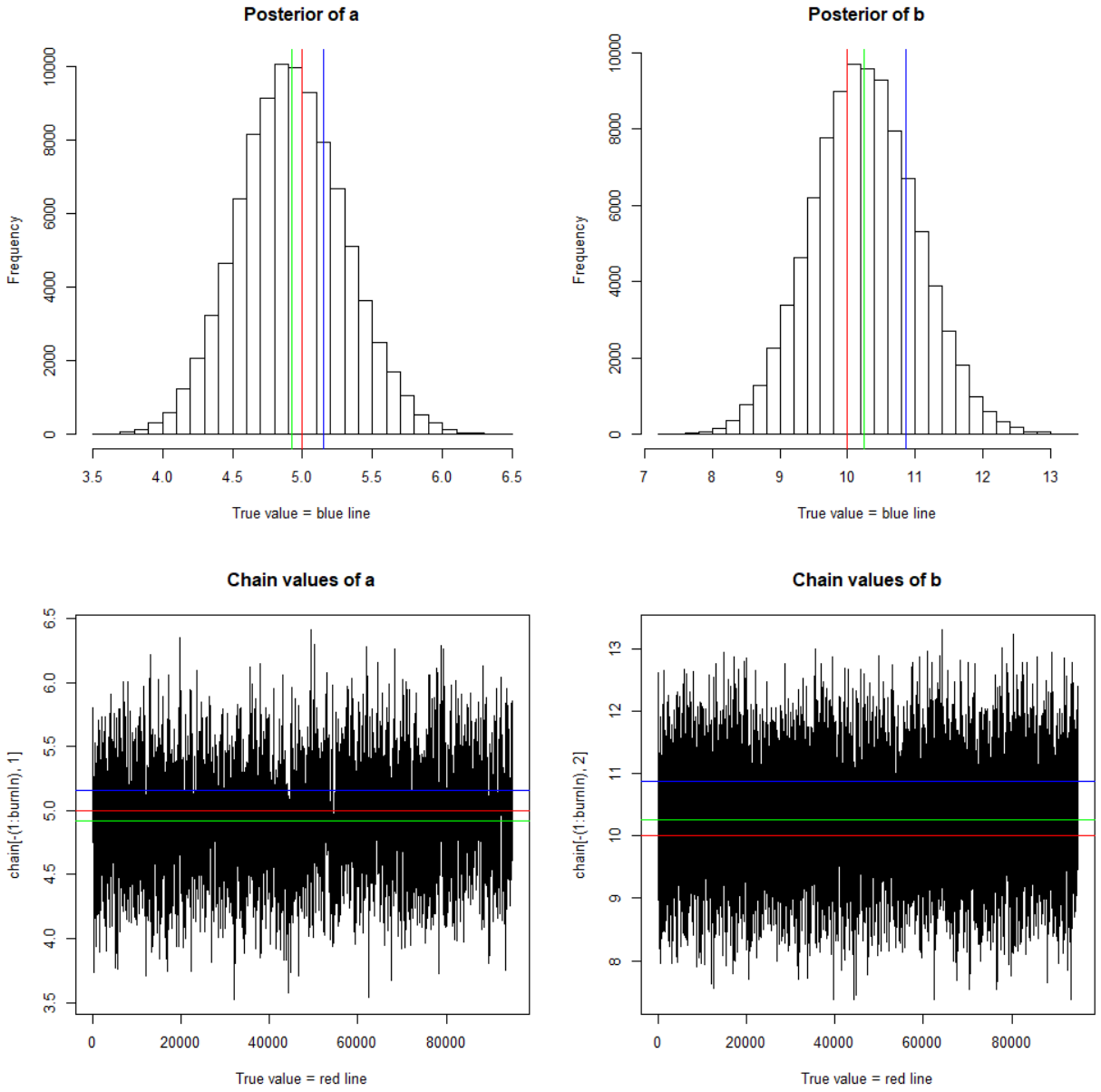


Figure A-1 Posterior histogram (upper) and trace plot (lower) of parameters a (left) and b (right), as well as true parameter values (red line), directly fitted parameter values (blue line), and mean of the MCMC posterior samples (green line) for Cycle 2.

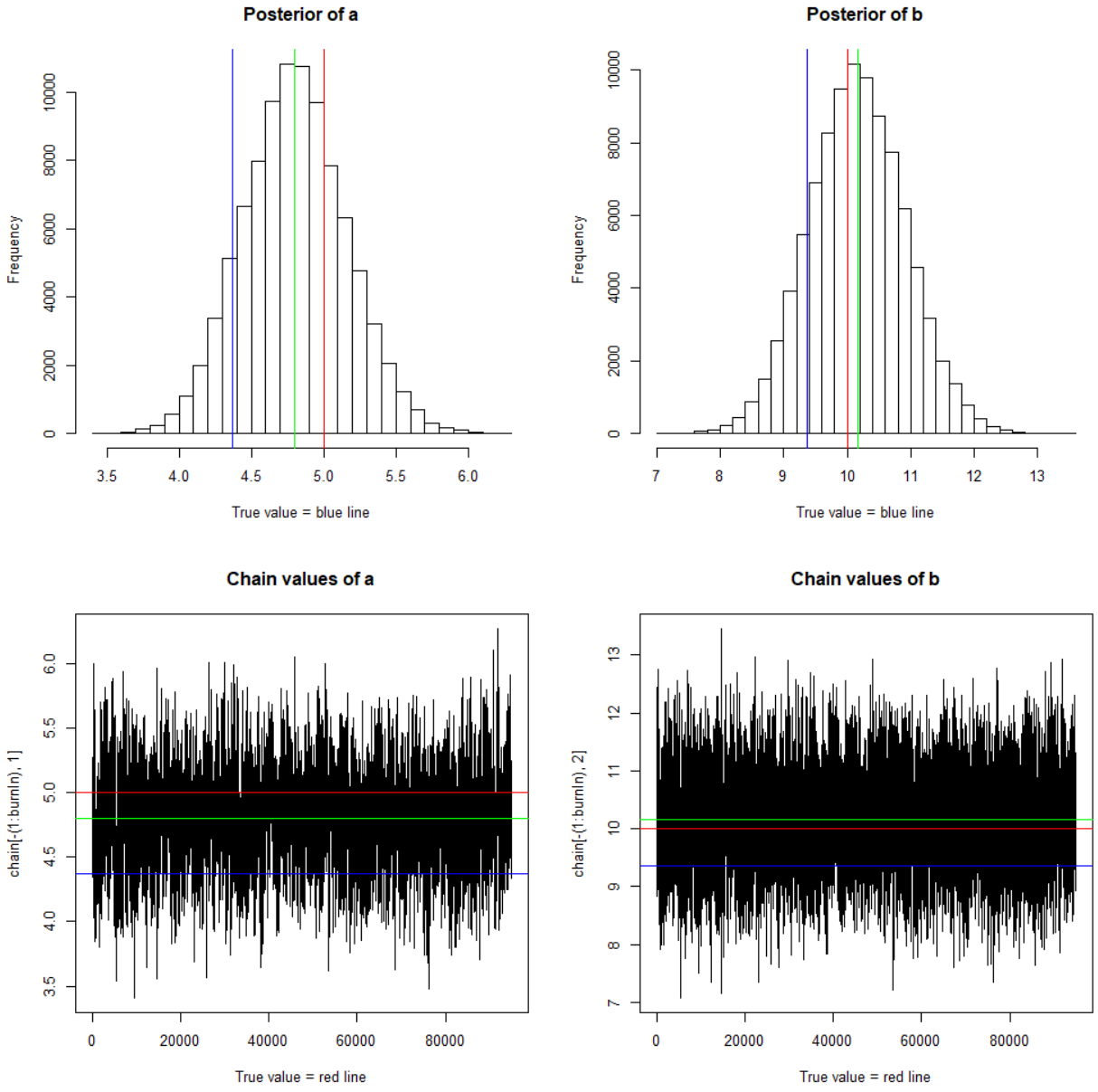


Figure A-2 Posterior histogram (*upper*) and trace plot (*lower*) of parameters a (*left*) and b (*right*), as well as true parameter values (*red line*), directly fitted parameter values (*blue line*), and mean of the MCMC posterior samples (*green line*) for Cycle 3.

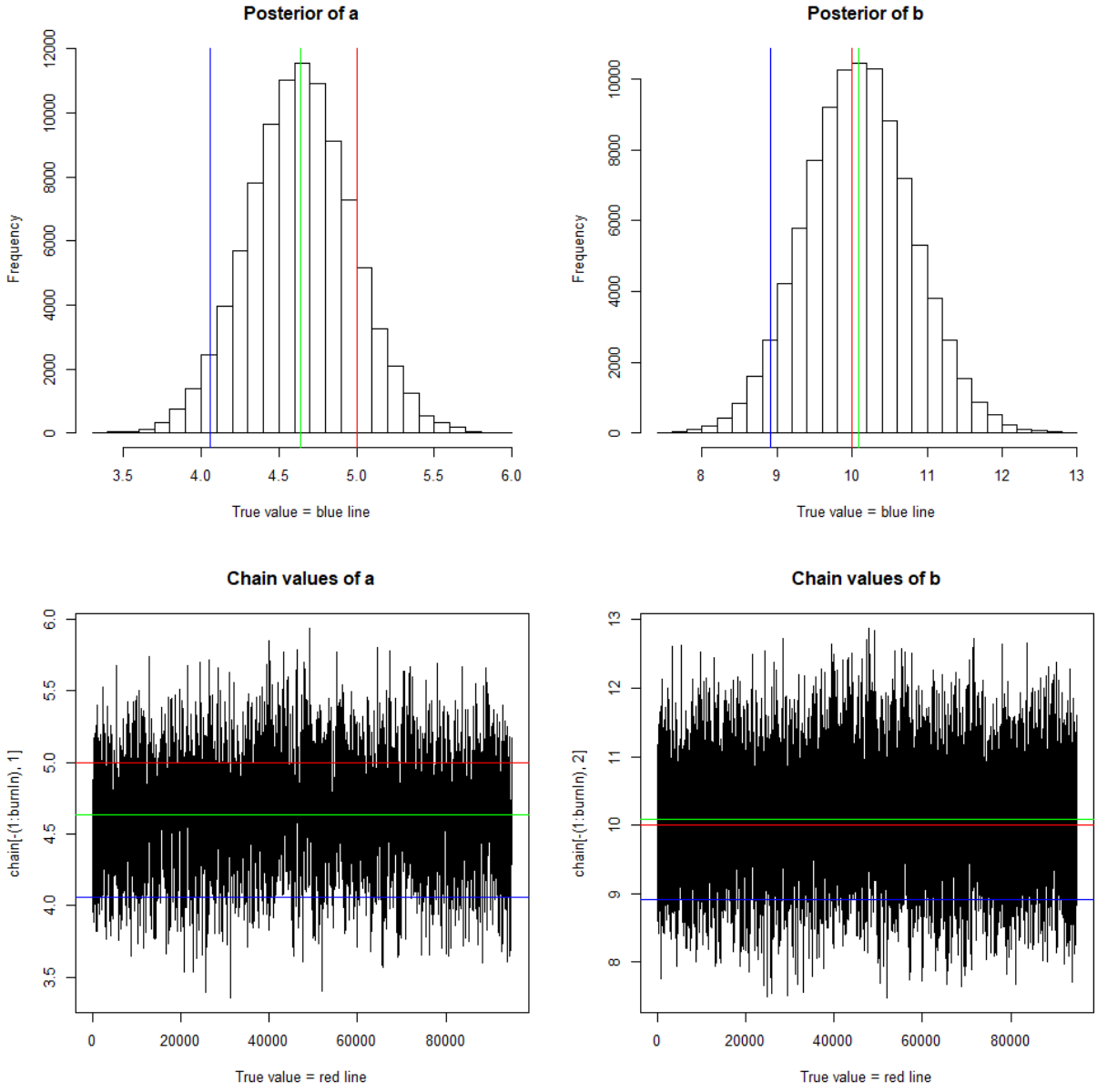


Figure A-3 Posterior histogram (*upper*) and trace plot (*lower*) of parameters a (*left*) and b (*right*), as well as true parameter values (*red line*), directly fitted parameter values (*blue line*), and mean of the MCMC posterior samples (*green line*) for Cycle 4.

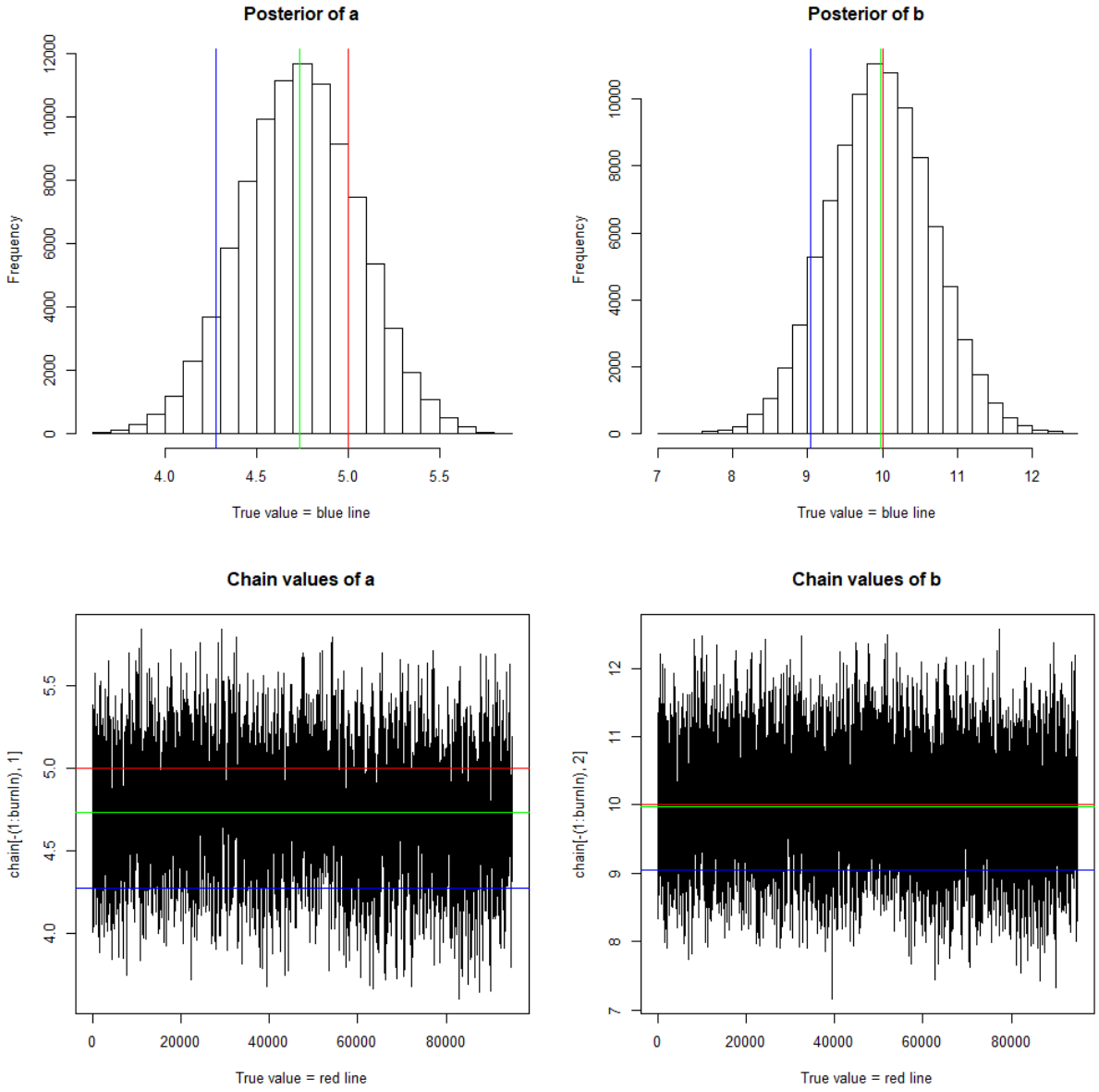


Figure A-4 Posterior histogram (*upper*) and trace plot (*lower*) of parameters a (*left*) and b (*right*), as well as true parameter values (*red line*), directly fitted parameter values (*blue line*), and mean of the MCMC posterior samples (*green line*) for Cycle 5.

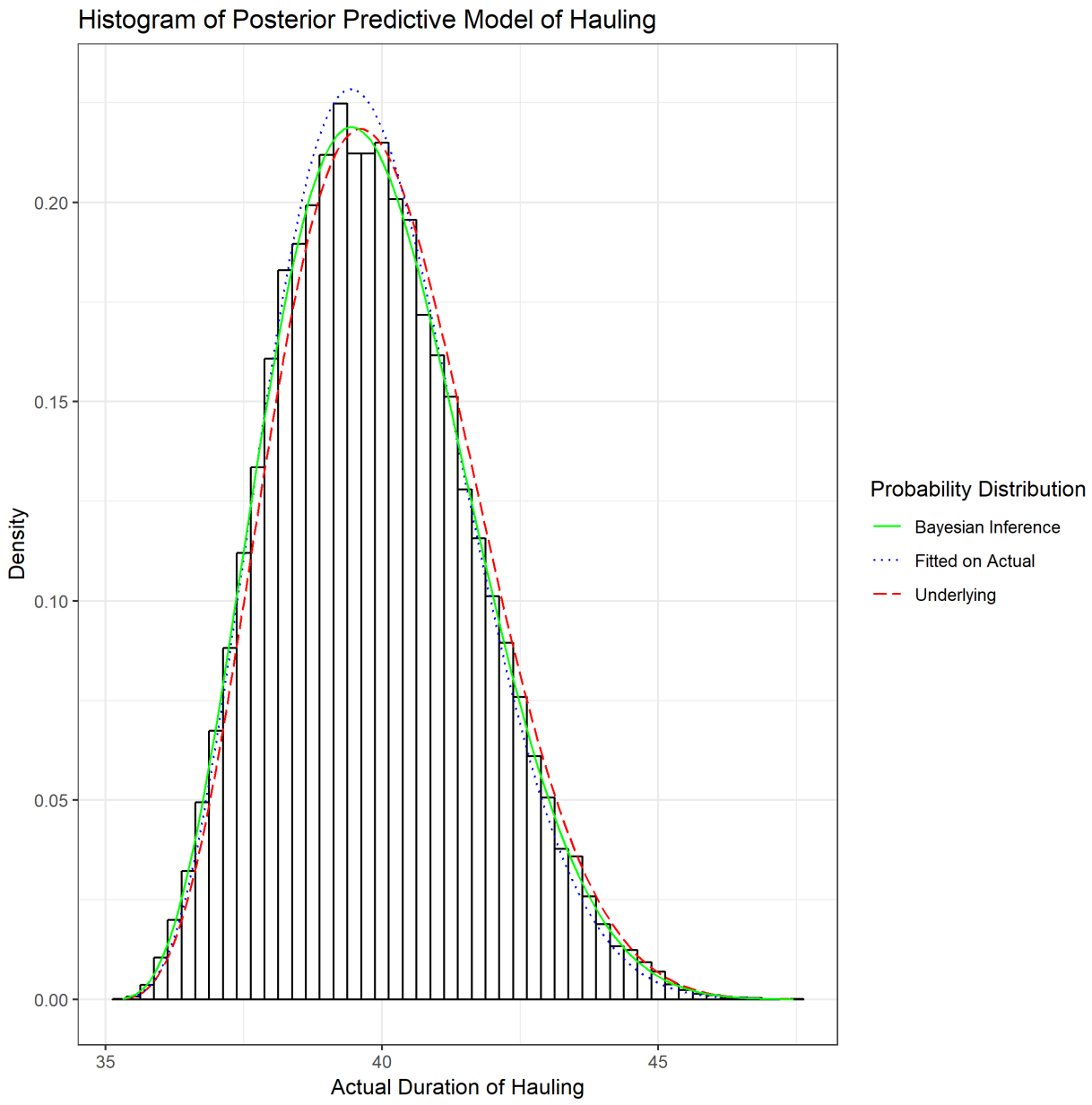


Figure A-5 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (*red line*), the input model fitted from cumulative observations (*blue line*), and the updated input model derived using proposed method (*green line*), for Cycle 2.

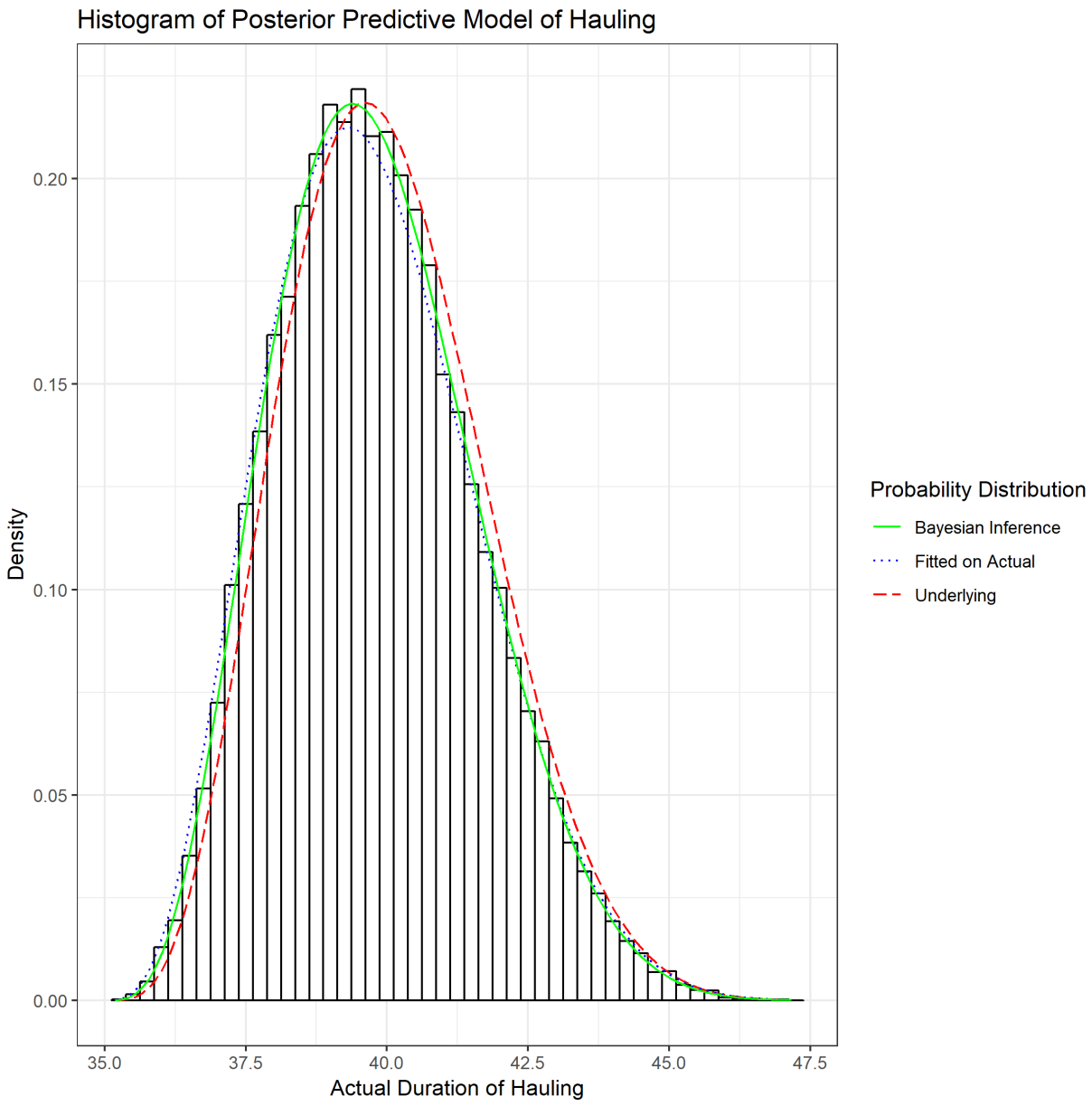


Figure A-6 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (*red line*), the input model fitted from cumulative observations (*blue line*), and the updated input model derived using proposed method (*green line*), for Cycle 3.

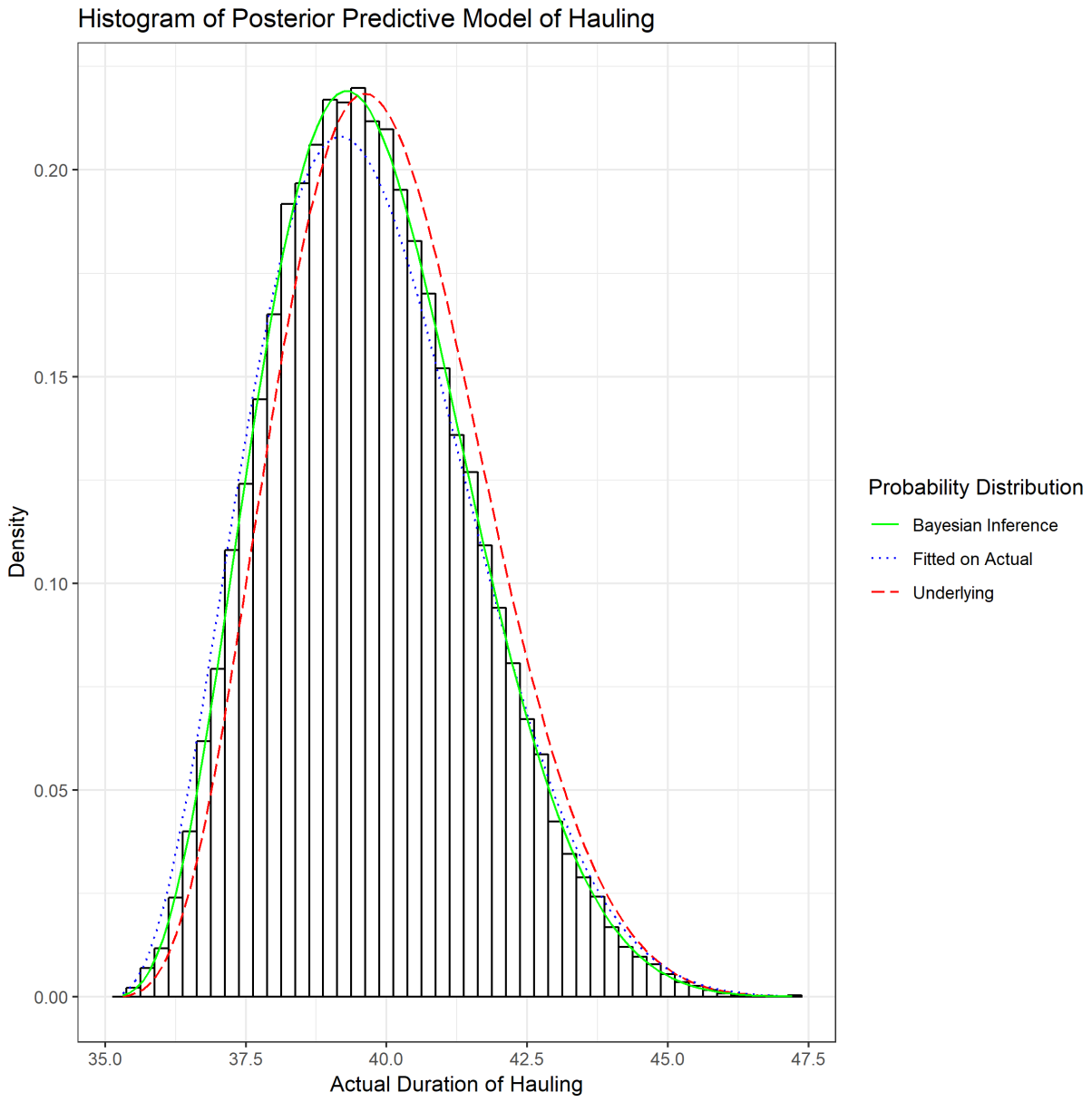


Figure A-7 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (*red line*), the input model fitted from cumulative observations (*blue line*), and the updated input model derived using proposed method (*green line*), for Cycle 4.

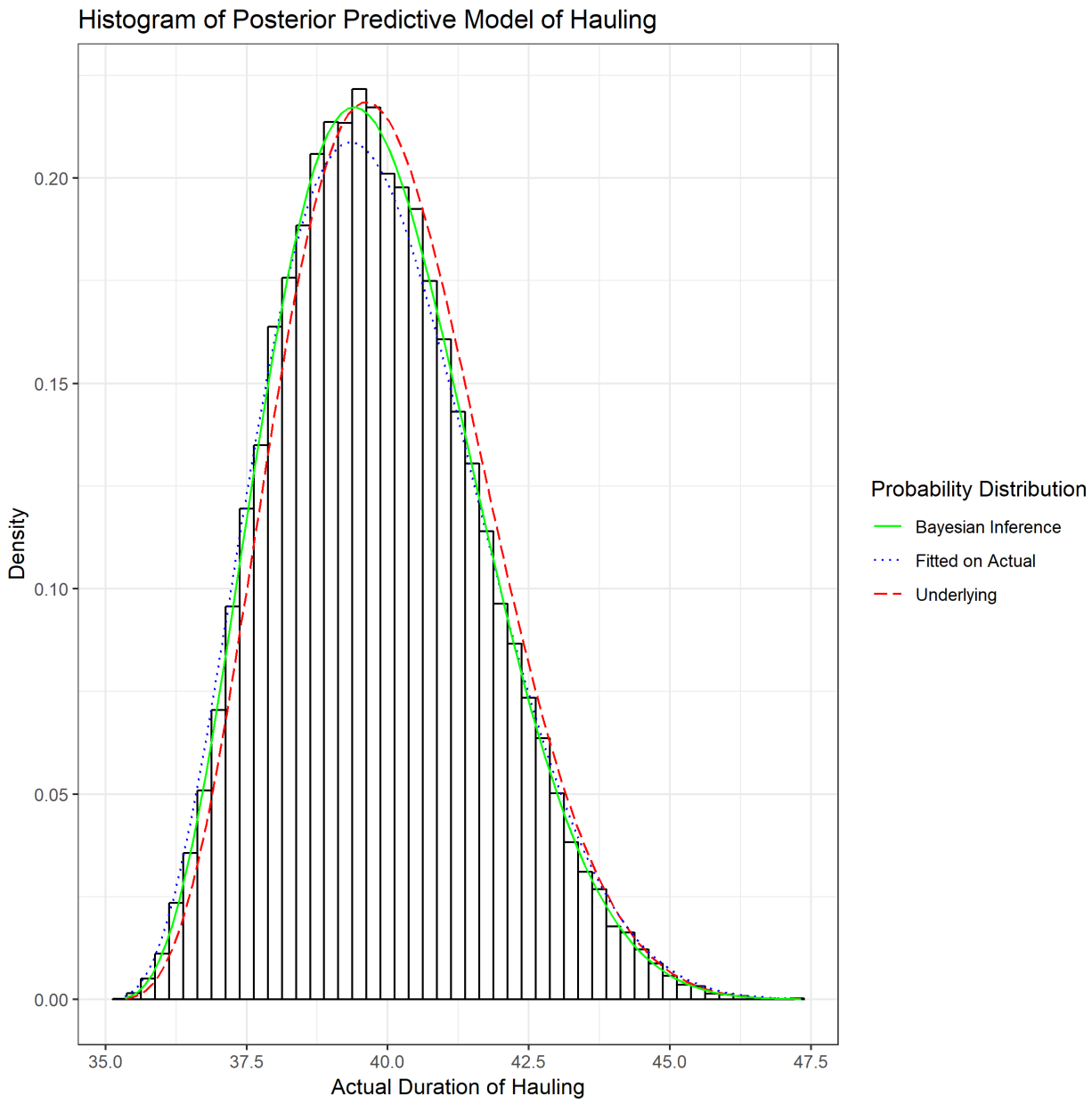


Figure A-8 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (*red line*), the input model fitted from cumulative observations (*blue line*), and the updated input model derived using proposed method (*green line*), for Cycle 5.

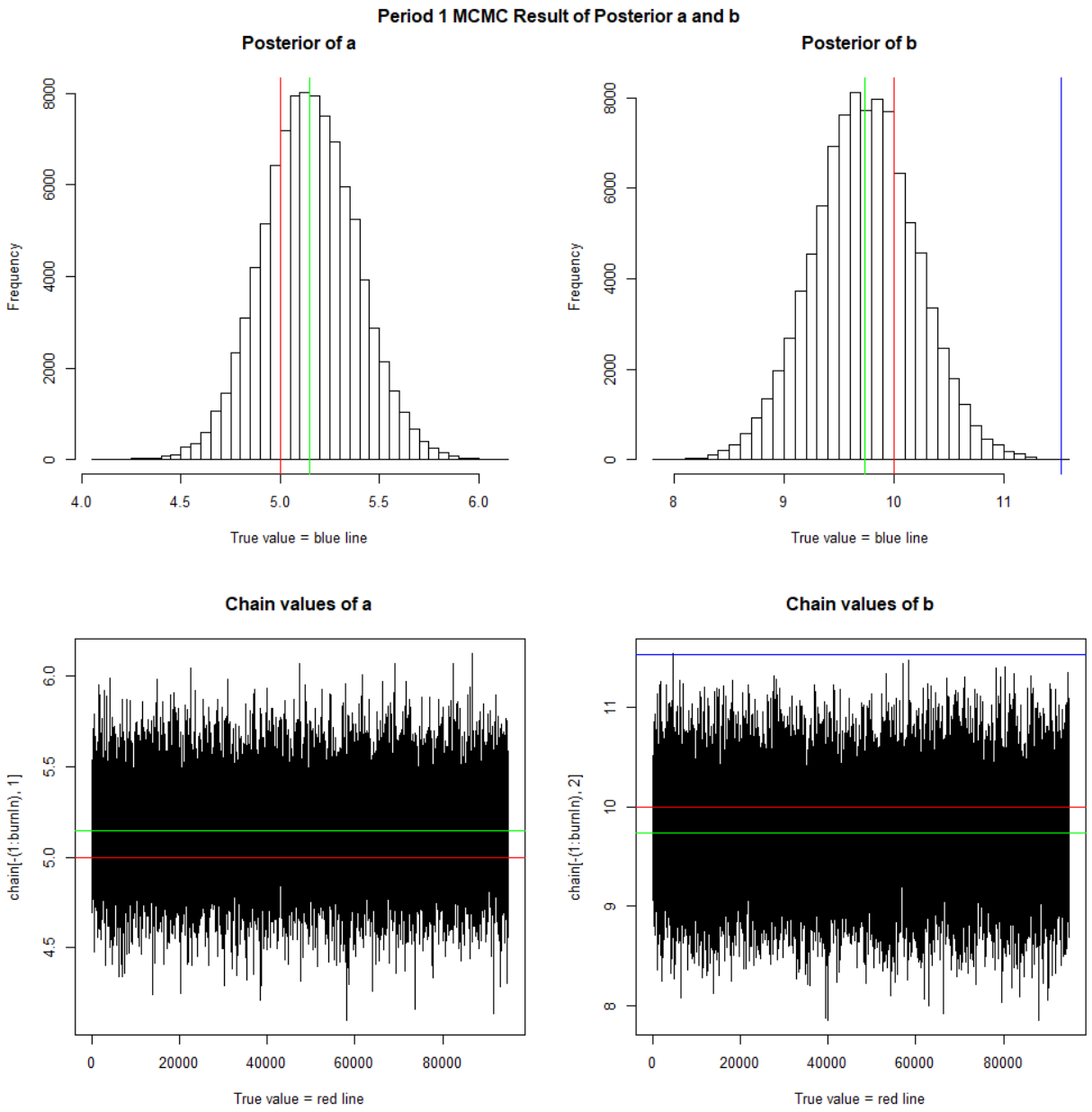


Figure A-9 Posterior histogram (*upper*) and trace plot (*lower*) of parameters a (*left*) and b (*right*), as well as true parameter values (*red line*), directly fitted parameter values (*blue line*), and mean of the MCMC posterior samples (*green line*) for Cycle 1.

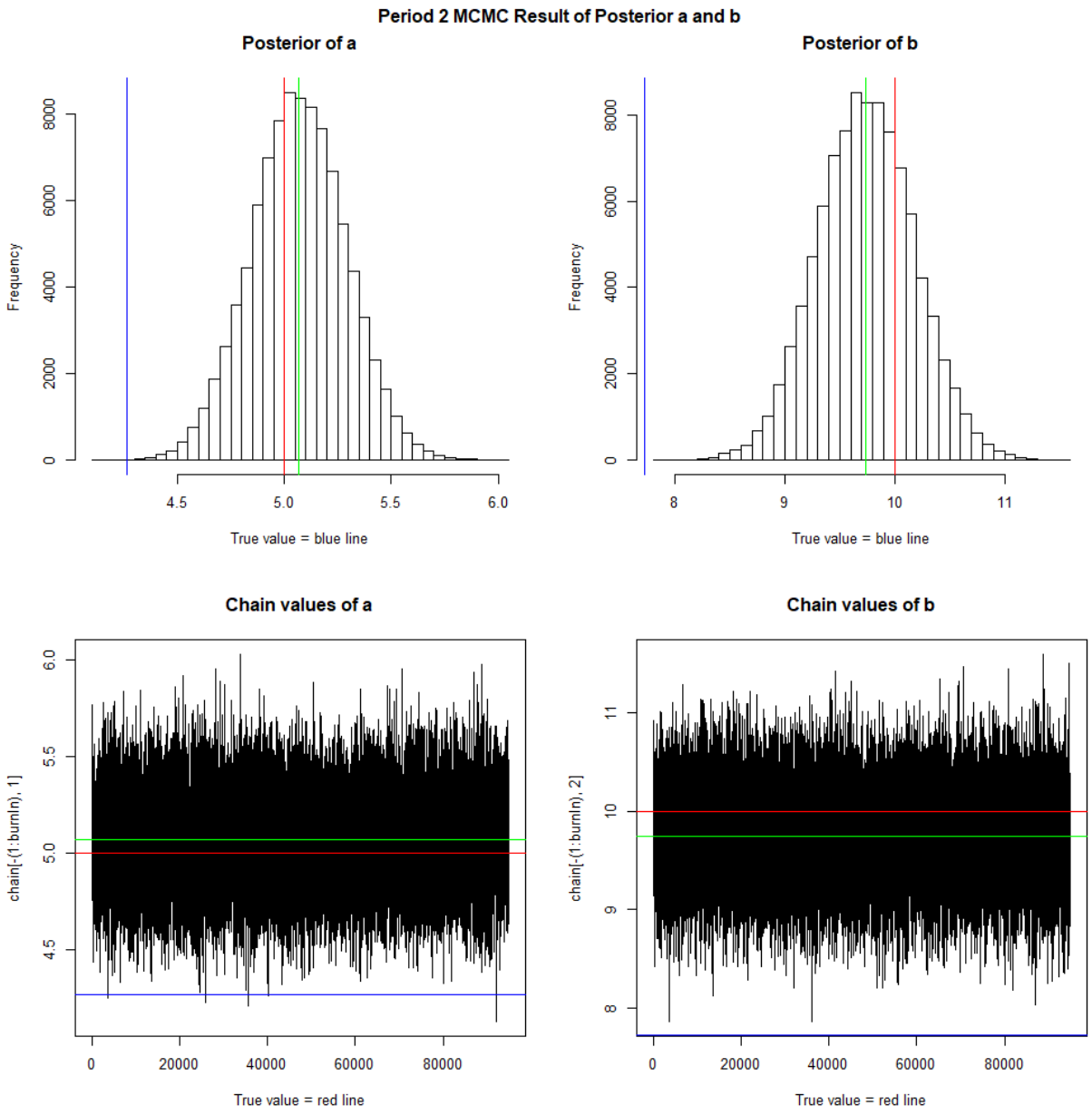


Figure A-10 Posterior histogram (*upper*) and trace plot (*lower*) of parameters a (*left*) and b (*right*), as well as true parameter values (*red line*), directly fitted parameter values (*blue line*), and mean of the MCMC posterior samples (*green line*) for Cycle 2.

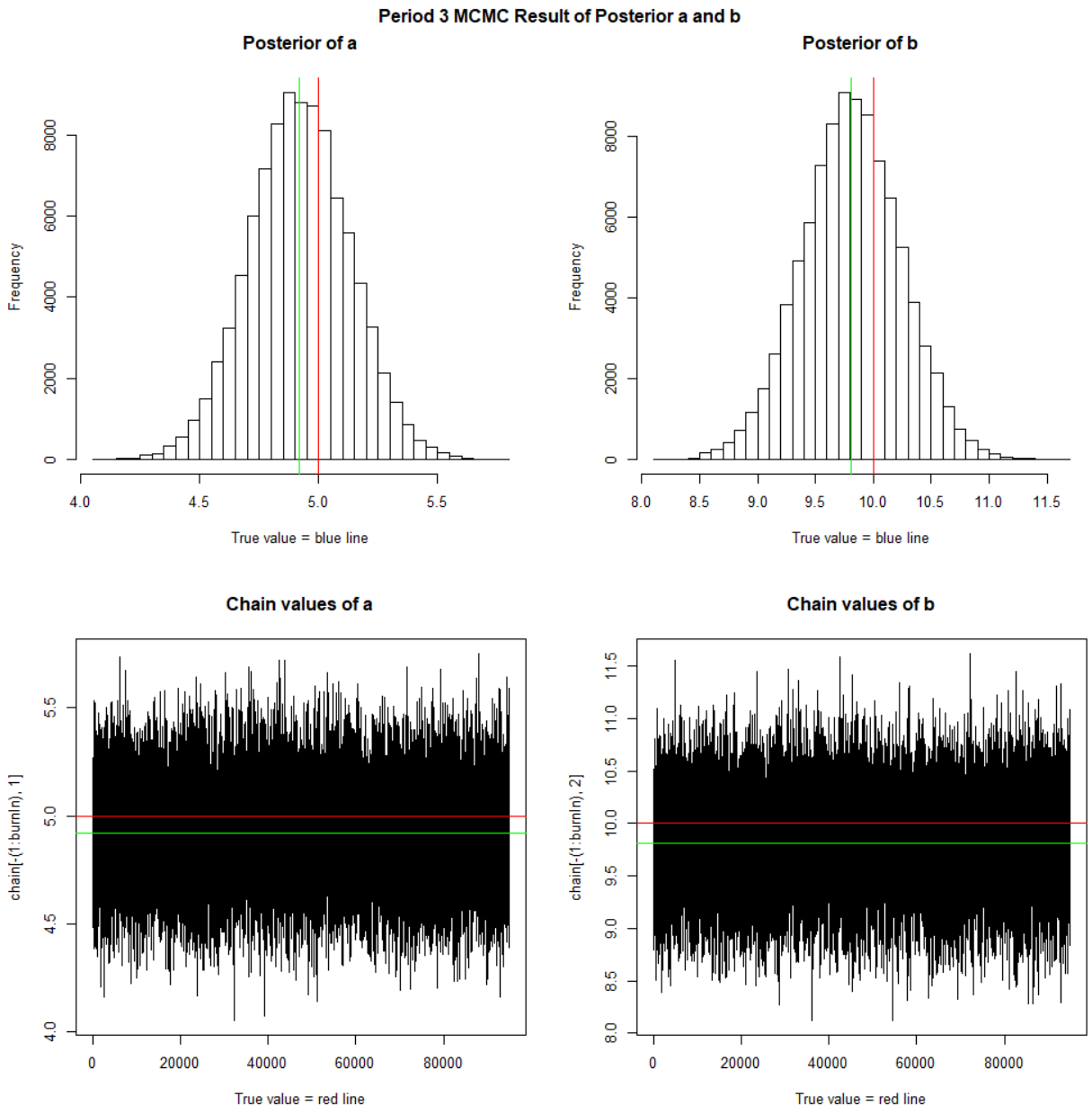


Figure A-11 Posterior histogram (*upper*) and trace plot (*lower*) of parameters a (*left*) and b (*right*), as well as true parameter values (*red line*), directly fitted parameter values (*blue line*), and mean of the MCMC posterior samples (*green line*) for Cycle 3.

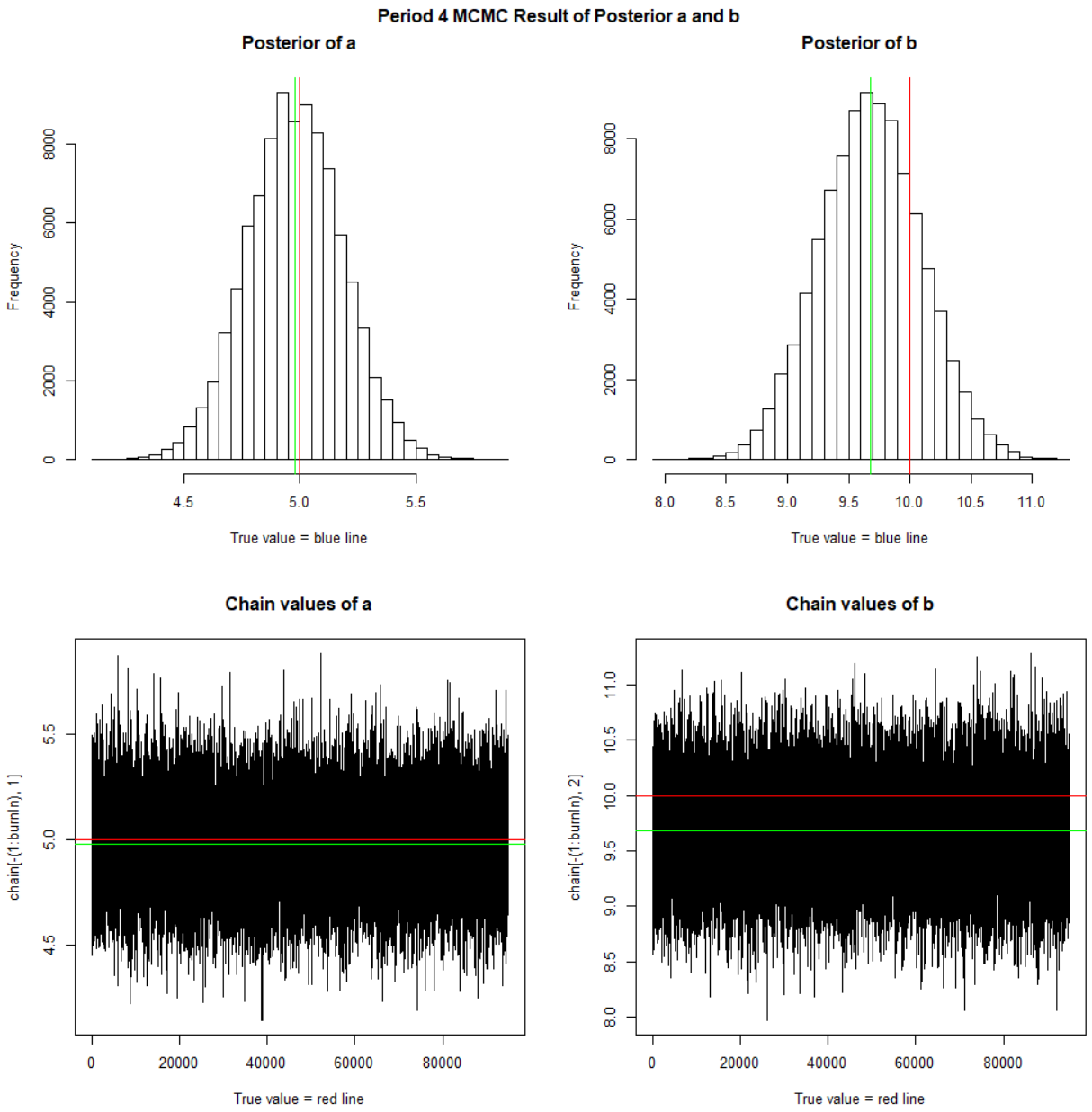


Figure A-12 Posterior histogram (*upper*) and trace plot (*lower*) of parameters a (*left*) and b (*right*), as well as true parameter values (*red line*), directly fitted parameter values (*blue line*), and mean of the MCMC posterior samples (*green line*) for Cycle 4.

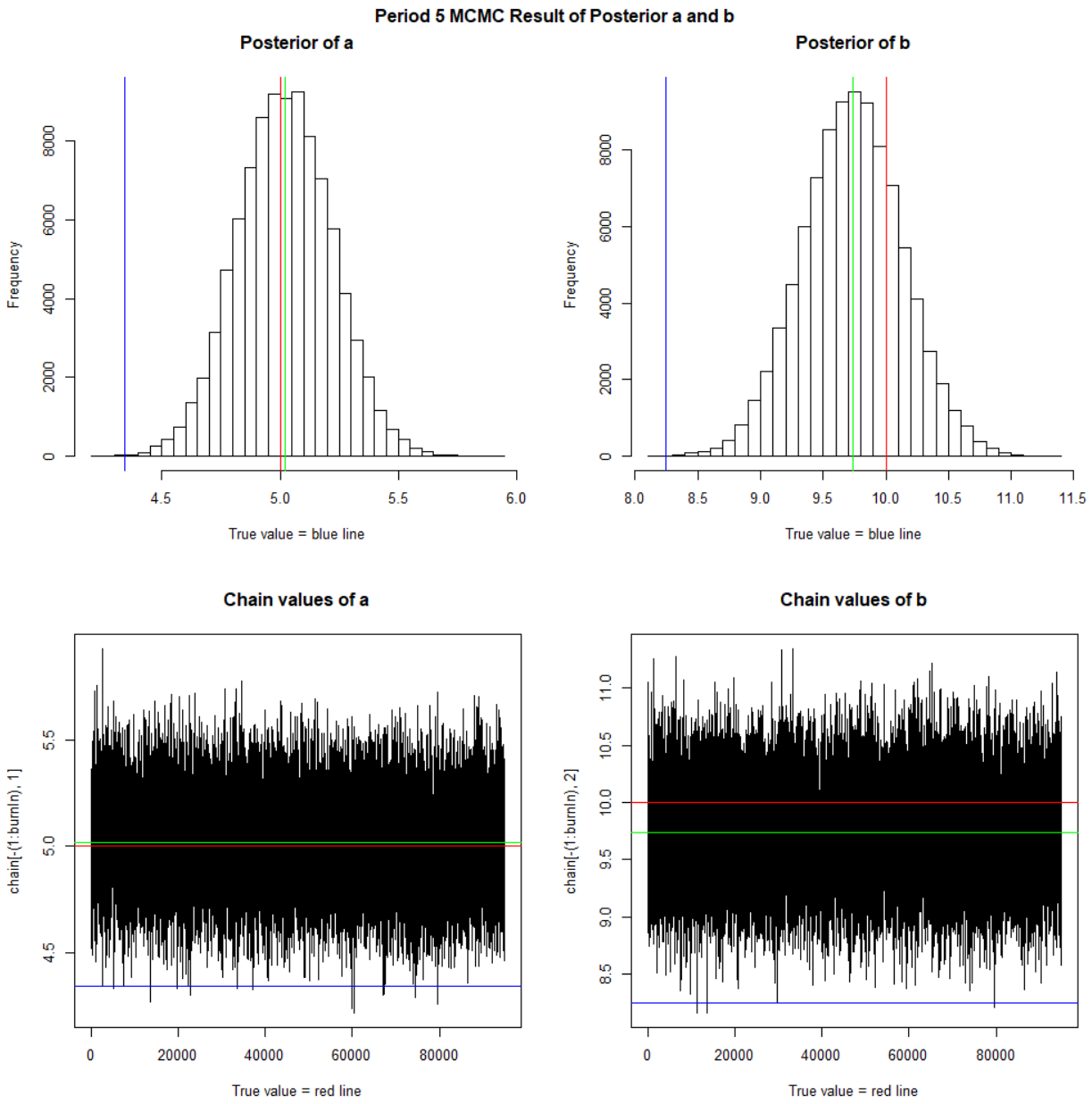


Figure A-13 Posterior histogram (*upper*) and trace plot (*lower*) of parameters a (*left*) and b (*right*), as well as true parameter values (*red line*), directly fitted parameter values (*blue line*), and mean of the MCMC posterior samples (*green line*) for Cycle 5.

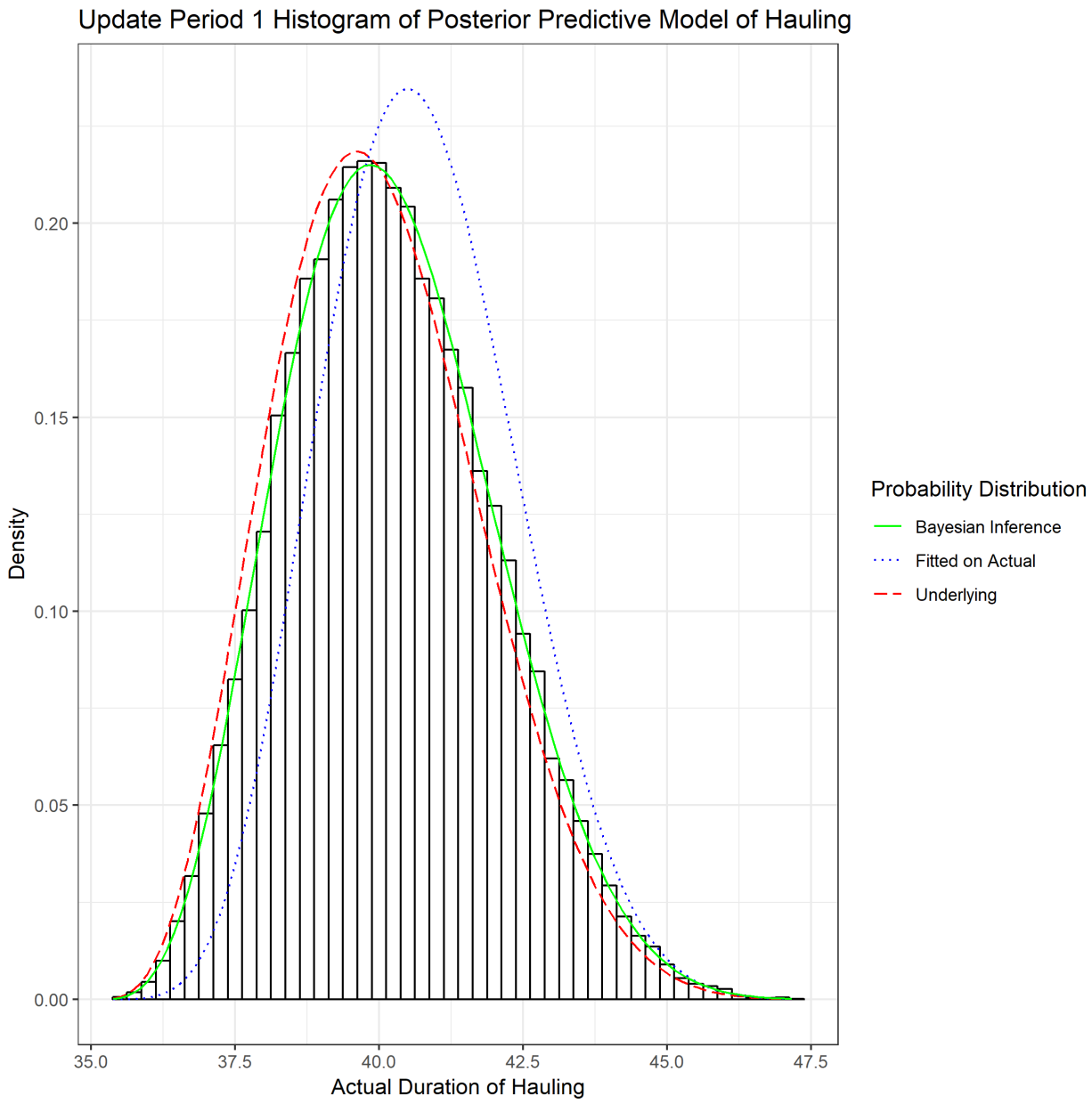


Figure A-14 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (*red line*), the input model fitted from cumulative observations (*blue line*), and the updated input model derived using proposed method (*green line*), for Cycle 1.

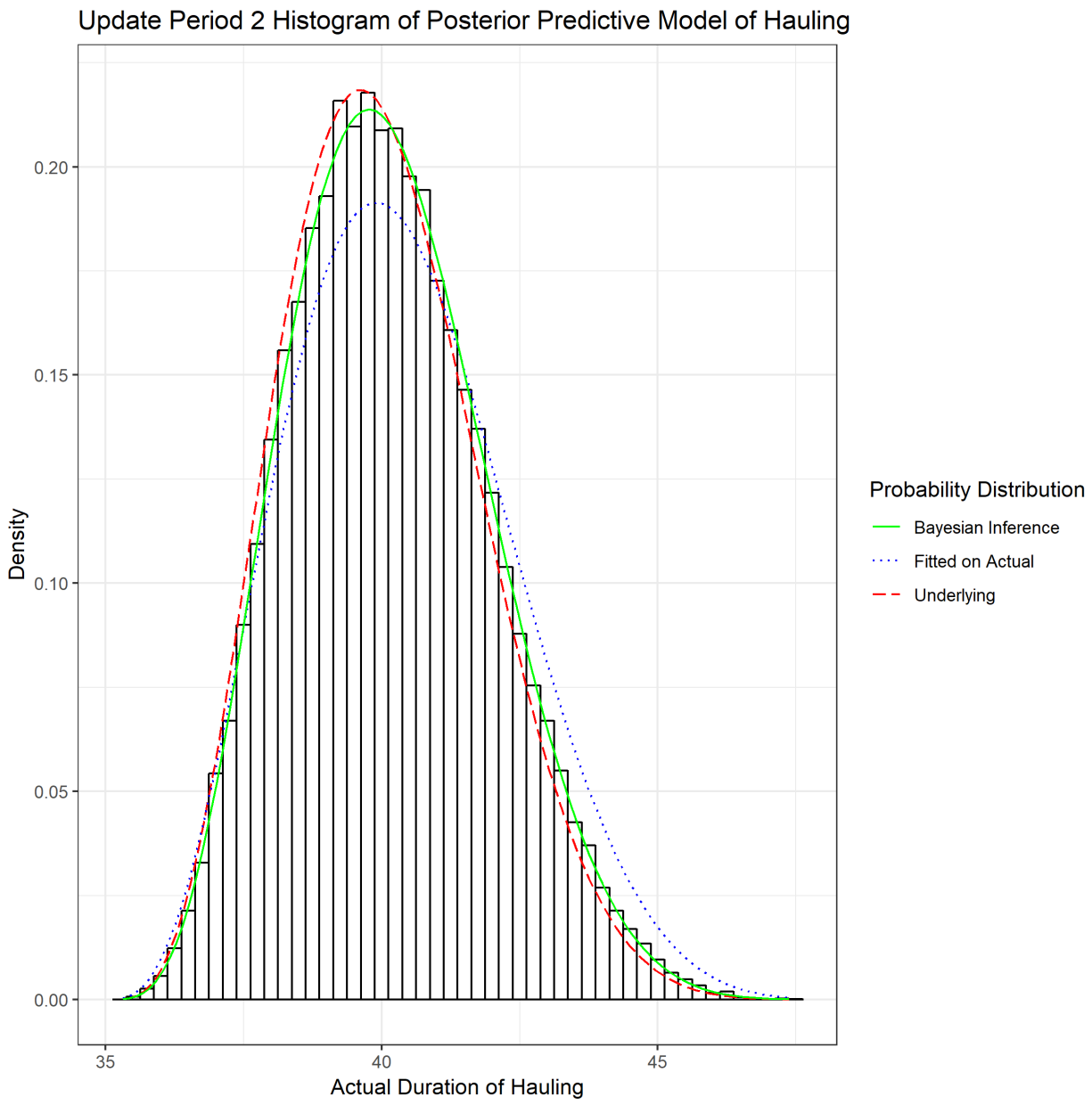


Figure A-15 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (red line), the input model fitted from cumulative observations (blue line), and the updated input model derived using proposed method (green line), for Cycle 2.

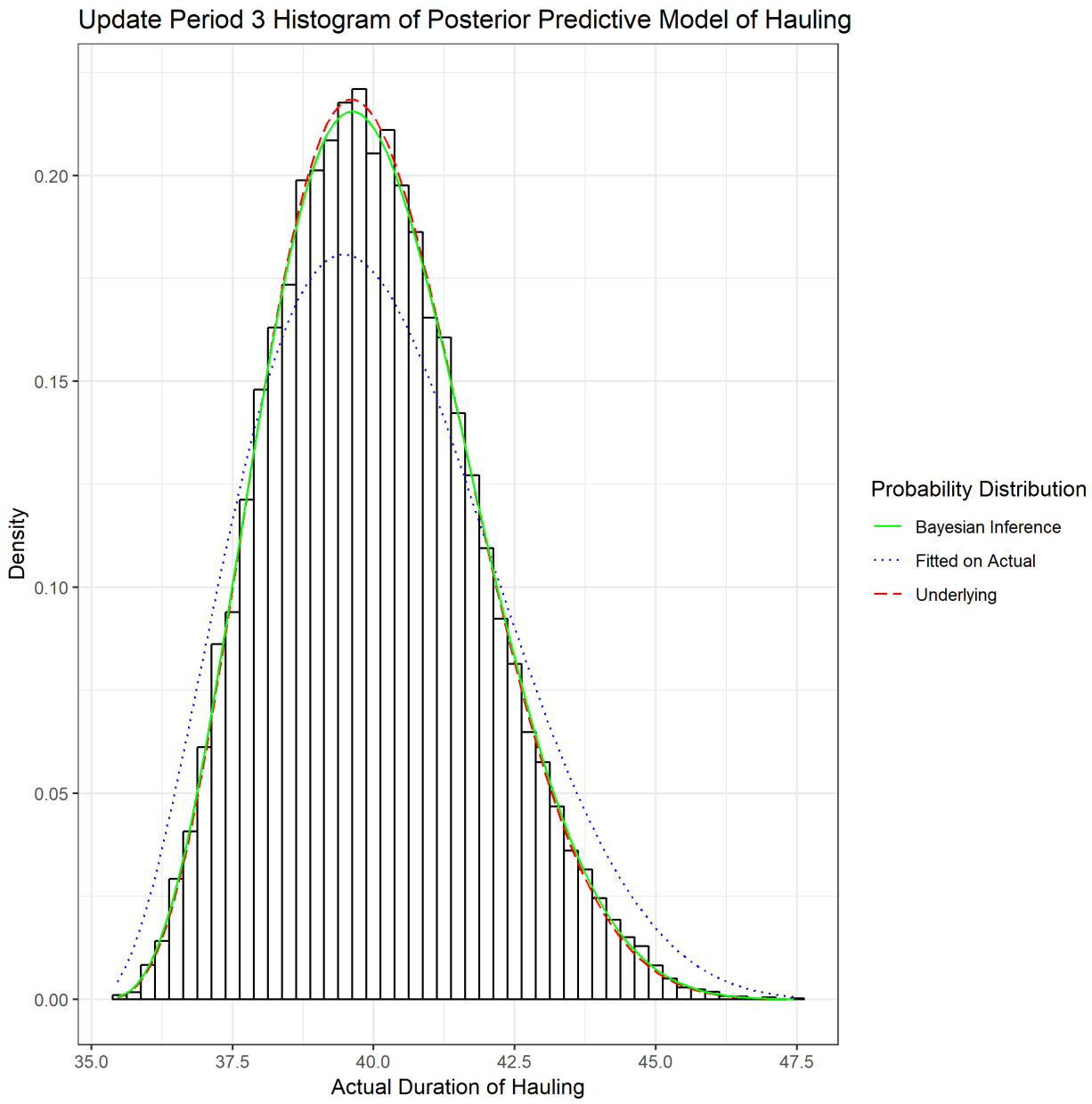


Figure A-16 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (red line), the input model fitted from cumulative observations (blue line), and the updated input model derived using proposed method (green line), for Cycle 3.

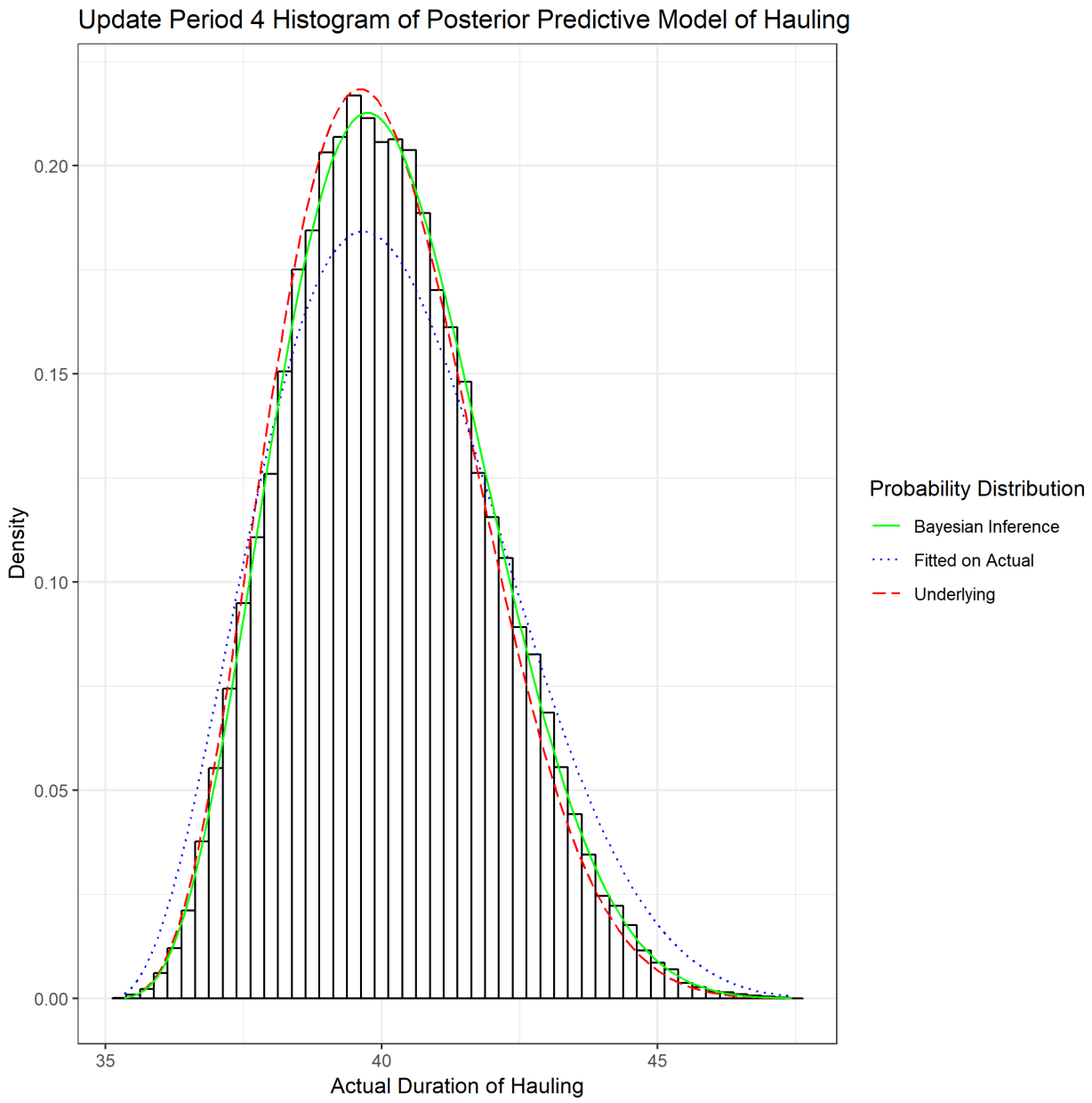


Figure A-17 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (*red line*), the input model fitted from cumulative observations (*blue line*), and the updated input model derived using proposed method (*green line*), for Cycle 4.

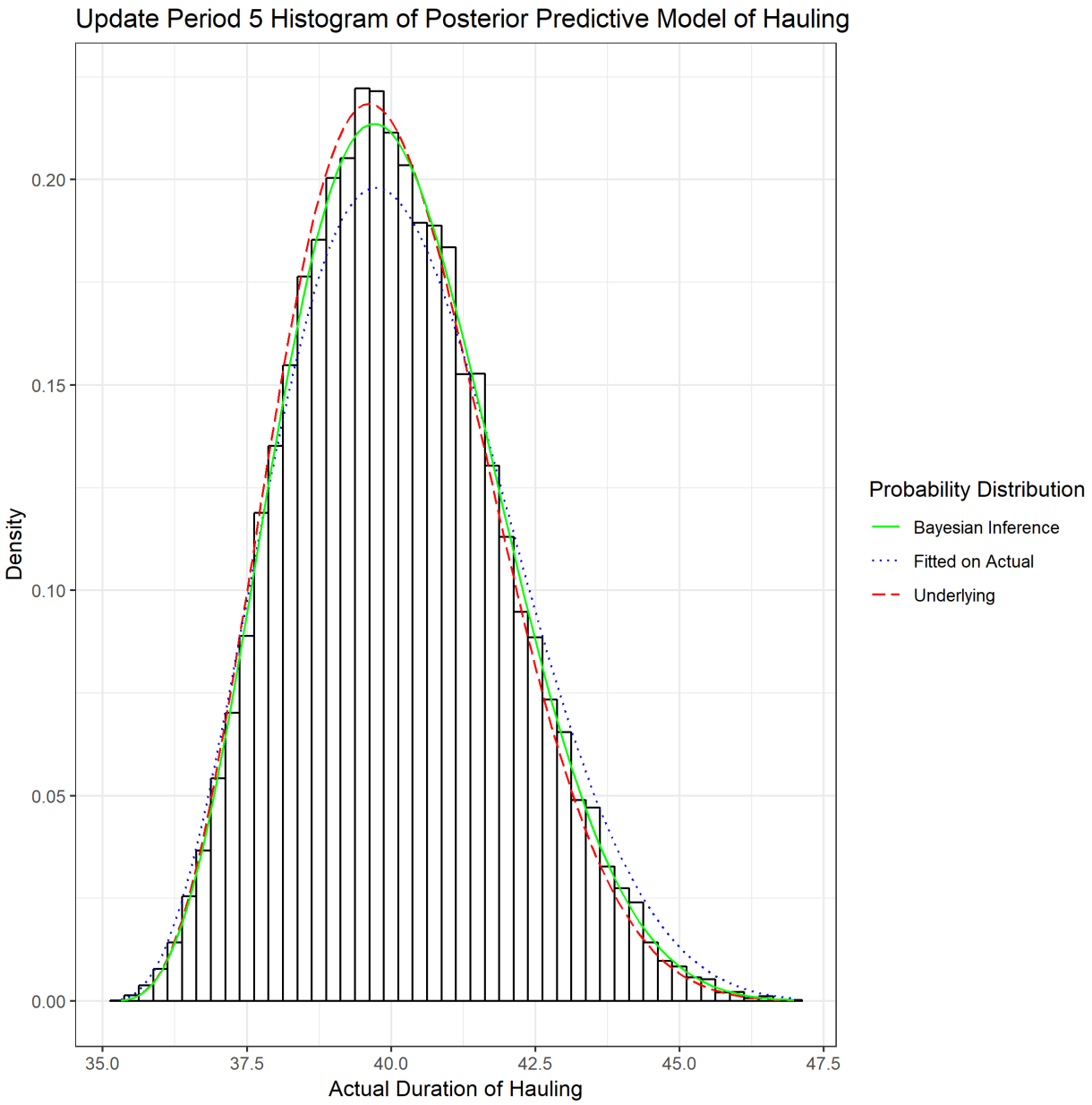


Figure A-18 Histogram of posterior predictive hauling model, as well as the true underlying probability distribution (*red line*), the input model fitted from cumulative observations (*blue line*), and the updated input model derived using proposed method (*green line*), for Cycle 5.

APPENDIX B: CODE OF THE TWO DEVELOPED R LIBRARIES

The *R* library "XiaomoLing/Detect3DRelation" captures the custom function of interval-based 3D objects relationship detection algorithm. The *R* code for this function is listed as follows.

```
#' A 3D Object relationship detection function
#'
#' This function allows you to compare objects saved in two list
#'
#' @param tablefrom, tableto are tables with; closedfrom, closedto
#' @keywords 3D
#' @return A table of combined the input tables with result
#' @export


detecte3Dr <- function (tablefrom, tableto, closedfrom, closedto) {

   #### tablefrom is Check Against List, with 6 column, with column names:
   #### c ("MIN.X","MAX.X", ("MIN.Y","MAX.Y", ("MIN.Z","MAX.Z")
   #### closedfrom is a vector, for example c(T,T)

  tableto$RNO <- seq.int(nrow(tableto))

  # INCLUDED Interval method

  ## X-axis
  ### From is Check Against List
  tablefrom.X <- as.matrix(dplyr::select(tablefrom,MIN.X,MAX.X))

  ### convert to intervals and name each row
  From.X <- intervals::Intervals_full(tablefrom.X, closed = closedfrom, type
= "R")
  rownames(From.X) <- tablefrom$RNO

  ### To is Checking List
  tableto.X <- as.matrix(dplyr::select(tableto,MIN.X,MAX.X))
  To.X <- intervals::Intervals_full(tableto.X,closed = closedto, type = "R")

  ### Result List of X included
  list.X <- intervals::interval_included(From.X, To.X)
  names(list.X) <- c(seq.int(nrow(tablefrom.X)))
  df.list.X <- data.frame(WithinRNO = rep(names(list.X), sapply(list.X,
length)),
                          tabletoRNO = unlist(list.X))
  df.list.X <- tibble::as_tibble(df.list.X)

  ## Y-axis
```

```
  ### From is Check Against List
  tablefrom.Y <- as.matrix(dplyr::select(tablefrom,MIN.Y,MAX.Y))

  ### convert to intervals and name each row
  From.Y <- intervals::Intervals_full(tablefrom.Y, closed = closedfrom, type
= "R")
  rownames(From.Y) <- tablefrom$RNO

  ### To is Checking List
  tableto.Y <- as.matrix(dplyr::select(tableto,MIN.Y,MAX.Y))
  To.Y <- intervals::Intervals_full(tableto.Y,closed = closedto, type = "R")

  ### Result List of Y included
  list.Y <- intervals::interval_included(From.Y, To.Y)
  names(list.Y) <- c(seq.int(nrow(tablefrom.X)))
  df.list.Y <- data.frame(WithinRNO = rep(names(list.Y), sapply(list.Y,
length)),
                          tabletoRNO = unlist(list.Y))

  df.list.Y <- tibble::as_tibble(df.list.Y)

  ## Z-axis
  ### From is Check Against List
  tablefrom.Z <- as.matrix(dplyr::select(tablefrom,MIN.Z,MAX.Z))
  From.Z <- intervals::Intervals_full(tablefrom.Z, closed = closedfrom, type
= "R")
  rownames(From.Z) <- tablefrom$RNO

  ### To is Checking List
  tableto.Z <-  as.matrix(dplyr::select(tableto,MIN.Z,MAX.Z))
  To.Z <- intervals::Intervals_full(tableto.Z,closed = closedto, type = "R")

  ### Result List of Z included
  list.Z <- intervals::interval_included(From.Z, To.Z)
  names(list.Z) <- c(seq.int(nrow(tablefrom.X)))
  df.list.Z <- data.frame(WithinRNO = rep(names(list.Z), sapply(list.Z,
length)),
                          tabletoRNO = unlist(list.Z))
  df.list.Z <- tibble::as_tibble(df.list.Z)

  ## Bind the all result table of XYZ, then filter for count = 3
  df.list.XYZ <- dplyr::bind_rows(df.list.X,df.list.Y,df.list.Z)
  df.list.XYZ <- dplyr::group_by(df.list.XYZ, tabletoRNO, WithinRNO)
  df.list.XYZ <- dplyr::summarise(df.list.XYZ, n=n())
  Within <- dplyr::filter(df.list.XYZ, n >2)


  # OVERLAP Interval method

  tableto.rest <- dplyr::filter(tableto,!RNO %in% Within$tabletoRNO)
  tableto.rest$RRNO <- seq(1:nrow(tableto.rest))

  ## X-axis
  ### To is the rest of the checkList
  tableto.rest.X <- dplyr::select(tableto.rest,MIN.X,MAX.X)
  To.rest.X <- intervals::Intervals(tableto.rest.X,closed = closedto, type =
"R")
```

```
  ### List of X overlapped
  list.rest.X <- intervals::interval_overlap(From.X, To.rest.X)

  names(list.rest.X) <- c(seq.int(nrow(tablefrom.X)))
  df.list.rest.X <- data.frame(OverlapRNO = rep(names(list.rest.X),
sapply(list.rest.X, length)),
                               tabletoRRNO = unlist(list.rest.X))

  df.list.rest.X <- tibble::as_tibble(df.list.rest.X)

  ## Y-axis
  ### To is the rest of the checkList
  tableto.rest.Y <- dplyr::select(tableto.rest,MIN.Y,MAX.Y)
  To.rest.Y <- intervals::Intervals(tableto.rest.Y,closed = closedto, type =
"R")

  #### List of Y overlapped
  list.rest.Y <- intervals::interval_overlap(From.Y, To.rest.Y)

  names(list.rest.Y) <- c(seq.int(nrow(tablefrom.X)))
  df.list.rest.Y <- data.frame(OverlapRNO = rep(names(list.rest.Y),
sapply(list.rest.Y, length)),
                               tabletoRRNO = unlist(list.rest.Y))

  df.list.rest.Y <- tibble::as_tibble(df.list.rest.Y)

  ## Z-axis
  ### To is the rest of the checkList
  tableto.rest.Z <- dplyr::select(tableto.rest,MIN.Z,MAX.Z)
  To.rest.Z <- intervals::Intervals(tableto.rest.Z,closed = closedto, type =
"R")

  ### List of Z overlapped
  list.rest.Z <- intervals::interval_overlap(From.Z, To.rest.Z)
  names(list.rest.Z) <- c(seq.int(nrow(tablefrom.X)))
  df.list.rest.Z <- data.frame(OverlapRNO = rep(names(list.rest.Z),
sapply(list.rest.Z, length)),
                               tabletoRRNO = unlist(list.rest.Z))

  df.list.rest.Z <- tibble::as_tibble(df.list.rest.Z)

  ### Bind the rows of the result table of XYZ, then filter for count = 3
  df.list.rest.XYZ <-
dplyr::bind_rows(df.list.rest.X,df.list.rest.Y,df.list.rest.Z)

  df.list.rest.XYZ <- dplyr::group_by(df.list.rest.XYZ, tabletoRRNO,
OverlapRNO)
  df.list.rest.XYZ <- dplyr::summarise(df.list.rest.XYZ, n = dplyr::n())
  Overlap <- dplyr::filter(df.list.rest.XYZ, n >2)

  ## Combine overlap and included results
  Within.Result <- dplyr::left_join(tableto, Within,by = c("RNO" =
"tabletoRNO") )

  Overlap.R <- dplyr::left_join(tableto.rest, Overlap,by = c("RRNO" =
"tabletoRRNO") )
```

```
  Overlap.R <- dplyr::select(Overlap.R, 8:11)
  Overlap.Result <- dplyr::left_join(tableto,Overlap.R, by = "RNO" )

  Result <- dplyr::left_join(Within.Result, Overlap.Result, by = c("MIN.X",
"MAX.X", "MIN.Y", "MAX.Y", "MIN.Z", "MAX.Z", "Manual", "RNO"))

  return(Result)

}
```

The *R* library "Chrisfufu/LongestCommonSubString" captures the custom function of dynamic programming-based longest common substring algorithm. The *R* code for this function is listed as follows.

```
#' Longest Common Substring
#'
#' it takes two strings
#' @param aString and bString are two strings
#' @return the Longest Common SubString.
#' @import stringr
#' @export
#'



LCStr <- function(aString, bString, minLen){

  LCS = matrix(data = 0,nrow = nchar(aString)+1, ncol = nchar(bString)+1)
  lengthOfSubstring = -1
  finalIndex = -1

  for(i in 1:nchar(aString)+1){
    a<-stringr::str_sub(aString, i-1, i-1)
    for (j in 1:nchar(bString)+1){
      b<-stringr::str_sub(bString, j-1, j-1)
      if(a==b){
        LCS[i,j] = LCS[i-1,j-1]+1
        if (lengthOfSubstring < LCS[i,j]){
          lengthOfSubstring = LCS[i,j]
          finalIndex = i
        }
      }
      else{
        LCS[i,j] = 0
      }
    }
  }

  if (lengthOfSubstring > minLen){
    return (stringr::str_sub(aString, finalIndex-lengthOfSubstring,
finalIndex-1))
  }
  else{
    return('no result')
  }
}
```