Summer 2021

Innovative modeling and management of infrastructure systems, engineering and construction operations, and offsite construction technology using computational data analytics

Rayan Hassane Assaad

Department: Civil, Architectural and Environmental Engineering

INNOVATIVE MODELING AND MANAGEMENT OF INFRASTRUCTURE SYSTEMS, ENGINEERING AND CONSTRUCTION OPERATIONS, AND OFFSITE CONSTRUCTION TECHNOLOGY USING COMPUTATIONAL DATA ANALYTICS

by

RAYAN HASSANE ASSAAD

A DISSERTATION

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

2021

Approved by:

Islam H. El-adaway, Advisor
Joel G. Burken
Kamal H. Khayat
Mohamed A. ElGawady
Cihan H. Dagli

ABSTRACT

The construction industry has been facing considerable challenges due to the inadequacy of the traditional methods in executing, managing, and modeling infrastructure and construction projects. While many techniques have been developed to improve the decision-making process in the industry, there is no evidence of sufficient and continuous improvements in the industry's adoption and implementation of innovative techniques such as new management approaches, modern modeling methods, and emerging computational data analytics. To this end, the goal of this research is to address some of the recent challenges faced in the industry with a focus on infrastructure asset management, construction engineering and management operations, and offsite construction technology. The research goals and objectives were achieved through multiple management, modeling, and computational analytical methods; including artificial intelligence and supervised machine learning algorithms, mathematical and risk modeling, statistical and multivariate time series analysis, clustering techniques and unsupervised data mining algorithms, and surveys and industry panel meetings. The research has numerous intellectual merits, methodological contributions, and practical implications as it addresses critical research areas that have not been investigated before and strengthens areas which needed in-depth examination and further advancements. The findings, outcomes, and conclusions of this research will contribute in further improving the cost, time, productivity, and safety considerations in the industry; leveraging innovative management, modeling, and computational analytics in infrastructure and construction projects; devising data-driven decision-making processes; and administrating and preparing the workforce of the future.

ACKNOWLEDGMENTS

Writing this dissertation has been fascinating and extremely rewarding. First, I would like to express my sincere and my deepest appreciation to my advisor Dr. Islam H. El-adaway–Hurst-McCarthy Professor of Construction Engineering and Management and Professor of Civil Engineering–for his excellent and invaluable guidance and caring; for all the invested time and insightful conversations during the development of this work and the supervision of my dissertation; and for providing me with an excellent atmosphere for doing research. His dynamism, vision, sincerity, and motivation have deeply inspired me, and he has been a tremendous mentor for me both on the professional and personal side. It was a great privilege and honor to work and study under his guidance. Second, I would like to express my gratitude to the dissertation committee members for their valuable comments, suggestions, and engagement, as represented by Dr. Joel Burken, Curators' Professor and Department Chair of Civil, Architectural, and Environmental Engineering; Dr. Kamal Khayat, the Vernon and Maralee Jones Professor of Civil Engineering; Dr. Mohamed ElGawady, Professor of Civil Engineering and Benavides Faculty Scholar; and Dr. Cihan Dagli, Professor of Systems Engineering and Engineering Management. Third, words cannot express how grateful I am to my family who experienced all the ups and downs of my journey. I am extremely grateful and forever indebted to my beloved family for their love, caring, and sacrifices for educating and preparing me for a bright future as well as for providing me with unfailing support. Their continued encouragement was what sustained me thus far. Finally, my thanks go to all people, colleagues, and friends who supported me, directly or indirectly, to complete this research work.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1. OVERVIEW

The construction industry is a significant contributor to different markets as well as to the economic health (Shrestha et al., 2020). Construction is a diverse industry and a project-based sector that includes different markets and project types such as infrastructure, non-residential buildings, mixed-use developments, commercial, residential, and industrial projects, among others (Arocho et al., 2014; Alashwal et al., 2017; Xia et al., 2018). According to the US Bureau of Labor Statistics (2019), the US construction industry employs 4.5% of the total US workforce, which is equivalent to around 7,289,300 workers, and it projects that more construction jobs will be created. Moreover, it is estimated that the construction sector will be one of the fastest growing industries with a predicted growth rate of 4.5% over the next years (Duffy Group, 2018), which makes the construction industry the leading sector in wage and employment growth.

The recent advancements in computational methods and modeling have provided exceptional capabilities to tackle and resolve many challenges faced by the construction industry. In relation to that, Cook (2019) provided that recent use of technologies and computational analytics have changed the landscape of the construction industry and have provided opportunities to leverage innovations and to disrupt the construction sector. In fact, according to Slaughter (1998), innovation can provide multiple advantages including: increase in economic growth, increase in the technical feasibility of construction undertakings, enhanced productivity, improved efficiency, social benefits, market growth, reductions in costs, improved reputation, ease of work, proposition of new solutions to

identified problems or technical constraints, attraction of promising new hires, enhanced competitive advantage, and improved technologies, among others. Although innovation has existed for several decades as a field of study, the focus was mainly on the manufacturing industry rather than on the construction industry; which led to a historical perception that innovation rarely occurs in the construction industry (Slaughter et al., 1998). In fact, innovation in the construction is traditionally been considered informal, unrecorded, and bespoke to one project (Lim et al., 2010). Pellicer et al. (2014) stressed that by stating that innovation has not yet been fully explored in the construction management literature. Also, Holt (2015) stated that innovation in the construction industry is considered ad hoc and project specific.

Construction projects are large, very complex, long lasting, and are created and built by temporary alliance of disparate organizations (Slaughter et al., 1998). In addition, the construction industry is known to be a fragmented sector (Assaad et al., 2020f) where many project participants, including: owners, engineers, architects main contractors, subcontractors, suppliers, and manufacturers, have different obligations towards the successful execution of projects (Assaad et al., 2020g). This has created many challenges, uncertainties, risks, claims, and disputes in the industry (Khalef et al. 2021). In relation to that, Loosemore and Richard (2015) provided that "the construction sector has come under particular scrutiny around the world as being a low-innovation sector." The low innovation in the construction sector has contributed to the fact that construction companies were not able to capitalize on the benefits and pronounced advantages of modern management and modeling approaches. Nevertheless, with all the recent advancements in technologies, computational methods, management techniques, and modeling procedures, it was

inevitable for the construction companies but to invest in these modern and innovative practices and processes. Given that construction is a very diverse industry as it includes multiple sectors and project types, there is no one single way in which innovation occurs (Ozorhon and Oral, 2017). Innovation could be defined as the acquisition, generation, development, and implementation of non-trivial ideas, processes, or systems that are new to the state of the art and that have practical benefits (Lijauco et al., 2020).

According to The National Academies Press (2020), data science is emerging as a field that is revolutionizing science and industries, and the work across nearly all domains is becoming more data driven; which affects processes, operations, jobs that are available, and the skills that are required. Computational analytics is a field of data science that is concerned with computer-based analysis of data and includes approaches and methods such as artificial intelligence, machine learning, data mining, algorithms, statistics, and theory (Charfreitag, 2020). Computational analytics involve solving real-world problems through the capture, management, organization, and visualization of data to inform better decisions, or to embed data into mathematical/statistical models for an automated or optimized decision-making process (Issuu, 2018). That said, there has been a recent interest in leveraging innovative management and modeling techniques as well as computational analytics in different sectors of the construction industry such as infrastructure asset management, engineering and construction operations, and modern/smart construction methods.

Infrastructure networks are one of the most critical construction projects and systems. In fact, cities and their surrounding urban and suburban areas rely on interwoven infrastructure networks such that one entity's function and performance affect its connected

counterparts (Dong et al., 2020). The US infrastructure industry is forecasted to substantially grow over the couple few years. In relation to that, according to GlobalData (2018), the total output value of the infrastructure construction market reached $326.6 billion in 2017 - which is up from $321.2 billion in 2012 - and is expected to rise to $396 billion in 2022, which corresponds to a 3.9% annual average growth rate. The importance of the infrastructure industry is reflected also by the fact that each dollar spent on the US infrastructure brings an approximate economic benefit of up to $2.20 (Congress Budget Office, 2015). Also, the US Council of Economic Advisers has estimated that a $1 billion investment in transportation-infrastructure can support 13,000 jobs for a year (FHWA, 2020).

High-quality infrastructure networks have become a prerequisite for delivering social benefits and achieving sustained economic growth due to the increase in urbanization and population growth (Ruiz and Guevara, 2020). This is especially with the increased demand for and interest in green and sustainable systems in recent years where such initiatives are considered among the most noticeable practices in today's world of applications for design and construction (Assaad et al., 2021b). Sustainable infrastructure involves conceiving, designing, constructing, operating, maintaining, or repairing/rehabilitating the existing infrastructure facilities to maintain or upgrade the existing social, economic, or environmental conditions (Mirza and Ali, 2017). The topic of sustainable infrastructures is very broad and includes multiple elements and aspects including sustainable infrastructure systems. According to Dale and Hamilton (2007), "[t]he importance of sustainable infrastructure to a community and its capacity for innovation is similar to the foundation the human skeleton plays in the overall structuring,

functioning and health of the body." Hence, sustainable development must be encouraged in all phases of a project, and the industry shall move towards more sustainable practices and shall ensure that these practices continuously adapt to emerging changes and needs (Mirza and Ali, 2017).

On the other hand, engineering and construction operations are subject to interruptions due to the inherent nature of the construction environment itself (AbouRizk and Halpin, 1990). Construction operations refer to any construction and engineering activity that contributes to the delivery of construction facilities. Thus, construction operations do not only cover construction field activities or activities on the project level but also activities on the company level. To this end, construction operations include a wide range of aspects, and they are considered complex processes (Gong and Caldas, 2010). Construction operations are influenced by internal factors—that companies usually have some control over them—as well as by external events which are usually not fully controlled by the company (Lee et al., 2010). Moreover, the increasing complexity of many constructed facilities and escalating demands for project performance are driving significant changes in the construction engineering and management operations (Tatum, 2005). In relation to that, examples of construction engineering and management-related operations that are experiencing substantial disruptions include project performance, productivity, and safety. For instance, productivity data has been widely used as performance indicators to evaluate construction operations (Bai et al., 2012). Also, safety is not considered as an add-on to construction operations but rather it is essential to, and an integral part of, operations in the construction industry (Burkart, 2002).

Furthermore, the construction industry is changing at a rapid rate, and new construction methods have emerged. In fact, the current construction operations include hybrid construction processes relying on both traditional onsite construction methods as well as offsite construction technologies. Offsite construction is defined as "the practice of preassembly or fabrication of components both off the site and onsite at a location other than at the final installation location" (CII, 2017). Furthermore, different offsite construction typologies exist including single-trade pre-fabrication, multi-trade pre-fabrication, preassembly, and modularization (FMI, 2018, Jang and Lee, 2018). Numerous studies have shown that the modular/offsite construction market will continue to grow in the coming years. In relation to that, according to Research and Markets (2018), the modular construction market had a value of $112.42 billion in 2018, and this value is expected to reach $157.19 billion by 2023; which corresponds to a compound annual growth rate of 6.9%.

Recent reliance on technologies, innovations, and computational analytics is helping professionals work smarter, not harder, which improves construction productivity and helps in doing much more with less resources (Cook, 2019). Although the construction sector has been recently responding to multiple things like the skilled labor shortage, new sustainability regulations, and advancements in information technology and software (Stannard, 2020), innovations in the construction industry still tend to be less than the innovations in other sectors (Lijauco et al., 2020). Thus, if the construction industry does not embrace innovation, technological, and computational initiatives, it will fall behind other sectors, and it would take several decades for the construction industry to recover, if any. Hence, currently, there are a lot of opportunities that the construction industry could

exploit, and this is the right time for leveraging innovative management and modeling techniques in the construction sector.

1.2. PROBLEM STATEMENT

Despite its substantial contributions to other sectors and the overall economy, the construction industry has been facing considerable challenges for decades. One of the main challenges experienced by the construction sector is the inadequacy of the traditional methods in executing, managing, and modeling construction projects. In relation to that, and while many techniques have been developed to improve the management and the decision-making process in the construction industry, more advanced management and modeling methods are needed (Yang et al., 2019) to solve and address the critical problems that the construction industry is currently facing.

In fact, Lim et al. (2010) provided that there is no evidence of continuous improvements in the industry's adoption and implementation of innovative techniques. This especially seen as related to new management approaches, modern modeling methods, and emerging computational analytics. One of the main reasons behind this is that construction organizations are usually not willing to spend much money on that since it is considered as a cost intensive investment within definite returns (Lim et al., 2010). This has led the construction industry to be traditionally seen as a low-technology sector with low levels of expenditure on aspects related to innovative approaches (Noktehdan et al., 2019). Nevertheless, construction companies have started to realize that there is no other option but to invest in more advanced management and modeling approaches to address the devastating challenges that are being faced by the industry.

One of the critical challenges faced by the construction industry is related to infrastructure asset management (Noktehdan et al., 2019), especially with the inability to devise cost-effective, accurate, and efficient methods. Infrastructure asset management could be defined as the science, the knowledge, and the program to manage infrastructure systems to function in a sustainable, efficient, and effective way (Suprayitno and Soemitro, 2018). While infrastructure asset management is the best approach for balancing growing demands, aging infrastructure, and constrained resources (Flintsch, 2002), infrastructure asset management is still a relatively new discipline and as such lacks well-grounded practices, theories, and innovations (Too, 2010).

Too (2010) adds that "while receiving relatively more interest and attention from empirical researchers, the advancement of this field [infrastructure asset management], particularly in terms of the volume of academic and theoretical development is at best moderate [because] many researchers and practitioners are still unaware of, or unimpressed by, the contribution that asset management can make to the performance of infrastructure asset." Infrastructure asset management systems include innovative and intelligent ways for managing infrastructure maintenance and rehabilitation (Neves et al., 2016). Therefore, governments need to re-evaluate their infrastructure systems and adopt more innovative and sustainable approaches in managing their infrastructure assets (Mirza and Ali, 2017). Innovation is specifically needed for the rehabilitation and repair of infrastructure assets, since infrastructure systems are facing insufficient investment for their maintenance and upgrading (Mirza and Ali, 2017). In fact, according to ASCE (2016), around $4 trillion are needed to repair the current state of the US infrastructure by 2025. In addition, According to Katseff et al. (2020), public-infrastructure spending has fallen, and there is an unfunded

infrastructure gap, or a backlog, of more than $2 trillion in 2016; this figure may currently be an underestimate. In fact, many departments of transportation (DOTs) in the US are reported that they are facing huge challenges in funding their infrastructure projects (Elsayegh et al., 2020). Also, America's deteriorating infrastructure imposes enormous costs on the US economy as well as lower productivity and reduced competitiveness (Meisels, 2020). Therefore, better innovation practices are needed in the management of sustainable infrastructure assets.

Another challenge faced by the construction industry is related to poor and old management of engineering and construction operations. In fact, it has been shown, long time ago, that construction operations are prone to interruptions due to multiple reasons (Remold, 1989). Thus, many previous research efforts have aimed to improve construction operations either from the perspective of the overall construction environment or the construction process (Choy and Ruwanpura, 2005). While the area of construction operations has been an active field of research for decades (AbouRizk and Halpin, 1990), many innovative management and modeling initiatives could be still be leveraged in construction operations. One of reasons behind the inability of previous studies to fully leverage new methods in construction operations could be attributed to the fact that they are subject to multiple interacting factors that produce unpredictable outcomes as well as to stochastic events that are difficult to anticipate (AbouRizk et al., 2011). Also, construction operations are highly diverse, are performed under very different conditions, require many types of resources, and present a range of risks (Tatum, 2005). This is also further magnified by the uniqueness of construction operations and the fact that a model built to simulate one operation cannot be automatically used for a different one (AbouRizk

et al., 2011). In fact, these differences in the construction industry, its products, operations, and technology suggest differences in the process of innovation (Tatum, 1989). While there are diverse construction engineering and management operations, new methods and approaches could still be leveraged in the following construction-related operations: assessment and prediction of project performance (Dulaimi et al., 2005), variables impacting construction productivity (Bröchner and Olofsson, 2012), and enhancing safety performance (Esmaeili and Hallowell, 2012; Assaad and El-adaway, 2021a), among others.

One challenge faced by the construction industry is the reliance on traditional stick-built construction methods which have showed to possess many disadvantages as compared to more advanced, modern, and innovative methods such as offsite and modular construction (Liu et al., 2017). In relation to that, the construction industry has recently started to use offsite construction to boost the performance of their projects and to overcome many challenges experienced on construction projects (Barbosa et al., 2017). In addition, offsite construction is considered as an innovative solution to alleviate workforce issues since it reduces the reliance on the scarce skilled workforce and maximizes labor productivity by relying on automation (Nasirian et al., 2019). Offsite construction brings with it many disruptions to the construction projects in terms of its reliance on technological advancements; manufacturing processes; integrated planning, design, construction; transportation considerations; and supply chain optimization. These disruptions are believed to substantially influence the demand for and the skillset of the construction workforce (Arashpour et al., 2016). Nevertheless, the physical scale of the construction components and of the completed projects establish certain operational constraints that affect the development and use of innovations in the construction industry

(Slaughter, 1998). More specifically, most of the construction activities are still being performed using the stick-built method such that construction works are executed, and components are assembled, primarily at the final location of the project. While offsite construction can bring many advantages and innovative techniques to the delivery and execution of construction projects, innovations in offsite construction are still considered to be limited. In relation to that, Slaughter (1998) stated that innovations that require controlled environments or conditions (i.e. offsite construction) during implementation are limited in their application since offsite fabrication and assembly activities in shops are limited to the space available and to the transportation requirements of the unit.

In summary, while various research efforts have been conducted to address and solve many challenges faced by the construction industry and infrastructure projects, there is still an undeniably room for improvement. While challenges exists in different infrastructure- and construction-related fields, this dissertation focuses on using innovative management methods, modeling techniques, and computational analytics in (1) the safety of dam infrastructures; (2) different construction engineering and management operations including project performance, productivity, and safety; and (3) offsite construction technology. These aspects were chosen due to the substantial knowledge gaps that were found in the existing research streams and current body of knowledge. In relation to that, the knowledge gaps that are identified and addressed in this dissertation are as follows:

1. Knowledge Gap A: Most of the previous research efforts focused on one special structural or geotechnical behavioral aspect of dams while disregarding the important potential hazard aspect. As such, no previous research work has been conducted to develop an artificial intelligence decision tool that can be adopted

quickly and easily to provide accurate forecasts for the hazard potential level of dams in the US.

2. Knowledge Gap B: While many previous research studies attempted to predict the project performance of construction projects, most of the research efforts focused on one aspect of project performance: schedule or cost. In addition, other research studies attempted to forecast both time and cost at completion but without incorporating any direct relation between different project risks and project performance. Further, some studies incorporated possible inputs affecting project performance; however, the utilized inputs fall short of covering the varied risks present in construction projects. As such, no research work has offered an integrated approach to estimate the performance of construction projects. This indicates that the construction industry lacks the formulation of predictive models that factor a wide spectrum of project performance risks. Therefore, a holistic model that incorporates the different project risks that affect project performance in terms of both cost and schedule is needed.
3. Knowledge Gap C: While previous research studies focused on studying labor productivity in the construction industry, no previous research attempted to study the causalities and relationships between dynamic workforce and workplace variables and the productivity of the industry as a whole. Consequently, there is a knowledge gap in the literature in relation to the quantification of the impacts of different workforce and workplace variables on the productivity of the entire construction industry. As such, there is a need to take the earlier research directions a step further by rigorously investigating the relationship between dynamic

workforce and workplace variables and the overall productivity of the construction industry.

4. Knowledge Gap D: Previous safety fatality studies in the construction industry focused mainly on the individual safety fatality factors rather than on analyzing the critical combinations, associations, and interconnectivities between the different fatality causes. Also, there is a lack of studies that have integrated clustering methods and data mining techniques to study the fatal safety accidents in the construction industry in specific. The traditional way of looking at safety incidents focuses on analyzing the weakest link in the chain of events by identifying the only one main cause for the accident and what went wrong that allowed the incident to occur. However, the rigid adherence to this way of thinking can lead to some significant errors in improving safety performance. As such, there is a knowledge gap in the literature in terms of focusing on the individual fatality factors rather than possible combinations and associations between them.
5. Knowledge Gap E: The increased use of offsite construction technologies has intensified the debate on its future and the impact of automation, manufacturing, and robotics on the workforce. While previous research works have focused on different aspects of offsite construction technology, no previous studies were conducted to investigate the impact of offsite construction on the different workforce occupations involved in offsite construction projects. In fact, very little research studies were directed to study the workforce-related aspects of offsite construction. More specifically, no previous study has been performed to study the impact of offsite construction on the onsite and offsite construction workforce, on

different labor characteristics, and on the engineering, construction, and administrative workforce. To this end, there is a knowledge gap in the current body of knowledge as related to helping industry practitioners in prioritizing training needs and programs for the different workforce occupations involved in the offsite construction-related activities as well as in understanding the implications of offsite construction on multiple, rather than specific, workforce-related characteristics.

1.3. RESEARCH GOAL, OBJECTIVES, METHODOLOGIES, AND OUTCOMES

The goal of this research dissertation is to address some of the recent challenges faced in the industry by tackling the previously identified knowledge gaps with a focus on infrastructure systems, construction engineering and management operations, and offsite construction technology.

This research dissertation has 5 main objectives, each objective corresponding to each of one of the identified knowledge gaps.

1. Develop a data-driven model to evaluate and predict the hazard potential level of dams in the US.
2. Create a holistic model to predict project performance in terms of cost and time at completion.
3. Study and model the impact of dynamic workforce and workplace variables on the productivity of the construction industry.
4. Determine the critical combinations of causes or factors leading to safety fatality accidents on construction job sites.

5. Study the impacts of offsite construction technology on the different workforce occupations and on labor-related characteristics.

1.4. RESEARCH PLAN

Although this dissertation has a single goal; it focuses on three research areas: infrastructure asset management, construction engineering and management operations, and offsite construction technology. Each one of these areas could be divided into modules to address the previously presented 5 objectives as shown in Figure 1.1.

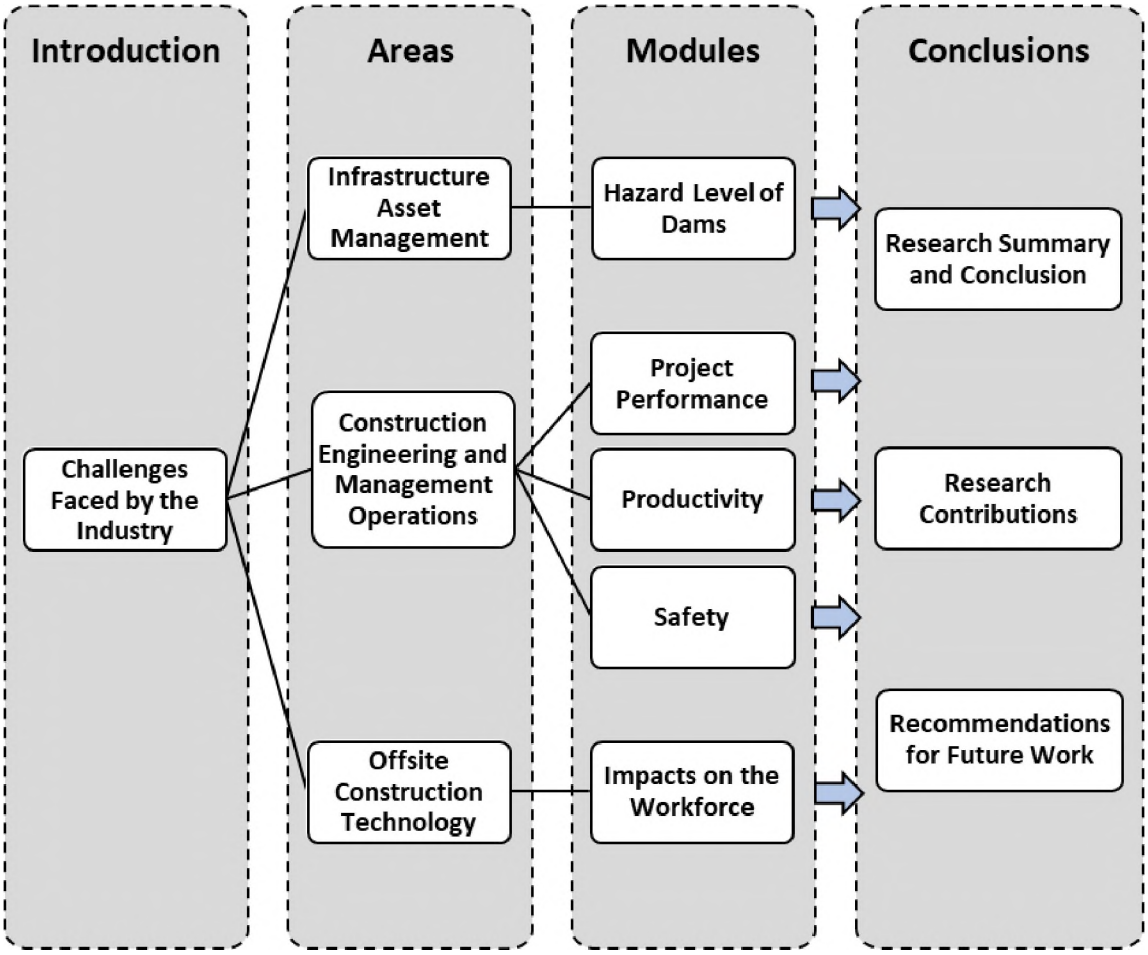

Figure 1.1 Summary of research plan.

Each one of the previously presented 5 research objectives is tackled in a separate section in this dissertation. Figure 1.2 shows each objective and its associated methodology as well as the outcomes.

The first section is an introductory one; discussing the problem statement, presenting the knowledge gaps that need to be addressed, and defining the goal and objectives of this research.

The second section presents a supervised computational artificial intelligence model for the evaluation and prediction of the hazard potential level of dam infrastructures; thus, covering the first objective. The model is developed to equip dam owners and authorities with a valuable data-driven framework that could help dam regulatory organizations to evaluate and predict the hazard potential level of their dams with a good accuracy while minimizing the effort, time, and costs associated with the needed periodic formal inspections of the dams by authorized engineers, as well as to address the deficiency in resources, funding, and staff for dam infrastructures.

The third section proposes a mathematical and risk model for the prediction of project performance in the construction industry; thus, covering the second objective. The model allows practitioners to make better and improved predictions of project performance by incorporating all pertinent and available information on a wide spectrum of project risks. Furthermore, the presented model allows construction professionals to assess the impact of their decisions on the performance of the project and helps them take the appropriate corrective and preventive actions to minimize cost and schedule overruns.

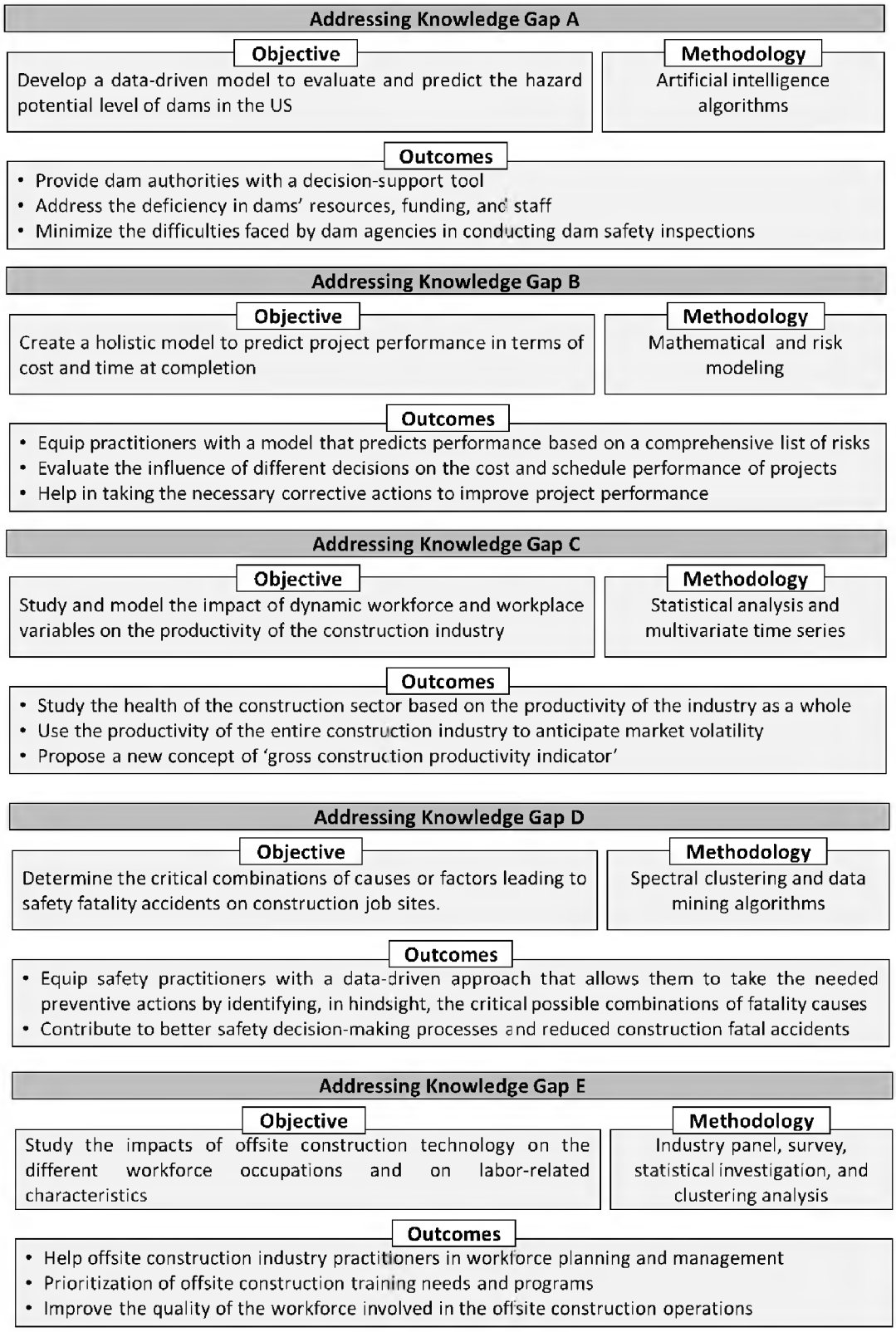

Figure 1.2 Objective, methodology, and outcomes

The fourth section develops a statistical and time series model to study the impact of dynamic workforce and workplace variables on the productivity of the construction industry; thus, covering the third objective. The model opens opportunities in relation to using the productivity of the entire construction industry to anticipate market volatility in the construction industry as well as to identify potential improvements in the overall productivity of the industry.

The fifth section presents an unsupervised model to determine the critical combinations of safety fatality causes in the construction industry; thus, covering the fourth objective. This section equips safety practitioners with a data-driven approach that can take into consideration the fact that, while safety accidents could happen due to factors or causes that are individually critical, fatalities on construction sites could also result due to a combination of factors that might not be perceived to be critical on the individual level but rather become critical when combined with other factors.

The sixth section investigates the impact of offsite construction technology on the different workforce occupations and labor-related characteristics; thus, covering the fifth objective. This section plays a critical role in helping offsite construction industry practitioners in workforce planning and management, in the prioritization of training needs and programs, and in improving the quality of the workforce involved in the offsite construction operations.

1.5. RESEARCH BENEFITS

This research is different from existing efforts as related to the focus, methods, and purpose. After successful completion of the objectives, the research will have significant contributions to the body of knowledge from multiple perspective. First, the research develops a novel data-driven approach for the evaluation of the hazard level of dam infrastructures in the US. This will pinpoint the key variables that dams' authorities shall considered as related to the hazard potential level of dams. Also, this will provide the agencies responsible for the management of dams in the US with a decision-support tool that could be used to accurately predict the hazard potential of their dams. Ultimately, this decision-support tool could be used to address the deficiencies in dams' inspection resources, funding, and staff. Second, the research develops an advanced model for the prediction of project performance for construction projects based on a wide range of risks. This will provide project stakeholders with the capability of evaluating the impact of multiple decisions on the cost and time performance. Ultimately, this will lead to an enhanced decision-making process that helps in identifying the needed corrective actions for an enhanced project performance. Third, the research provides a better understanding of the different variables that affect the productivity of the construction industry. This will equip construction practitioners with a new information that could be used to anticipate market volatility and changing conditions. Ultimately, this provide the opportunity to study the health of the construction industry based on the fluctuations in productivity. Fourth, the research presents a data-driven approach to identify the critical combinations of safety fatality causes in the construction industry. This will equip safety practitioners with a robust method that allows them to take the needed preventive actions by identifying, in

hindsight, the factors or causes that could lead to fatal accidents. Ultimately, this will lead to a better safety performance on the construction job sites and to reduced fatal accidents in the construction industry. Fifth, the research provides construction companies with a better understanding of the impacts of offsite construction on their different workforce occupations that are involved in offsite construction projects. This will help offsite construction industry practitioners in workforce planning and management, in the prioritization of offsite construction training needs and programs, and in improving the quality of the workforce involved in the offsite construction operations. Ultimately, this will help in devising effective workforce strategies that could be implemented for a better leverage of offsite construction technology.

Furthermore, the following 4 key elements play important roles in the infrastructure and construction industry: data, people, technology, and process. The mapping between these elements and the focus of this dissertation is shown in Figure 1.3. The data element refers to the available information and insights, the people element refers to the individuals who are doing the work, the process element refers to how the operations are performed, and the technology element refers to the tools used to function. Obviously, these different elements are connected. For instance, people in general follow processes and can leverage technologies. Also, the use of technologies enables better processes. In addition, data creates value, knowledge, and better understanding of people aspects, process functions, and technology resources.

To this end, the research benefits of the different sections in this dissertation will collectively help in addressing challenges faced in infrastructure asset management, construction engineering and management operations, and offsite construction technology

as well as in enhancing the decision-making process in these areas. The benefits are listed in more depth in each section as well as in the Conclusion section (Section 7.2).

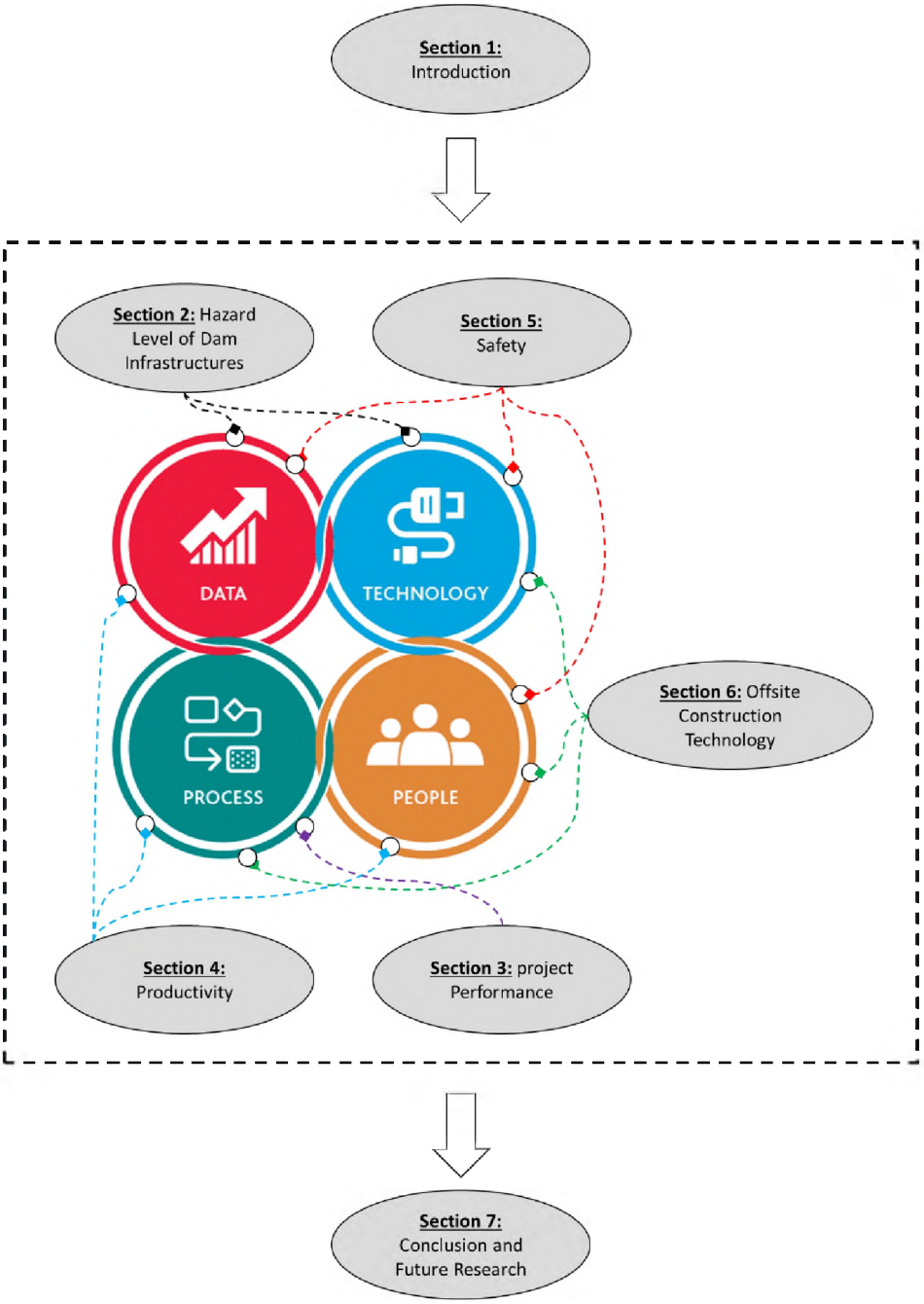

Figure 1.3 Mapping between sections and data, people, technology, and process aspects.

2. SUPERVISED COMPUTATIONAL ARTIFICIAL INTELLIGENCE MODEL FOR THE EVALUATION AND PREDICTION OF THE HAZARD POTENTIAL LEVEL OF DAM INFRASTRUCTURES

2.1. OVERVIEW

Dams are considered a vital and beneficial part of a nation's infrastructure (Fagel, 2011), and they are defined as structural barriers built to obstruct or control the streamflow of water in rivers and streams (Encyclopedia.com, 2019). According to the Association of State Dam Safety Officials (2019a), dam failures have been documented in every US state, and 173 dam failures and 587 incidents (that would have resulted in dam failure if no intervention took place) were reported by state dam safety programs from January 2005 to June 2013. In general, dam failures cause immense property and environmental damages and take thousands of lives, and the potential for deadly dam failure is growing as the nation's dams are aging and the population is increasing (Association of State Dam Safety Officials, 2019a). To address the critical aging high-hazard-potential dams, an investment of $45 billion is needed (ASCE, 2017).

Historical dams' failures in the US were a prelude to a number of federal actions to create FEMA, and Congress to pass Public Law 104-303 to provide the country with a legislatively mandated National Dam Safety Program (NDSP) (FEMA, 2016). The aim of NDSP is to reduce the risk of life and property from dam failure in the US by establishing and maintaining an effective program to achieve national dam safety hazard reduction (FEMA, 2016). That being said, some states developed emergency action plans (EAPs) for their dams to reduce the likelihood of loss of life and property damages resulting from dams' failures (FEMA, 2015).

To achieve the aforementioned NDSP's purpose, the hazard potential level or hazard rating of dams was created, and it is defined as the possible adverse incremental consequences that result from the failure of the dam or misoperation of the dam or appurtenances (FEMA, 2004b). FEMA created the hazard potential classification system to categorize dams based on the probable loss of human life and the potential for economic losses, environmental damage, and/or disruption to lifelines resulting from dams' failures (FEMA, 2004b). This hazard system for dams is considered the common practice among federal and state dam safety offices to categorize a dam according to its potential impacts (Association of State Dam Safety Officials, 2019c). This hazard potential categorization for dams' evaluation is also adopted by most dam safety organizations and inspection professionals, and it is not based on the condition of the dam, but rather solely on the downstream impacts (Eisenbraun and LaRiviere, 2014).

Further, the 2001 Dam Safety Law requires regular dam inspections every 2 years for high-hazard dams and every 4 years for significant-hazard dams (Schalit and Christie, 2011). Multiple states in the US—such as Iowa, Kansas, South Dakota, Missouri, and Mississippi—possess very tight dam safety budgets that could be equivalent to a range of $22–$93 allocated annually per dam, which is very far from covering the costs of inspection or any needed maintenance (Bell, 2017). This has caused some states to chronically be behind in dam safety inspections where high-hazard dams can go longer than several years without inspection. In relation to that, Miller et al. (2012) provided that efforts to monitor and repair levees and dams are piecemeal and drastically underfunded, and hundreds of dams across the country whose failure would put lives in danger are years overdue for inspection.

ASCE (2009) stresses the funding issues as well as the inspection deficiencies by stating that "many state dam safety programs do not have sufficient resources, funding, or staff to conduct dam safety inspections, to take appropriate enforcement actions, or to ensure proper construction by reviewing plans and performing construction inspections. For example, Texas has only 7 engineers and an annual budget of $435,000 to regulate more than 7,400 dams. That means each inspector is responsible for more than 1,050 dams. Worse still, Alabama does not have a dam safety program despite the fact that there are more than 2,000 dams in the state."

The actual hazard potential of dams is determined based on a safety dam inspection conducted by authorized professional engineers. The professional engineer's judgment and common sense must ultimately be a part of any decision on a dam's hazard potential classification (FEMA, 2004). In general, each dam safety inspection report shall document the observations made during the inspection and the engineer's opinion on the condition of the dam, in addition to any relevant information to the safety of the dam including any items requested by the chief engineer before the inspection (Kansas Department of Agriculture—Division of Water Resources, 2019). Further, the engineer should include a list of the hazards and a map showing the location of the identified hazards. That said, the hazard potential is established according to the engineers' evaluation of the potential impact that the dam failure (breach) or misoperation (unscheduled release), should it occur, would have on the upstream and/or downstream areas or at locations remote from the dam (Association of State Dam Safety Officials, 2019c). An example of an upstream condition is the construction of another dam or water conveyance system that would affect the inflow of water into the reservoir, and a downstream condition could be the development in the

dam's floodplain (Brown et al., 1988). Because the performance of a formal inspection requires considerable efforts, time, and cost, this has led multiple states in the US—such as Iowa, Kansas, South Dakota, Missouri, and Mississippi—to possess very tight dam safety budgets that could be equivalent to a range of $22–$93 allocated annually per dam (Bell, 2017). These budgets are very far from covering the costs of inspection or any needed maintenance. This has caused some states to chronically be behind in dam safety inspections where high-hazard dams can go longer than several years without inspection. In summary, it is well reported that many state dam safety programs do not have sufficient resources, funding, or staff to conduct dam safety inspections (ASCE, 2009).

Based on the above, it could be concluded that there is a critical need to provide federal and state agencies with alternatives that are inexpensive, easy to use, and effective to help them in evaluating and predicting the hazard potential levels of their dams. As such, this section of the dissertation addresses this critical need by developing a decision support tool that can be adopted quickly and easily to provide accurate forecasts for the hazard potential levels of US dams.

2.2. OBJECTIVE

The goal of this section of the dissertation is to evaluate and predict the hazard potential level of dams in the US using a comparative approach based on computational artificial intelligence (AI) algorithms. The associated objectives include: (1) investigating the performance of two AI algorithms: artificial neural networks (ANNs) and k-nearest neighbors (KNNs) for the evaluation and prediction of the hazard potential levels of US dams; (2) developing a decision support tool that could be used by the agencies responsible

for the management of dams in the US with the capability to predict the hazard potential with good accuracy; and (3) identifying the best subset of variables that affect the hazard potential level of dams.

2.3. CURRENT STATE OF LITERATURE ON DAMS AND ASSOCIATED LIMITATIONS

Many previous efforts have been carried out to study and analyze the different characteristics of dams from various perspectives. Because new research should build on previous work efforts to convey prospective findings that add to the body of knowledge (El-adaway et al., 2019), this subsection provides an overview of the previous research efforts that studied dams. In relation to that, Wen et al. (2019) performed a comparative study of concrete cutoff walls' responses for earthen dams on alluvium foundations by gathering a data set of 58 cases. Lee et al. (2019b) conducted a seismic deformation analysis of embankment dams using the total-stress approach and presented an assessment of a finite-difference computer program using embankment dam prototypes tested in dynamic centrifuges. In addition, Savage et al. (2019) modeled the erosion and swelling of the sides of transverse cracks in embankment dams through carrying out laboratory tests on five representative soils used in the core of embankment dams. Jamali et al. (2018) developed a spatial multicriteria evaluation decision model to identify suitable sites for underground or subsurface dams for water supplies in an arid watershed using Boolean and fuzzy logic. Moreover, Wang et al. (2018) conducted a seismic dynamic analysis of gravity dams of different heights in the time domain based on the fluid-structuring model. Zhang et al. (2019a) studied the viscous damping and contraction joint friction in underwater explosion-resistant design of arch dams. Further, Saichi et al. (2019) investigated the

effects of rock foundation roughness on the shear strength of dam–rock interfaces and dam sliding stability. Zhou et al. (2019) proposed a new wetting deformation simulation method for core-wall rock-fill dams.

In light of the preceding information, most of the previous research efforts focused on one special structural or geotechnical behavioral aspect of dams while disregarding the important potential hazard aspect. As such, this section of the dissertation bridges this knowledge gap by developing an AI decision tool that can be adopted quickly and easily to provide accurate forecasts for the hazard potential level of dams in the US. Ultimately, it is important to have a data-driven decision support tool that could help regulatory organizations in evaluating and predicting the hazard potential level of their dams for a better allocation of funds and for reducing the time, efforts, and costs associated with the needed periodic dam inspections by authorized engineers.

2.4. GENERAL INFORMATION ON DAMS

This subsection aims to provide general information on US dams. An EAP is a written document that (1) identifies incidents that could lead to potential emergency conditions at a dam, and (2) identifies the areas that can be affected by the reservoir and specifies preplanned actions to be followed to minimize property damage, potential loss of infrastructure and water resources, and potential loss of life caused by the failure or misoperation of a dam (Association of State Dam Safety Officials, 2019b). In fact, the failure of dams could have enormous impacts.

After performing a statistical descriptive analysis of the USACE (2019) data, it was revealed that the top five US states that possess the highest number of dams as a percentage

of the total dams in the US are Texas with 8.03% of the US dams, Kansas with 7.02%, Mississippi with 6.67%, Missouri with 5.90%, and Georgia with 5.82%. The aforementioned states alone contain around 33.4% of the total dams in the US. Figure 2.1 and 2.2 visualize the regulatory information of dams for different US states (shown on the maps) as well as across the entire US (shown as pie charts).

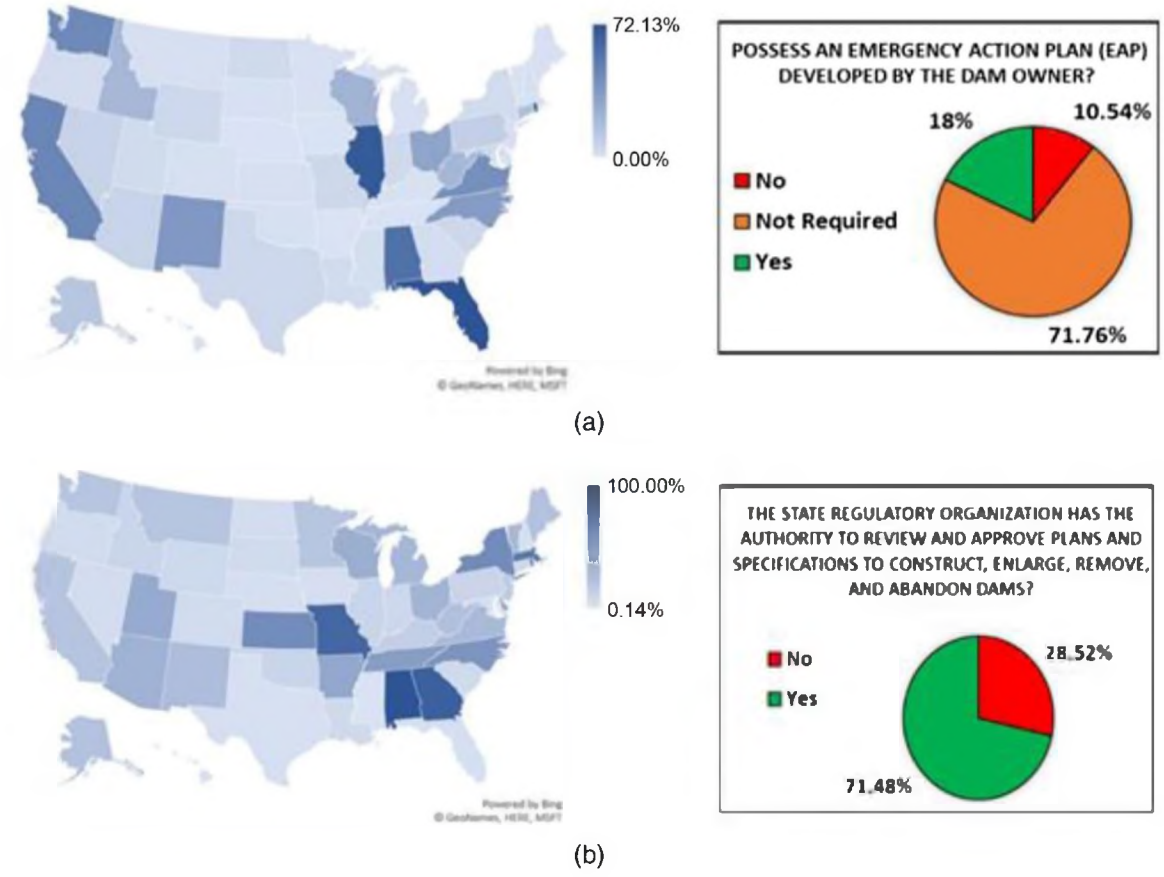

Figure 2.1 Descriptive statistics on dams' EAP and review/approval authority. a) Percentage of dams with no EAP developed by the dam's owner; and b) percentage of dams for which the state regulatory organization does not have the authority to review and approve plans and specifications, construct, enlarge, remove, and abandon dams.

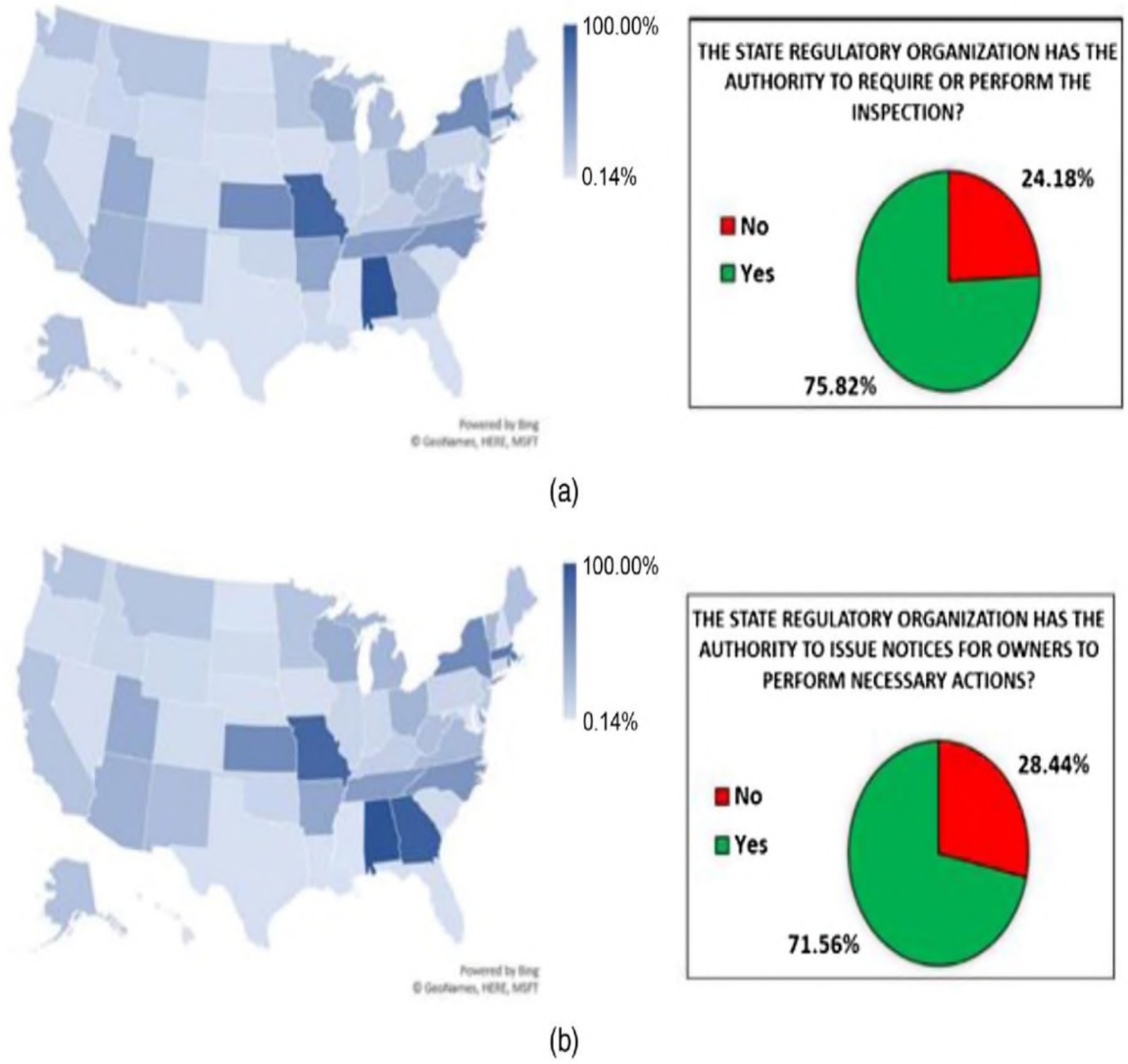

Figure 2.2 Descriptive statistics on dams' inspection authority and notice authority. a) Percentage of dams for which the state regulatory organization does not have the authority to require or perform inspection; and b) percentage of dams for which the state regulatory organization does not have the authority to issue notices for owners to perform necessary actions.

According to Figure 2.1(a), around 71.76% of the dams in the US do not require an EAP, only 18% of the dams possess EAPs, and 10.54% of the dams do not possess EAPs. In addition, Florida and Rhode Island are the two US states that possess the highest percentages of dams not having an EAP at 72.13% and 70.59%, respectively. The lack of

an EAP could be very problematic in the event of a dam failure. Although worst-case scenarios are rare, they have happened in the past, reflecting that an exercised and well-prepared EAP is a valuable tool to help save lives (Tata and Howard, 2016).

For a state to qualify for assistance in the National Dam Safety Program, state appropriations must be budgeted to carry out state legislation. According to FEMA (2013), for a state to be eligible for assistance, the state dam safety program must be working toward meeting several criteria listed in Public Law 109-460, including: (1) the authority to review and approve plans and specifications to construct, enlarge, modify, remove, and abandon dams; (2) the authority to issue notices, when appropriate, to require owners of dams to perform necessary maintenance or remedial work, revise operating procedures, or take other actions, including breaching dams when necessary; and (3) the authority to require or perform the inspection, at least once every 5 years, of all dams and reservoirs that would pose a significant threat to human life and property in case of failure to determine the continued safety of the dams and reservoirs, and a procedure for more detailed and frequent safety inspections.

According to Figure 2.1(b), the state regulatory organization has the authority to review and approve the plans and specifications for 71.48% of the dams in the US. In addition, Alabama and Georgia possess the highest percentages of dams for which the state regulatory organization does not have the authority to review and approve the dams' plans and specifications, at 100% and 91.63%, respectively. Moreover, according to Figure 2.2(a), the state regulatory organization has the authority to require or perform the inspection for 75.82% of the dams in the US. In addition, Alabama and Missouri possess the highest percentages of dams for which the state regulatory organization does not have

the inspection authority, at 100% and 87.10%, respectively. Further, according to Figure 2.2(b), the state regulatory organization has the authority to issue notices for 71.56% of the dams in the US. In addition, Alabama and Georgia, possess the highest percentages of dams for which the state regulatory organization does not have the authority to issue notices for owners, at 100% and 91.63%, respectively.

2.5. MODEL DEVELOPMENT BASED ON ARTIFICIAL INTELLIGENCE ALGORITHMS

The research methodology followed in this section of the dissertation is divided into three main steps: data collection, data preprocessing, and data processing. It is worth mentioning that the followed methodology is slightly different than the one reported in the work of Assaad and El-adaway (2020e). Therefore, slightly different results are obtained in this dissertation as compared to the work of Assaad and El-adaway (2020e).

2.5.1. Data Collection. This section of the dissertation uses the published data on dams in the continental US by the NID, which was populated using the 116th Congressional District information (USACE, 2019). The data were provided by state and federal dam regulators. The NID is maintained and published by the US Army Corps of Engineers. The data contain numerous pieces of information on dams' locations, types, and sizes, among others. A description of the data is provided in the next subsection. The data exist for all dams present in all US states.

According to FEMA (2004b), the hazard potential of dams is rated in the following three categories: low hazard potential, significant hazard potential, and high hazard potential. The description of each hazard potential level is given in Table 2.1.

Table 2.1 Possible dams hazard potential levels in the US. Data from FEMA (2004b).

Hazard potential levels for dams in the US	Description	Loss of human life	Economic, environmental, lifetime losses
Low	Dams where failure or misoperation results in no probable loss of human life and low economic and/or environmental losses; losses are principally limited to the owner's property	None expected	Low and generally limited to owner
Significant	Dams where failure or misoperation results in no probable loss of human life but can cause economic loss, environmental damage, or disruption of lifeline facilities, or can impact other concerns	None expected	Yes
High	Dams where failure or misoperation will probably cause loss of human life	Probable; one or more expected	Yes (but not necessary)

2.5.2. Data Preprocessing. The published NID data includes information on many dams' features and variables as shown in Table 2.2 that reflects the potentially most relevant variables to dams' hazard potential. As shown in Table 2.2, some variables are categorical (such as dam type), while others are numerical (such as distance to nearest city or town); hence, categorical variables shall be converted to numeric data through the data label encoder for efficient processing (Vincent et al., 2019). Since the one hot encoding method substantially increases the dimensionality of the data, categorical variables were transformed to numeric data (Khuriwal and Mishra, 2018) using the label encoder method because it provides fewer errors than other methods (Khan et al., 2019). The label encoder method works by transforming the categorical features to numerical values between 0 and the number of attributes minus 1.

Table 2.2 NID's variables and their descriptions. Data from USACE (2019).

Variable index	Variable or feature	Description	Units
1	Distance to nearest city or town	Distance from the dam to the nearest affected downstream city, town, or village	Miles
2	Dam type	Earth; rock-fill; gravity; buttress; arch; multiarch; roller-compacted concrete; concrete; masonry; stone; timber crib	—
3	Core	Indicates the position and type of watertight member. Position: upstream facing; homogeneous dam; core. Type: bituminous concrete; concrete; earth; metal; plastic	—
4	Foundation	Indicates the material upon which the dam is founded: rock; rock and soil; soil	—
5	Purposes	Indicates the current purpose(s) for which the reservoir is used: irrigation; hydroelectric; flood control and stormwater management; navigation; water supply; recreation; fire protection, stock, or small farm pond; fish and wildlife pond; debris control; tailings; grade stabilization	—
6	Age	Indicates the dam's age from its year of completion	Years
7	Modified	Indicates whether the dam was modified or rehabilitated or controlled	—
8	Dam length	Length of the dam, which is defined as the length along the top of the dam	Feet
9	Dam height	Height of the dam, which is defined as the vertical distance between the lowest point on the crest of the dam and the lowest point in the original streambed	Feet
10	Structural height	Structural height of the dam, which is defined as the vertical distance from the lowest point of the excavated foundation to the top of the dam	Feet
11	Hydraulic height	Hydraulic height of the dam, which is defined as the vertical difference between the maximum design water level and the lowest point in the original streambed	Feet

Table 2.2 NID's variables and their descriptions. Data from USACE (2019). (Continued).

12	NID height	Maximum value of dam height, structural height, and hydraulic height; accepted as the general height of the dam	Feet
13	Maximum discharge	Amount of water the spillway is capable of discharging when the reservoir is at its maximum designed water surface elevation	Cubic feet per second
14	Maximum storage	Maximum storage, which is defined as the total storage space in a reservoir below the maximum attainable water surface elevation, including any surcharge storage	Acre-feet
15	Normal storage	Normal storage, which is defined as the total storage space in a reservoir below the normal retention level, including dead and inactive storage and excluding any flood control or surcharge storage	Acre-feet
16	NID storage	Maximum value of normal storage and maximum storage; accepted as the general storage of the dam	Acre-feet
17	Surface area	Surface area of the impoundment at its normal retention level	Acres
18	Drainage area	Drainage area of the dam, which is defined as the area that drains to a particular point (in this case, the dam) on a river or stream	Square miles
19	Inspection frequency	Scheduled frequency interval for periodic inspections of the dam	Years
20	Spillway width	Width of the spillway available for discharge when the reservoir is at its maximum designed water surface elevation	Feet
21	Volume	Total space occupied by the materials used in the dam structure	Cubic yards
22	Number of locks	Number of existing navigation locks for the project	—
23	Length of locks	Length of the primary navigation lock	Feet
24	Width of locks	Width of the primary navigation lock	Feet
25	Number of separate structures	Number of separate structures associated with this dam project, including saddle dams (or dikes) as defined in FEMA 148 (FEMA, 2004a), as a subsidiary dam of any type constructed across a saddle or low point on the perimeter of a reservoir	—

In addition, since the output variable which is the 'hazard potential level' of dams is a very critical and important aspect of dams, dams with missing as well as undetermined hazard potential level were completely dropped from the dataset rather than imputed using data imputation methods to avoid assigning a value which is not totally precise. This has resulted in a total of 79,470 readings or dams.

The next data preprocessing step was the division of the dataset into 80% training/validation set (63,573 dams) and 20% testing set (15,894 dams). The testing data (15,894 dams) was held out for final evaluation of the developed decision support tool on unseen data to ensure its robustness.

The next data preprocessing steps include data scaling, data imputation, feature selection, and data oversampling. Each one of these steps is discussed in the following paragraphs. Also, it is worth mentioning that these steps were performed within the *k*-fold cross-validation loop.

Starting with data scaling, because the variables and features presented in Table 2.2 have varying magnitudes, units, and ranges (for instance, the dam length is in feet and the volume is in cubic yards), this will create magnitude issues and thus there is a need to bring all features to the same level of magnitudes (Asaithambi, 2017). As such, the min-max data scaling method was used as shown in Equation (1) which rescales the range of features to the range of [0, 1].

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X_{scaled} = scaled value of the unscaled value X; X_{min} is the minimum of the considered sample of the variable X; and X_{max} is the minimum of the considered sample of the variable X.

Further, because of some incomplete or missing NID data points, data imputation was performed to ensure a sufficient and representative amount of data is processed into the AI algorithms (Camm et al., 2019). As such, multivariate imputation was used in this research. It is to be noted that multivariate imputation is different that multiple imputation that was used in Assaad and El-adaway (2020e).

As for feature selection, a subset of the features presented in Table 2.2 shall be used in a way that it includes the variables and features that contribute the most to the predicted variable "hazard potential." This is known as feature selection, which is the process of selecting the most useful features for building models with better generalization ability (Solorio-Fernández et al., 2020). In addition, the importance of feature selection lies in the fact that it: (1) reduces overfitting by minimizing the redundancies in data, (2) improves modeling accuracy, (3) decreases training time, and (4) reduces computational complexity and memory usage (Ding et al., 2018). Wrapper methods are perceived to be more effective than filter methods because they consider the specific interactions between the feature subset search and the learning model (Yan et al., 2019). As such, a wrapper method was used to select the most relevant features. More specifically, the Boruta feature selection algorithm was used because it is considered one of the best ways for implementing feature selection with wrapper methods (Kaushik, 2016).

Since the output variable "hazard potential" included three classes (Low, Significant, and High) which are not represented equally, the dataset is considered as

imbalanced because it has many more instanced of certain classes than others. This is considered to be an issue since classifiers tend to make biased learning model that has a poorer predictive accuracy over the minority classes compared to the majority classes, where test samples belonging to the minority classes are misclassified more often than those belong to the majority classes (Zheng and Jin, 2020). In such cases, different methods are present to help in alleviating such cases including oversampling techniques for the minority classes, undersampling approaches for the majority classes, or a combination of both. In relation to that, the synthetic minority oversampling technique (SMOTE) algorithm was used on the training set to oversample the minority classes as to reduce the imbalance in the samples.

2.5.3. Data Processing. This subsection offers all data processing details.

2.5.3.1. Choice of the computational artificial intelligence algorithms. Many AI predictive algorithms are present in the literature, and it would be infeasible to examine all of them. As such, the widely used algorithms that proved to yield good prediction accuracies compared to other techniques were reviewed. In relation to that, Ryu et al. (2019) and Jebelli et al. (2019) provided the following short list of AI algorithms: KNNs, multilayer ANNs, decision trees (DTs), and support vector machines (SVMs). Although all these AI algorithms were proven to possess good performance based on the domain application, each has its own benefits and drawbacks. In relation to that, one of the major drawbacks of SVMs is their unsuitability for large data sets because they take a considerably long training time (Viswanathan et al., 2019). Given that this section of the dissertation aims to train an AI algorithm for the entire 79,470 dams in the US, SVM is not perceived to be the best algorithm for the scope and goal of this section of the dissertation.

On the other hand, one key disadvantage of DTs is that a small change in the data can cause a large change in the structure of the tree, causing instability (Kumar, 2019). Given that this section of the dissertation aims to develop a robust decision support tool for existing dams as well as for newly constructed dams in the US, DTs are not perceived to be the optimal choice.

KNN is believed to overcome these limitations because it performs well for large training sets (Topak et al., 2018; Richman, 2011) and it is able to make new predictions of unseen data owing to its ability to determine similarity among data (Miner et al., 2015). Furthermore, KNN is simple and easy to implement, does not require the tuning of several parameters, does not make additional assumptions, and is versatile, meaning it could be used for different applications (Horrison, 2018). In fact, the KNN algorithm is considered one of the top 10 AI algorithms (Gou et al., 2019), and it is usually used as the baseline algorithm in many domain problems (Hu et al., 2016). Furthermore, compared to other algorithms, KNN is able to generalize as unseen data with potentially complex geometry (Chen et al., 2019). On the other hand, ANNs are also believed to overcome and address the limitations of SVMs and DTs, among others (Ibrahim et al., 2019), because they are able to model complex problems since they possess the ability to control multidimensional data. In addition, they have the capability to achieve reliable prediction performance by changing their structures and internal information during the training phase to model highly nonlinear systems.

There is no clear-cut answer to which AI algorithm (ANN or KNN) is the best because it is highly dependent on the particularities of the data and problem at hand. The absence of a clear-cut answer has led to an increased research interest in performing a

comparative approach between different AI algorithms to decide which one performs better than the other for a specific application domain. That said and given that there is no previous research that tried to predict the hazard potential level of dams, there is no definite guidance on the best AI algorithm to such application domain. To this end and based on all the previous discussion on the applicability and drawbacks of the different AI algorithms, KNNs and ANNs were used, which is similar to the reasons specified in the work of Assaad and El-adaway (2020c).

Other AI algorithms might perform better than the two investigated algorithms; nevertheless, the work presented in this section of the dissertation is believed to provide the foundation for, and encourage, future research work to try other AI algorithms. This is because the work presented in this section of the dissertation is the first research work that relies on the published NID data set, and thus it would not be reasonable to implement and compare all AI algorithms. At the end, the body of knowledge and the agencies responsible for the management of dams is the US would substantially benefit from the opportunity to investigate the performance of different AI algorithms in future research efforts.

2.5.3.2. K-nearest neighbors. KNN is an AI algorithm that could be used for regression as well as for classification purposes. In this section of the dissertation, KNN is used for classification (prediction of the hazard potential level/class). The output of the prediction is determined as the value or class with the highest frequency from the *k*-most similar instances. In relation to that, the dam's hazard potential is predicted based on the plurality vote of its *k* neighbors, with the dam being assigned to the hazard potential level that is the most common among its *k* neighbors (where *k* is a positive integer that is data dependent). As such, different values of *k* ranging from 1 to 40 were investigated in this

section of the dissertation, and the best value was chosen based on the highest average accuracy on the validation set using grid search. On the other hand, to determine which of the *k* instances in the training set are most similar to the new input, a distance measure or function is used. In relation to that, there exist different distance measures with the two main ones being the Euclidean distance and the Manhattan distance. Using Python's and sklearn's K Neighbors Classifier, this could be tuned using the power parameter *p* where if p = 1 then this is equivalent to using the (l_1) Manhattan distance, and if p = 2 then this is equivalent to the (l_2) Euclidean distance. It is worth mentioning that for any other arbitrary *p* values, then the (l_p) Minkowski distance could also be used, but this was not considered in this section of the dissertation. In other words, only $p = 1$ and $p = 2$ are investigated.

In addition, two weighing methods for the KNN algorithm exist: the uniform method and the distance method (Scikit-learn, 2019). The uniform method equally weighs all points in each neighborhood, whereas the distance method weighs the points by the inverse of their distance such that closer neighbors have a greater influence than those that are further away. These two weighing methods were considered. In relation to that and in addition to the distance functions and *k* value, these weighing methods were considered as parameters for the grid search as to choose the one(s) with the highest accuracy.

2.5.3.3. Artificial neural networks. ANNs are one of the most widely used AI algorithms that can recognize the pattern between input(s) and output(s) (Dawood et al., 2018). The structure of ANNs includes individual parameters or weights for each node in the network's architecture, transfer or activation function that shapes the generated outputs or results, and learning rules that calculate the relative impact of the individual inputs (Mishra et al., 2017; Assaad and El-adaway, 2020b). ANNs rely on adjusting individual

parameters or weights by error minimization through learning from experience (Avci and Abdeljaber, 2016). The adaptive moment estimation (known as Adam) solver was used to train the ANNs because practice has proved that it is better than other adaptive learning solvers (Li et al., 2019). Adam is an algorithm that can be used in place of the classical stochastic gradient descent (SGD) to update the weights in an iterative way based on the training data. It refers to a stochastic gradient-based solver (Scikit-learn, 2019) proposed by Kingma and Ba (2014). Further, Adam works well on relatively large data sets in terms of both training time and validation score (Scikit-learn, 2019).

The configuration or architecture of ANNs highly affects their performance (Negnevitsky, 2005; Keller et al., 2016). As such, this section of the dissertation investigates a multilayer perceptron with different configurations in terms of the number of hidden layers and number of neurons. The architecture of ANNs is generally composed of (1) an input layer that accepts input signals and redistributes them to all neurons in the hidden layers; (2) one or more hidden layers, which include computing neurons for the processing of the input patterns; and (3) an output layer that accepts output signals from the hidden layers and reports the output pattern of the entire architecture (Negnevitsky, 2005). The number of neurons in the input layer is equal to the number of input variables and features, and the number of neurons in the output layer is equal to the number of output variables (three in this section of the dissertation: low, significant, and high hazard potential levels). As such, the determinants of an ANN's architecture are the numbers of hidden layers and the number of neurons. In this section of the dissertation, the architecture of an ANN is represented as $(N_1, N_2, \ldots, N_n)$, where n is the number of hidden layers and N is the number of neurons in the corresponding hidden layer. The number of hidden layers and the

number of neurons in each hidden layer were considered the parameters for the grid search for the selection of the best parameters.

In general, the numbers of hidden layers and their neurons that provide the ultimate prediction accuracy are determined through trial and error, and they are problem dependent (Zhang et al., 2019c). However, well-established guidelines exist to help researchers narrow their search space. That said, the guidelines provided by Heaton (2008) were used as shown in Table 2.3. Based on Table 2.3, two, three, four, five, and six hidden layers were investigated with different neurons within each one because they can represent arbitrary decision boundary and complex representations.

Table 2.3 Number of hidden layers and their descriptions. Data from Heaton (2008).

Number of hidden layers	Description
None	Only capable of representing linear separable functions or decisions
1	Can approximate any function that contains a continuous mapping from one finite space to another
2	Can represent an arbitrary decision boundary to arbitrary accuracy with rational activation functions and can approximate any smooth mapping to any accuracy
> 2	Additional layers can learn complex representations (sort of automatic feature engineering)

For the number of neurons within each hidden layer, there are no clear-cut answers to what the appropriate number of neurons is because it is dependent on the problem and the data at hand (Chao and Kim, 2019). In fact, the number of neurons in the hidden layer(s) determines the complexity of the ANNs. In other words, when the number of neurons is small, the associated ANN is considered to be simple, and when the number of neurons is large, the associated ANN is considered to be complex. While simple ANNs are faster and

more efficient, complex ANNs are believed to better represent sophisticated relationships. As such, both simple and complex architectures for ANNs have their own advantages. To this end, the best way to choose the number of neurons within the hidden layer(s) is to consider a range of neurons that is able to cover both simple and complex architectures. That said, 10 cases for the number of neurons were investigated within each hidden layer: 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. While the number of hidden layers and neurons is determined through trial and error of all possible combinations (Zhang et al., 2019c), only equal numbers of neurons were considered in each one of the hidden layers to reduce the training time. For instance, for two hidden layers, the following numbers of neurons were considered: (1,1), (2,2), (3,3), (4,4), (5,5), (6,6), (7,7), (8,8), (9,9), (10,10). The same applied if there are three, four, five, or six hidden layers. In relation to that, and since it is not feasible to report the accuracies for all the combinations of hidden layers and neurons, this section of the dissertation reports the obtained highest accuracy among all these combinations. Other hyperparameters were also tuned, including the L_2 penalty (regularization term) parameter with possible values of 0.0001 and 0.05, the size of minibatches with possible values of 32, 64, 128, 256, and the initial learning rate with possible values of 0.1 and 0.001. Not many values were tried for these hyperparameters so that to reduce the training time.

2.5.3.4. Cross validation, hyperparameters' tuning, and model evaluation. This subsection aims to present all details pertaining to the performed steps in relation to cross validation, hyperparameters' tuning/selection, and model evaluation. The exhausted grid search method was used to loop over the hyperparameters' values for the developed AI algorithms (both KNN and ANN) because it is the most widely used approach

(Chowdhury et al., 2019). Grid search is the process of building a model on each parameter combination and iterating through each combination accordingly to select the optimal parameters of the given model (Zhao and Jiang, 2019). In addition, *k*-fold cross validation was performed on the training data set by splitting it into *k* folds (or *k* parts) where the model is trained on *k* minus 1 folds and its accuracy is reported or validated on the remaining cross-validation set or fold to obtain the optimal parameters for the AI algorithms. This is repeated *k* times as to ensure that the cross-validation set was run on the entire spectrum of the data set. In other words, the five-fold cross validation was used where the training set was divided into five equal folds, with the first fold being used as a cross-validation set and the other four folds being used to train the AI algorithms. In this first iteration, the accuracy on the cross-validation set (CV1) is reported. In the second iteration, the second fold is used as a cross-validation set and the four other folds (50,858 dams) are used to train the AI algorithms. In this second iteration, the accuracy on the cross-validation set (CV2) is reported. This is repeated five times to ensure that the cross-validation set was run on each of the five folds. The results of the fivefold cross-validation methods are five accuracies (CV1, CV2, CV3, CV4, and CV5) as shown in Figure 2.3.

Once the cross-validation process has been repeated *k* times, as shown in Figure 2.3, the average accuracy over the *k* rounds shall be calculated (Singh, 2019a), and it is used as the basis for optimal parameter selection (Hammerla and Plötz, 2015). The used *k*-fold cross-validation set or method in this section of the dissertation is perceived to be better than other methods because it ensures the generalization and robustness of the created decision support tool (Hong et al., 2018) and it avoids overfitting in the predictive AI algorithms (Chemchem et al., 2019).

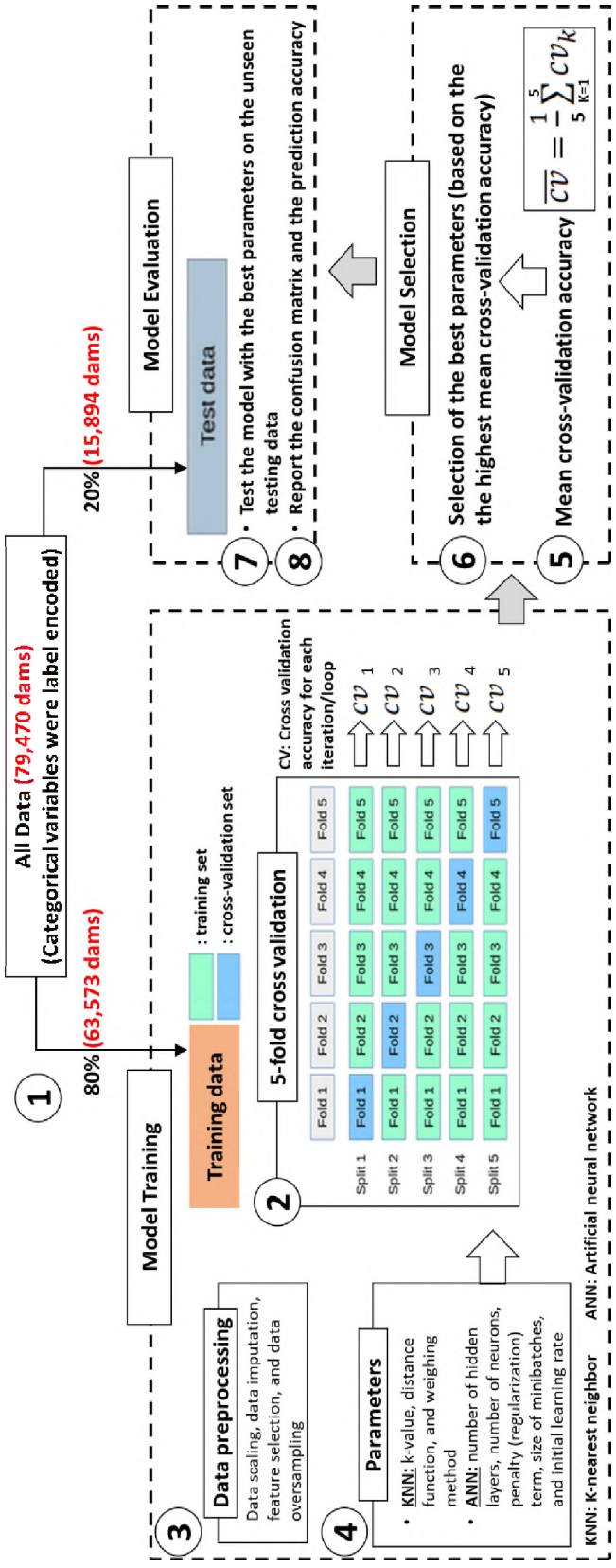

Figure 2.3 Simplified steps in relation to data division, model training, model selection, and model evaluation.

Once the optimal parameters are selected based on the highest mean k-fold cross-validation accuracy obtained using the previous steps, the model with the identified optimal parameters is evaluated on the unseen testing data set (15,894 dams) for final testing of the created decision support tool. In relation to that, the confusion matrix is obtained, and the accuracy is calculated because it is the most widely used method for accuracy assessment among the various discussed approaches in the literature (Mukherjee, 2019).

2.5.4. Coding and Software Packages. The data preprocessing and processing on the collected data were performed using Python, which is an interpreted, high-level, general-purpose programming language. In addition, the predictive AI algorithms were coded using Project Jupyter, which is a nonprofit organization created to develop open-source software, open standards, and services for interactive computing across different programming languages including Python. Moreover, Python's scikit-learn library, which is a free machine learning library, was used to develop the predictive AI algorithms. Additionally, Python's open-source NumPy library was used to add support for large, multidimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on the arrays. Pandas library was used for data manipulation and analysis, and Matplotlib was utilized for data plotting and visualization.

2.6. RESULTS AND ANALYSIS

This subsection provides the obtained results and analyzes them.

2.6.1. Model Selection and Evaluation. The KNN AI algorithm was applied to predict the hazard potential of dams. Based on the 5-fold cross validation and the grid search approach described previously, the optimal value of k was obtained to be 2 because

it corresponded to the highest average accuracy on the validation set. In a similar way, the best other KNN's parameters were determined where the optimal power parameter for the Minkowski metric was p =1 which corresponds to the Manhattan distance, and where the optimal weighing method was uniform method. In relation to that, the best KNN model yielded a highest mean accuracy of 82.63% on the validation set. On the other hand, the ANN AI algorithm was also assessed to predict the hazard potential of dams. The highest average accuracy of 77.86% on the validation set was obtained for the following hyperparameters: an alpha value of 0.05, a batch size of 128, a hidden layer and neurons architecture of (10, 10, 10), and an initial learning rate of 0.001.

Comparing the obtained highest mean accuracies for KNN (82.63%) and for the ANN (77.86%), the best performing model is the KNN. In relation to that, and to evaluate and validate the performance of the selected best model on unseen dams' data, the prediction accuracy of the model was computed on the testing set. The obtained confusion matrix on the unseen testing set is presented in Figure 2.4.

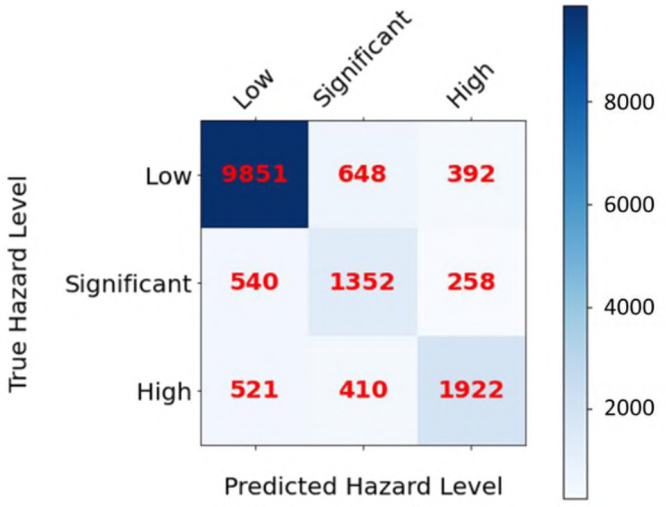

Figure 2.4 Confusion matrix of the developed decision support tool.

Figure 2.4 shows that the highest numbers exist on the diagonals of the confusion matrix where the rows represent the true hazard potential and the columns represent the predicted hazard potential using the developed AI decision support tool. This reflects that most of the predictions are correct. The total prediction accuracy of the model is 82.58%.

2.6.2. Feature Selection. The Boruta algorithm was applied to select the best features subset that is relevant to the hazard potential level of dams in the US. The obtained results are shown in Figure 2.5.

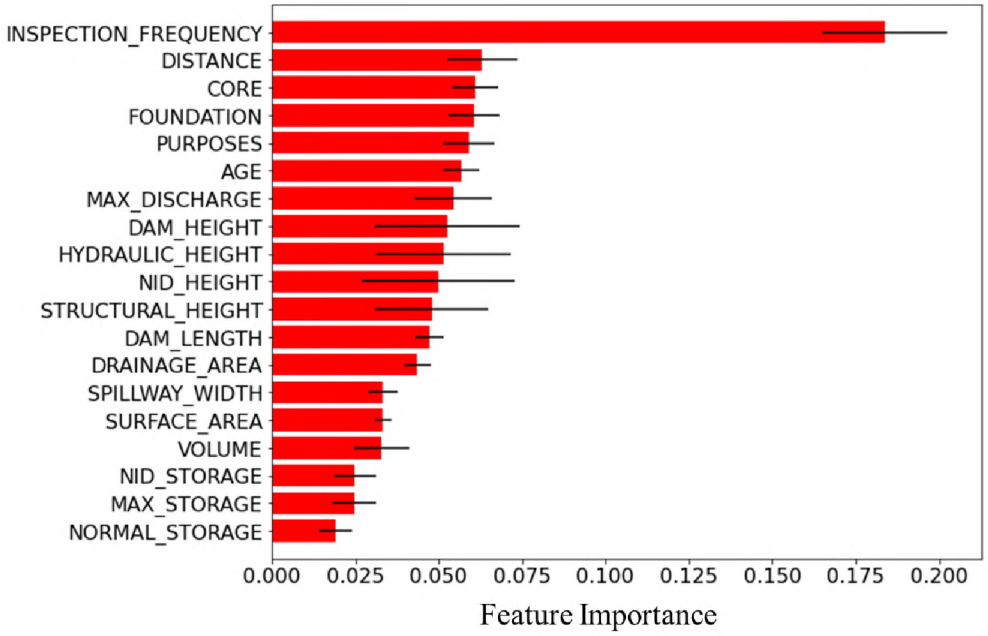

Figure 2.5 Results of the Boruta algorithm for feature selection.

As can be seen from Figure 2.5, a total of 19 input variables or features was selected to be the best subset relevant to the dams' hazard potential. As can be seen from the obtained feature selection results in Figure 2.5, the variable 'inspection frequency' has the

highest weight. This is not to be mistaken for the formal dam's inspection that the specialized or authorized engineers perform to determine the potential hazard level of the dam because this section of the dissertation aims to find an inexpensive and easy-to-use decision support tool as an alternative to the expensive formal inspection, and thus it is not logical to have formal inspection as a candidate variable for the feature selection. That said, the most important variable is the inspection frequency, which is the scheduled frequency interval for periodic inspections of the dam in years as defined by the USACE (2019) in Table 2.2.

2.7. SUMMARY

This section of the dissertation developed a data-driven AI-based decision support tool to evaluate and predict the hazard potential of dams in the US. The adopted research approach has: (1) identified the best subset of 19 variables that affect the prediction of the potential hazard level of US dams; (2) investigated the performance of two AI techniques, ANN and KNN, for the evaluation and prediction of hazard potential levels of dams; and (3) developed a decision support tool that could be used by the agencies responsible for the management of dams in the US with the capability to predict the hazard potential with good accuracy. The obtained results reflected that the KNN algorithm yielded better accuracy compared to the ANN algorithm. This section of the dissertation contributes to the body of knowledge by devising a data-driven framework or decision support tool that is valuable for dam owners and authorities because it could be used to evaluate and predict the hazard potential level of their dams with good accuracy while minimizing the effort, time, and costs associated with formal inspection of the dams. Future work could include the use of

advanced techniques for the management of dam infrastructures such as optimization and photogrammetry (Fayek et al., 2020a, b, 2021).

2.8. RELATED APPENDIX

Appendix A presents the used data and Python code for the developed supervised computational artificial intelligence model for the evaluation and prediction of the hazard potential level of dam infrastructures.

3. A MATHEMATICAL AND RISK MODEL FOR THE PREDICTION OF PROJECT PERFORMANCE IN THE CONSTRUCTION INDUSTRY

3.1. OVERVIEW

To ensure the successful execution of projects, many control systems and methods were developed and implemented to enable project managers to assess the deviations in time and cost from the established target objectives. The project performance in the construction industry is significantly affected because projects are susceptible to several uncertainties and are very complicated in nature (Flanagan and Norman, 1993; Mills, 2001). In fact, the construction industry is ripe for disruption because the large asset class projects are up to 80% higher than the planned budget and require 20% more time than scheduled (Agarwal et al., 2016). In addition, only an estimated approximately 2.5% of companies successfully complete 100% of their projects (Betz, 2018). Further, despite the negative project performance in terms of cost and schedule overruns, only 28% of companies use project performance techniques (PMI, 2017a). Therefore, equipping companies with effective tools to predict their projects' cost and schedule overruns earlier in the project life cycle is crucial for the successful execution of construction projects.

Schedule and cost overruns in construction projects have become an industry-wide status quo. This is reflected by a study conducted by KPMG (2015) showing that only 31% of construction projects were performed within 10% of the budget, and only 25% of projects were completed within 10% of their original deadlines. Moreover, many studies linked the poor project performance in the construction industry to the inefficiencies in existing project performance assessment and prediction techniques. For instance, PMI (2018) reflected that "the traditional measures of scope, time, and cost are essential but no

longer sufficient in today's competitive environment." Further, Olawale and Sun (2010) stated that "despite the wide use of these methods [project control techniques] and software packages in practice, many construction projects still suffer time and cost overruns." Ibadov (2016) stated that "more than half of the owners of construction projects came in contact with at least one problematic project despite the belief that they apply the right tools to plan and control their projects." In addition, Samuel (2017) provided that "it is time to reassess the approach to effective performance of major construction projects."

Triggered by the increase in global competition within the construction industry and the rapid advancements in technologies, most construction companies have started to direct substantial efforts to improve project control (Kivilä, et al., 2017). The changes are being experienced at the internal and external project levels, thus, leading to a growing interest in project performance prediction, monitoring, and control (Kim et al., 2003). In addition, construction projects possess a unique nature that is considered to create many risks and uncertainties (Vaagen et al., 2017). The risks generate numerous scenarios in which one or more project objectives (such as time and cost) can be affected by uncertain events or conditions (Tereso et al., 2018). Although construction projects are being executed, for construction companies to possess tools that allow for the evaluation, assessing, and prediction of the project performance is crucial, including project parties being able to effectively act to avoid or limit such slippages in the case of cost or time overruns (Hazır, 2015).

Many previous research studies highlighted the urgency to develop holistic models for predicting the performance of construction projects. For instance, Hoffman et al. (2007) emphasized the need for a comprehensive model to more accurately predict the

performance of projects. Olawale and Sun (2013) underlay the necessity of having improved cost and time control models that factor practitioners' needs, requirements, and current issues in practice. Terry and Lucko (2012) highlighted that an urgent research need exists to create powerful and comprehensive tools to model projects' time and cost performance. Leon et al. (2018) stressed that project managers strive to make reliable predictions about project performance; however, they are faced with the complexity of a performance prediction process that includes many indices that need to be modeled. As such, a comprehensive framework that factors the key risks that affect project performance is substantially needed in the construction industry. To this end, this section of the dissertation fills this knowledge gap by developing a framework using an unprecedented methodology that incorporates a list of project risks extracted from a previous study.

3.2. OBJECTIVE

The goal of this section of the dissertation is to create a holistic approach to evaluate project progress and predict its expected cost and time at completion by incorporating a wide spectrum of project risks. The associated objectives are to (1) quantify the impacts of the risks related to project performance in terms of cost and schedule; (2) formulate a holistic assessment model; and (3) correlate the developed system to predict cost and time at project completion. To this end, this section of the dissertation uses mathematical and risk modeling to devise explicit equations for the different project risks that affect the performance of construction projects to understand their behavior and predict the schedule and cost overruns in the construction industry. The outcomes will equip project stakeholders with a framework that evaluates the performance of construction projects

based on a list of project risks. In addition, the developed model empowers project managers to evaluate the influence of different decisions on the cost and schedule overruns of projects. In other words, the proposed framework enables project administrators to take the necessary corrective actions to improve the performance of their projects. Eventually, such actions will include decreasing project delays and avoiding unnecessary costs.

3.3. BACKGROUND ON PERFORMANCE MEASUREMENT

A number of techniques have been developed to assess the performance of construction projects or to improve the quality of the decision-making process. Two of the leading approaches are earned value management (EVM) and risk management (RM), which stand out from other techniques because they could be applied in an integrated manner (Hillson, 2004). EVM is a prevailing technique in project management for globally monitoring projects by measuring the performance and progress of the works (Fleming and Koppelman, 2002). Traditionally, EVM combines schedule, scope, and resource management (PMI, 2013); however, the most common uses for EVM provide forecasts of project performance problems in terms of cost and schedule. EVM illuminates the current status of the project and where the project is going relative to where it was supposed to be and where it was perceived to be heading (PMI, 2005). That is, EVM utilizes a performance measurement baseline for which changes from this baseline are tracked to assess the progress of the project. As such, EVM enables project managers to assess the status of the project at different points during the project lifecycle and, consequently, effectively manage projects and programs (Kwak and Anbari, 2010). Nevertheless, the EVM technique metrics are believed to possess some drawbacks, such as "they are based on [a]

monetary unit and not on time. They can behave in ways that are not normally expected of schedule indicators and predictors. Furthermore, it is also possible that an earned value analysis may show that the project is delayed; on the contrary, the project would be on time" (KhodaBandehLou et al., 2016).

The RM technique shares common aspects with EVM in the sense that it equips decision makers with important information when considering management strategies. Boyadzhieva-Georgieva (2014) also underlined this concept and provided that "both Earned Value Management (EVM) and Risk Management (RM) are directed toward solving the same problem [which is] measuring a project's performance by providing information that is used for a basis for making informed decisions and taking actions." Nevertheless, the major difference between RM and EVM is that the latter is based on previous project information and extrapolates these data to understand the project's future trends; however, the former focuses on the project's unknown future performance to identify risks and recommend early actions to minimize or limit the impacts of identified risks (APM, 2008). This point was stressed by Babar et al. (2017), who stated that "[RM] in nature is forward looking, whereas EVM forecasts the future performance in the form of an estimate at completion (EAC) based on past data. The basic limitation of EVM is its sole reliance on past performance without taking into account the strength of [RM] of being able to foresee future uncertainty."

3.4. CURRENT STATE OF LITERATURE ON PROJECT PERFORMANCE AND LIMITATIONS OF EXISTING PREDICTIVE MODELS

Many previous research studies attempted to predict the project performance of construction projects from various angles and by using different modeling techniques. For

instance, Ling et al. (2008) utilized multiple linear regression to develop a model that predicts the project performance in China using project management practices adopted by foreign architecture, engineering, and construction firms based on data collected from 33 projects. Lee et al. (2004) used the discriminant function analysis method to develop user-friendly interface software that predicts the performance of construction projects based on six practices: pre-project planning, constructability, project change management, design/information technology, team building, and zero accident techniques. Attalla et al. (2003) proposed a framework for infrastructure reconstruction projects based on 54 projects using statistical analysis and artificial neural networks and an identified list of critical factors that affect the performance of reconstruction projects. Kim and Reinschmidt (2011) presented a probabilistic method to forecast costs using Bayesian inference and the Bayesian model average technique. Babar et al. (2017) focused on estimating the cost at the completion of projects by risk integration. Du et al. (2016) concentrated on improving the prediction accuracy of the cost at completion in the construction industry using a Markov chain simulation. In contrast, Chang and Yu (2018) developed metrics to predict the completion time of projects in the construction industry. Moreover, Rudeli et al. (2017) used Markov chain models to predict possible deviations in projects' schedules and future progress. Jarkas (2016b) proposed a framework to predict project duration using the time-cost model and multiple linear regression based on 113 residential and 74 office buildings. In contrast, Mortaji et al. (2015) developed cost and performance indices for projects using a change point analysis to estimate the final cost and duration of projects. Lipke et al. (2009) predicted project outcomes based on data from 12 projects. Leon et al. (2018) utilized system dynamics and proposed a model for predicting project performance in the

construction industry based on eight indices. Chen (2014) used a linear modeling approach to increase the accuracy prediction of cost and time at completion using collected data on 131 sample projects. Ling et al. (2004) utilized multiple linear regression to develop a model to predict project performance based on data collected on 84 building projects. In addition, Leung et al. (2017) employed artificial neural networks to predict the project performance of architecture, engineering, and construction projects based on 108 project cases. Ko and Cheng (2007) proposed a model that predicts the success of construction projects using artificial intelligence.

These studies indicate that most of the research focused on one aspect of project performance: schedule or cost. This focus was stressed by Vanhoucke (2012), who reflected that although many tools were developed to predict project performance in terms of time and cost, most of the research focused on the cost aspect of projects. In addition, other research attempted to forecast both time and cost at completion but without incorporating any direct relation between different project risks and project performance. Further, some studies incorporated possible inputs affecting project performance; however, the utilized inputs fall short of covering the varied risks present in construction projects. In addition, other efforts developed models based on a limited number of data points on real construction projects. Moreover, other research studies employed existing modeling tools, such as system dynamics, linear regression, Markov chain, and others, to predict either negative or positive project performance but rarely both. As such, no research work is believed to have offered an integrated approach to estimate the performance of construction projects. This fact indicates that the construction industry lacks the formulation of predictive models that factor a wide spectrum of project performance risks or indices.

Therefore, a holistic model that incorporates the different project risks that affect project performance in terms of both cost and schedule is needed.

3.5. METHODOLOGY

As is subsequently detailed in the forthcoming subsections, a multistep interdependent methodology based on mathematical and risk modeling was used to achieve the research goal and objectives.

3.5.1. Step 1: Data Collection. This step provides the definitions of the terminologies used in this section of the dissertation and the methodology followed for quantifying project risks.

3.5.1.1. Definitions. According to Marle and Vidal (2016), the main result of the risk analysis process is to prioritize risks in terms of their criticality. Criticality is defined "quantitively" as the product of the probability and impact of each risk. Although no unified "qualitative" meaning of criticality exists, many studies attempted to define it based on its application. For instance, Stamatelatos et al. (2011) assigned the criticality rank of each risk based on (1) probability, (2) magnitude of the consequence, (3) the point at which the risk first surfaced, (4) the magnitude of uncertainties, and (5) the amount of time available to react. Rah et al. (2016) defined criticality as whether the entire system will fail, given that a part of it fails.

To this end, this section of the dissertation qualitatively defines criticality by reflecting the agglomeration of (1) the magnitude of uncertainty/probability of the event for which higher uncertainty means higher criticality; (2) the magnitude of the consequence(s) in the case that an event takes place for which higher magnitudes mean

higher criticality; (3) amount of time available to react when a shorter time indicates a higher criticality; and (4) the contribution of event failure toward the failure of the entire project in terms of cost and time for which the higher the contribution, the higher the criticality. This qualitative definition of criticality is used as the basis on which the users utilizing the developed model assess the criticality of the 25 identified project risks. In other words, the criticality of each project risk is entered as an input to the model after being assessed in terms of the magnitude of uncertainty, the magnitude of the consequences, the amount of time available to react, and the failure contribution.

Additionally, in this section of the dissertation, parametric fitted distributions refer to theoretical distributions existing in the literature. These distributions are determined by the relevant parameters; for instance, the normal distribution is defined by the mean and the standard deviation, and the beta distribution is defined by minimum, maximum, alpha1, and alpha2 (Clemen and Reilly, 2013; Palisade, 2018). In contrast, this section of the dissertation refers to nonparametric distributions for the manually fitted functions; these distributions do not possess predetermined parameters and, thus, could not be defined by parameters—instead, they are defined by explicit equations, such as polynomial functions. These definitions were devised from the literature in which a parametric function follows a theoretical distribution (Wilks, 2011). Thus, it considers that the data can be adequately modeled by a probability distribution that has well-determined parameter(s) (Geisser and Johnson, 2006). In contrast, the nonparametric and empirical distribution terms are used interchangeably (de Melo Mendes and Lopes, 2004) for which the data on the entire range of values is used to fit a cumulative distribution (Vose, 1996).

3.5.1.2. Quantification of project risks. In order to quantify the risks associated with the performance of projects in terms of cost and schedule, a list of project risks affecting the execution of construction projects is needed. According to research by Abotaleb and El-adaway (2018), 25 project risks that impact project performance were identified based on an extensive review of the literature. The literature used to identify the project risks included papers that are peer-reviewed, published in archived scholarly journals (i.e., journal articles), and directly related to construction projects. These project risks were identified based on a meta-analysis of the literature that studied the key factors affecting construction project performance. Moreover, Abotaleb and El-adaway (2018) highlighted the lack of and need for a comprehensive model that integrates the 25 identified project risks for the holistic management of construction projects. To this end, Abotaleb and El-adaway's (2018) identified project risks were used in this section of the dissertation because they were shown to be the most important based on a meta-analysis of the literature.

Because the term risk has been defined in numerous ways (Perrenoud et al., 2017), it is important to have a consistent nomenclature for the 25 identified project risks to ensure consistency when respondents fill in the survey. In other words, to ensure that the entered ratings of the respondents/users are consistent in using the Likert scale throughout the 25 project risks, the meaning of each of the project risks was extracted from Abotaleb and El-adaway (2018) but was modified to reflect the negative impacts on schedule and cost. The negative nomenclature of project risks was used to be "align[ed] with the common usage of the word risk… as the extent and impact of adverse occurrences causing a construction project to exceed its predicted budget or cost plan sum" (Adafin et al., 2016). This point

was also stressed by Khodeir and Mohamed (2015), who stated that "risk [is] generally recognized among those within the construction industry as the phenomenon of continually facing a variety of situations involving many unknown, unexpected, frequently undesirable and often unpredictable factors." Accordingly, the utilized negative nomenclature of project risks is only to ensure consistent terminology throughout the 25 risks (instead of having some risks reflecting positive impacts and others reflecting negative impacts). This unified nomenclature avoids creating confusion for the respondents when they provide their individual ratings because they do not have to change their understanding of the used Likert scale from one project risk to the other. Worth mentioning is that another used but less common definition of risk is one that is similar to that of ISO 31000 (ISO, 2018) for which a risk includes the effect of uncertainty on project objectives as a deviation from what is expected, with such an effect being positive or negative. Nevertheless, the negative nomenclature of project risks has no relation with the predicted project performance because the developed model can predict negative as well as positive project performance in terms of schedule and cost overruns, as provided in subsequent subsections. This is possible because, although entering the criticality value of each project risk, the respondent or user provides a high value if the corresponding project risk has a negative impact on the schedule and/or cost of the project and a low value if the project risk has a positive impact on the schedule and/or cost. Table 3.1 summarizes the project risks and their meanings.

After determining the project risks that affect project performance, a survey was developed to establish the probability of occurrence (Pr) and impact (Im) for each project risk; calculate the criticality (Cr) for each project risk; and, ultimately, fit distribution functions for each project risk based on the calculated Cr.

Table 3.1 Project risks impacting project performance.

Project risk	Meaning
Unrealistic scheduling	Failing to: recognize and incorporate uncertainties in duration estimation; add contingency buffer to activities; and determine proper logical sequencing of activities; among others.
Inappropriate schedule pressure	Failing to: take action when the project is behind schedule or when the project is on schedule but needs to be accelerate; assess the progress of the project compared with the planned progress; take the appropriate pressure action such as hiring new staff, using overtime, or adding shifts; among others.
Complexity	Includes: activities that are highly interdependent, overlapping, and complex (in terms of the required skill); and considerable concurrency between engineering and execution; among others.
Inefficient coordination and communication	Failing to: have good coordination and communication between owner, engineer, and contractor; abide by the client's progress-reporting demands and progress meetings; have a continuous review of the system definition and its required functionality; and have continuity in building information modeling (BIM) between the general contractor and the subcontractors; among others.
Deficiency in approval process	Includes: delay in the approval of changes; in replying to requests of information; or in replying to the contractor's queries/requests in general; among others.
Lack of trust and motivation	Failing to have: mutual trust between and within the parties internally (owner payment to contractor, contractor delivery on time, owner payment to workers and engineers); and incentives to increase the motivation of the staff; among others.
Ripple effects of schedule pressure	Includes: prolonged working hours; fatigue and decline in morale; reduced productivity; and increased errors; among others.
Unproductivity of workforce	Includes: inappropriate productivity rate; among others.
Inadequate constructability reviews	Failing to have: the contractor involved in the design stage to ensure that the designed works are constructible with minimal interruptions and costs related to the construction method; and the end users involved in reviewing and revising the project specifications early on; among others.
Incompetent resource development	Failing to have: experienced and reliable staff; and training for inexperienced staff; among others.

Table 3.1 Project risks impacting project performance. (Continued).

Inaccurate resource allocation	Failure to: allocate the available human resources to the tasks at hand, either engineering or execution; and accurately determine the needed human resources based on the project's performance and the time remaining; among others.
Absenteeism and turnover	Includes: high rate of absenteeism and turnover; among others.
Workplace congestion	Includes: use of more resources than what is required; and overmanning or overcrowding effect; among others.
Unsuitable overtime and added shifts	Failing to: use appropriate overtime and added shifts to make up for delayed progress or to accelerate work; among others.
Inferior technology	Failing to: use advanced technology in engineering (three-dimensional modeling and BIM); execution (modern construction equipment and automated construction methods); or management (electronic integrate management systems); among others.
Rework in execution	Includes: mistakes discovered during the execution that need to be reworked; and any rework that is made due to intended changes in design and not necessary due to mistakes; among others.
Rework in design	Includes: mistakes in design that require producing new drawings for already-made designs; and rework in drawings that are due to intended changes in design; among others.
Unreliability of quality assurance staff	Includes: delays by quality assurance staff to check and approve executed works; and high percentage of falsely approved erroneous works that are discovered later in the project; among others.
Out-of-sequence work	Includes: work performed out of its intended logical sequence; among others.
Controlled change	Includes: changes made intentionally by the parties, such as change orders, variations, and changes in construction sequence; among others.
Uncontrolled change	Includes: changes made as a reaction to external risks, such as weather conditions, unforeseen site conditions, and market fluctuations; among others.
Low fabrication quality	Includes: errors in the fabricated items and the quality approval of such items; among others.

Table 3.1 Project risks impacting project performance. (Continued).

Poor communication with fabricators	Includes: delay in ordering and delivery time; and inappropriate communication with vendors and fabricators; among others.
Unsound financial estimating	Failing to: estimate the cost of change and the earned value at any point in time; and consider financial limitations with respect to managerial decisions; among others.
Unreasonable budget contingency	Failing to: have appropriate contingency amounts and maneuvering through the project costs; among others.

To collect reliable results, a unified scale for Pr and Im needed to be communicated to the respondents. As such, because this section of the dissertation utilizes data published by Construction Industry Institute (CII), the five-point Likert scale present in CII's International Project Risk Assessment Implementation Resource (CII, 2013) was adopted in this section of the dissertation. This adoption reduces the subjectivity because the communicated Likert scale complies with objective and standard scaling methods embraced by practitioners in the construction industry. By doing so, it was possible to minimize the respondents' biases associated with the possibility of having different understandings of the scale. The used scales for Pr and Im are presented in Table 3.2 and 3.3, respectively.

After collecting Pr and Im for each project risk and each respondent, the corresponding c was calculated by multiplying Pr and Im. Worth mentioning is that, although one of the limitations of the product between Pr and Im is that it cannot reflect possible correlations between different factors, this method is still of great value in reflecting the criticality or score of each risk factor. This criticality or score is reflected by

its common use in the construction/project risk management field to assess, prioritize, and manage project risks. In fact, the Project Management Institute's PMBOK guide (PMI, 2017b) recommends this approach by stating that "where numeric values are used [for the probability and impact of risks], these can be multiplied to give a probability-impact score for each risk, which allows the relative priority of individual risks to be evaluated within each priority level." Dikmen et al. (2018) also stressed this point as follows: "risk assessment based on probability-impact (P-I) ratings is the most widely used approach in project-based industries such as the construction industry." In addition, Marle and Vidal (2016) stated that "the main output of risk analysis is prioritization of risks, often as a function of their criticality [with] criticality is often defined by the product of P [probability] and I [impact]." Based on the effectiveness of this method, many previous construction management studies provided exceptional knowledge and guidelines by calculating the risk score as the product between probability and impact; some of these efforts include Castro-Nova et al. (2018), Haider et al. (2016), Casanovas et al. (2014), Mostafavi et al. (2013), Russell et al. (2013), and Chan et al. (2011), among others.

Table 3.2 Used scale for Pr. Data from CII (2013).

Descriptions for communicated Pr	Ranges for communicated Pr
1. Very low probability and occurs only in exceptional circumstances	<10% chance
2. Low chance and unlikely to occur in most circumstances	10% <chance<35%
3. Moderate chance and will occur in most circumstances	35% <chance<65%
4. High chance and will probably occur in most circumstances	65% <chance<90%
5. Very high chance and almost certain and expected to occur	90% or greater chance of occurrence
NA- Not applicable to this project	NA

Table 3.3 Used scale for Im. Data from CII (2013).

Descriptions for communicated Im	Effect of communicated Im
1. Negligible and routine procedures sufficient to deal with consequences	5% < increase in cost or time
2. Minor and would threaten an element of the function	5% < increase in cost or time <10%
3. Moderate and would necessitate significant adjustment to overall function	10% < increase in cost or time <20%
4. Significant and would threaten goals and objectives; requires close management	20% < increase in cost or time <50%
5. Extreme and would stop achievement of functional goals and objectives	50% or greater increase in cost or time

3.5.2. Step 2: Fitting of Parametric Distribution Functions. This step provides the methodology followed for the normalization of data and the distribution fitting of the different project risks.

3.5.2.1. Normalization of data. Because both Pr and Im have a scale of 1–5, the scale for the calculated Cr for each project risk is 1 to 25. For a criticality ratio, the calculated Cr was normalized to a scale of [0, 1], which is equivalent to a scale of [0%, 100%]. The Min-Max normalization technique was used because it transforms a variable into a range of [0, 1]; is suitable for variables with known bounds (minimum and maximum); and retains the original distribution of the variable (Jain et al., 2005; He et al., 2010). Accordingly, the normalized Cr of each respondent was calculated using Equation (2).

$$\text{Normalized Cr} = \frac{\text{Cr}_i - \text{Cr}_{min}}{\text{Cr}_{max} - \text{Cr}_{min}} \quad (2)$$

where Cr_i = calculated Cr of respondent *i* on a scale of 1–25; Cr_{min} = minimum possible value of Cr, which is 1; and Cr_{max} = maximum possible value of Cr, which is 25. Worth noting is that, in this model, the average values of the normalized Cr for each project risk as calculated from the survey are referred to as "the reference point."

Worth mentioning is that the Min-Max normalization technique does not use the respondents' maximum and minimum values only. In fact, it uses each collected data point (irrespective of it being maximum or minimum), which is reflected by the Cr_i term in Equation (2). The minimum and maximum values are only used to normalize the obtained values from a 1 to 25 interval to a [0, 1] or [0%, 100%] interval. This use is also indicated from the fitted continuous distributions on the [0, 1] interval in subsequent subsections. For instance, for a collected Cr_i of 16 (which is neither the minimum value nor the maximum value) on the 1–25 scale, its corresponding normalized Cr using Equation (2) is (16−1)/(25−1)=(15/24)=0.625 on the [0, 1] scale.

3.5.2.2. Distributions fitting. Considering "Cr" as a continuous random variable with possible values denoted as "c" representing the criticality of each project risk, a distribution of "Cr" for each project risk is needed. Palisade statistical package was used to fit the normalized Cr to the appropriate continuous parametric distribution functions. This package allows fitting of continuous parametric distribution functions from a pool of possible distributions (Palisade, 2018). To assess the adequacy of the fitted distributions, the Anderson-Darling (AD) test was used because it (1) tests whether the data follow a specified distribution with a significant level (or p-value), (2) detects small variations of any one parameter between two distributions in a reliable manner; (3) identifies differences at the extreme ends of distributions in a reliable manner; and, (4) works well even for small

sample sizes (Engmann and Cousineau, 2011). The most commonly used p-value of 0.05 was adopted for the AD test (Andersson and Burberg, 2015). Because the p-value does not exist for all parametric distributions, the fitted parametric functions were assessed by starting with the distributions that best fit the data. Each distribution is assessed based on its p-value; if the p-value does not exist, the next best-fit distribution is assessed and so on until a satisfactory p-value is observed. Once a satisfactory p-value is reached, the previously assessed distributions are reconsidered only if no p-value was observed for these distributions. In the event that no satisfactory p-values are observed, nonparametric distribution fitting was used. The followed methodology is provided in Figure 3.1.

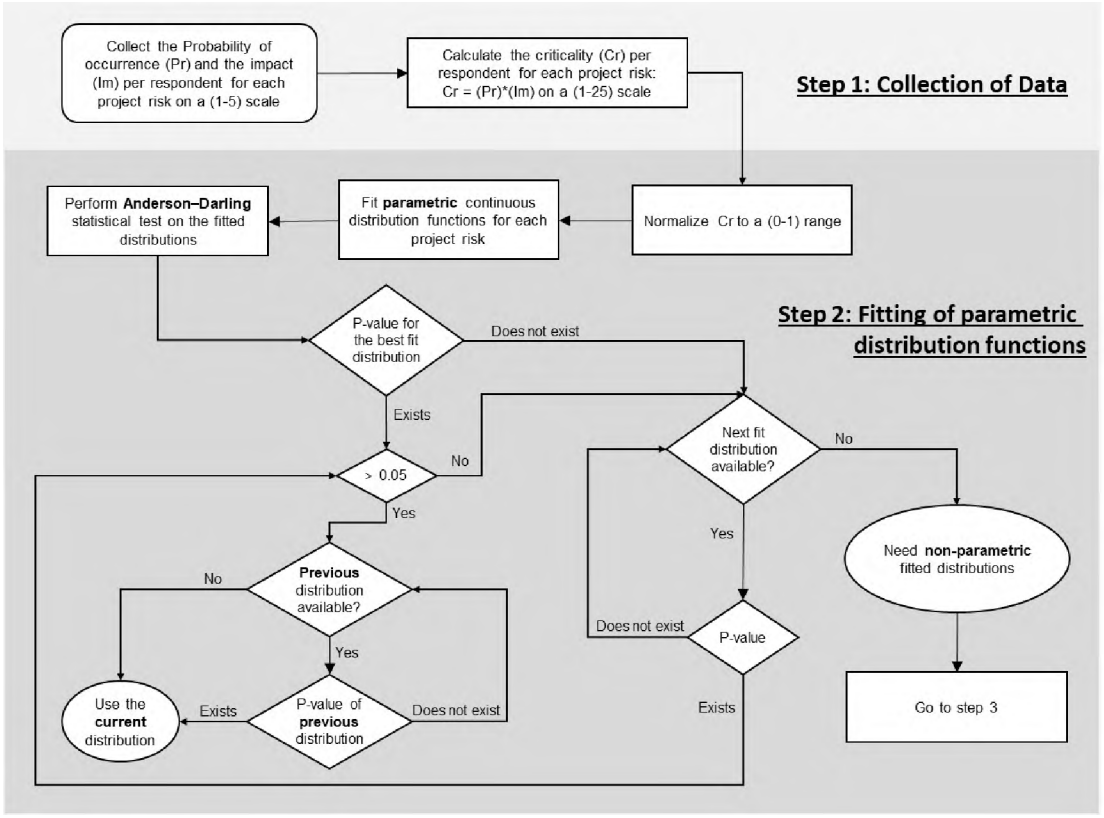

Figure 3.1 Followed methodology to fit parametric distributions.

After fitting the parametric distributions, it is important to make sure that the fitted distributions effectively model Cr in the [0, 1] range. As such, the fitted distributions with possible values outside the [0, 1] interval need to be truncated because Cr cannot take values outside this range. Accordingly, the fitted parametric distributions were truncated using Equation (3).

$$T(c) = \begin{cases} \frac{f(c)}{[\int f(c)dc]_k^l} & \text{for } k \leq c \leq l \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where f(c) = fitted parametric distribution without truncation and T(c) = truncated distribution over the interval [k; l] ⊆ [0; 1]. This approach is important to ensure proper boundary conditions for the model (having values in the [0, 1] interval). To be noted is that the parameters of the fitted distributions are estimated with a confidence interval of 95%, indicating that the true parameters of the distributions are calculated with a 95% certainty.

3.5.3. Step 3: Fitting of Nonparametric Distribution Functions. Historically, polynomial models are considered to be one of the most used empirical techniques for fitting functions (NIST/SEMATECH, 2018). Polynomial functions are effective because they (1) are useful when a distribution must be fitted empirically; (2) are very flexible with different data structures; (3) can represent complicated functions; (4) possess well understood and known properties; and (5) are computationally easy to use (NIST/SEMATECH, 2018). As such, polynomial functions were used to fit the distributions that could not be fitted using the procedure indicated in step 2 and Figure 3.1.

For any function to carry the attributes of a continuous distribution, the following properties are needed: (1) the function should be nonnegative over the entire data range; (2) the function should be continuous over the entire data range; and (3) the integral of the function should equal 1 over the entire data range—known as the unity property. The followed methodology first attempted to fit a single polynomial function with the previous properties over the entire data range. If a single polynomial function was not able to represent the entire data range, the data range was divided into two parts, and one polynomial function was fitted for each range (that is, a total of two polynomials). In this section of the dissertation, the first polynomial is labeled as f(c), and the second one as g(c). In such cases, one split point was needed to divide the data range into two parts. In this section of the dissertation, this split point was labeled Ω and was selected such as f(Ω) = g(Ω) (Figure 3.2).

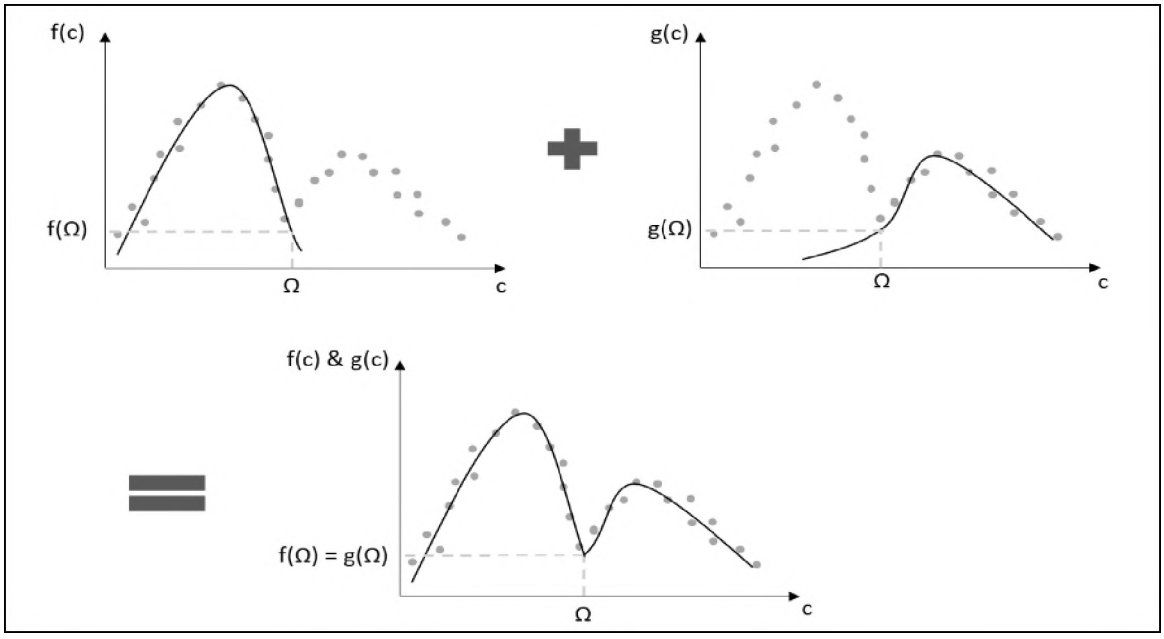

Figure 3.2 Fitting polynomial distributions by splitting the data range into two intervals.

The respective cumulative functions of f(c) and g(c) were labeled F(c) and G(c), respectively. To be noted is that the cumulative distribution functions rather than the probability distribution functions are relative to the developed model. As such, the derivation of the equations underlying the model is based on cumulative distribution functions calculated by integrating the probability density functions.

The first property of distribution functions was satisfied by selecting the polynomials that are greater than or equal to zero over the [0, 1] interval of the Cr. The second property of distribution functions was satisfied by having $F(\Omega) = G(\Omega)$ to preserve continuity. The value of the best-fitted cumulative parametric distribution function at Ω was used and is donated as η in this section of the dissertation. The third property of distribution functions was satisfied using Equation (4) and (5).

$$F(c) = \frac{\int a(c)dc}{\int_0^{\Omega} a(c)dc} . \eta \quad \text{over the } [0, \Omega] \text{ inerval} \tag{4}$$

$$G(c) = \frac{\int b(c)dc - [\int b(c)dc]_0^{\Omega}}{\int_{\Omega}^{1} b(c)dc} . (1 - \eta) + \eta \quad \text{over the }]\Omega, 1] \text{ interval} \tag{5}$$

where Ω = cut point that divides the entire [0, 1] data interval into two intervals [0, Ω] and [Ω, 1]; a(c) = fitted polynomial distribution function without the unity property over the interval [0, Ω]; F(c) = fitted cumulative polynomial distribution function over the interval [0, Ω]; b(c) = fitted polynomial distribution function without the unity property over the interval [Ω, 1]; G(c) = fitted cumulative polynomial distribution function over the interval

[Ω,1]; and η = value of the best fitted cumulative parametric distribution function assessed at the Ω data point. The inclusion of η ensures the continuity property of the cumulative distribution functions, that is, having F(Ω) = G(Ω).

In the event in which one polynomial function was able to be fitted for the entire data range, Equation (6), which is a modified version of Equation (4), is used to ensure the unity property of the fitted distribution, where a(c) denotes the fitted polynomial distribution function without the unity property over the entire [0, 1] data interval. Figure 3.3 shows a summary of the methodology for the aforementioned process.

$$F(c) = \frac{\int a(c)dr}{\int_0^1 a(c)dr} \quad \text{over the } [0, 1] \text{ interval} \tag{6}$$

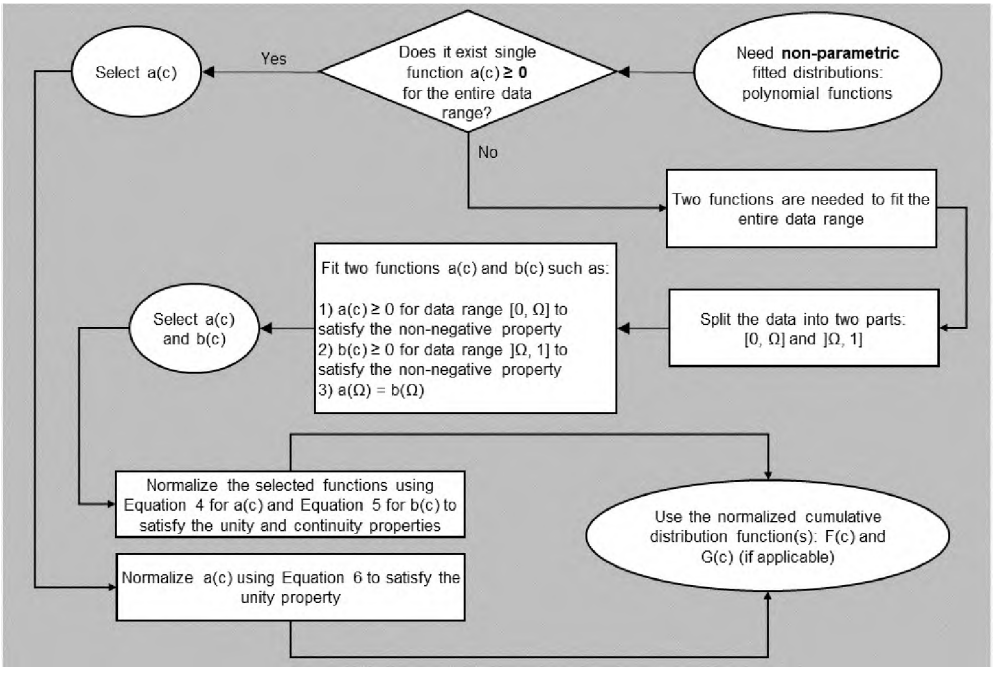

Figure 3.3 Followed methodology to fit non-parametric polynomial distributions.

3.5.4. Step 4: Calculation of Cost and Schedule Overruns. This step provides the methodology followed for the distributions fitting of cost and schedule overruns as well as for calculating the individual weights of the different project risks.

3.5.4.1. Fitting distributions for cost and schedule overruns. Considering the cost overrun as a continuous random variable denoted as "co" and the schedule overrun as a continuous random variable denoted as "so", fitting distribution functions for these two project performance indicators is crucial. However, to predict project performance, having reliable data on the cost and schedule overruns of past projects is essential. To this end, this section of the dissertation relies on the data published by CII in the Construction Owners Association of Alberta (COAA) major projects benchmarking summary report (CII, 2009). Worth noting is that CII's COAA publication includes a comparison between projects in Alberta and the United States on many aspects, such as project size, contingency budget, cost and schedule performance, change cost factor, and engineering productivity, among others. For the scope of this section of the dissertation, the used data is that of the US-related data proclaimed in CII's COAA publication for projects' cost and schedule growths, respectively (CII, 2009, p. 34). The reason for choosing this data source is that it is reliable, recent, based on real US projects, and contains the information needed to develop the model. The published data includes two box-and-whisker diagrams for industrial US projects that illustrate the cost and schedule overruns, respectively. The data present in the report allows to fit distribution functions for both the cost and schedule overruns of US industrial projects. A sample box-and-whisker plot is depicted in Figure 3.4 to better visualize the format of the data present in the CII's publication.

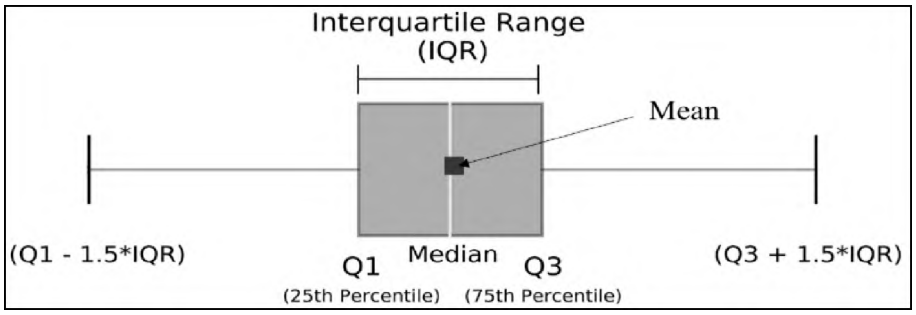

Figure 3.4 Sample box and whisker plot.

To this end, some distribution functions have flexibility, making it possible to fit them to a known set of data points whereas others do not possess this aspect, making them unable to be used to fit known data points. As such, this section of the dissertation explored possible distribution functions that possess the flexibility property (Clemen and Reilly, 2013) to fit distributions for cost and schedule overruns. The investigated distribution functions are beta general, beta, triangular, exponential, and normal. Compared with others, these functions allow the fitting of distributions based on data points (such as mean, median, Q1, and Q3, among others) rather than the entire dataset.

3.5.4.2. Calculation of project risk weights and performance prediction. The weights for the individual project risks were calculated using Equation (7), based on Cr's reference point. Equation (7) is simply a weighted average equation (individual criticality over the sum of criticalities of all project risks).

$$W_i = \frac{\overline{Cr_i}}{\sum_{i=1}^{i=25} \overline{Cr_i}} \tag{7}$$

where W_i = weight for project risk *i* and $\overline{Cr_i}$ = average criticality for project risk *i*.

The weights calculated using Equation (7) could be used as a baseline to predict the cost and schedule overruns; however, project administrators could assign different weights than those obtained (in the case of correlations between the different project risks) if enough information is available for the quantification of the individual contribution of each project risk toward project performance. In such cases, project administrators should ensure that the sum of the assigned individual weights equals 1.

To predict cost and schedule overruns, the following methodology is followed: (1) compute $\omega = \sum_{i=1}^{i=25} W_i * F_i(c)$ where $F_i(c)$ = value of the fitted cumulative distribution function for project risk *i* evaluated at point *c* specified by the project administrators for project risk *i*; (2) retrieving the value of the cost and schedule overruns such that F(co) = ω and F(so) = ω, where F(co) = fitted distribution for the cost overrun and F(so) = fitted distribution for the schedule overrun.

3.5.5. Step 5: Model Verification. Model verification is crucial because it helps uncover possible errors in the developed model and test the suitability of the model for its intended purpose. As such, any created model should be verified to ensure that it is a sound basis for decision making. For any developed model to be robust, it is important that it behaves similarly to the real system (Sterman, 2000). To this end, the extreme condition and surprise behavior tests were used to verify the developed model.

3.5.6. Step 6: Guidelines for Using the Developed Model in Industry Practice. This section of the dissertation presents guidelines for the actual industry usage of the developed model. The needed steps that users should follow to ease the model's implementation are depicted in the next subsections. In other words, the summarized

procedure that project administrators or users of the developed model use to predict cost and schedule overruns is presented in later subsections.

3.5.7. Step 7: Model Application. A hypothetical dataset was used to present the potential of the developed model by demonstrating its applied use by industry practitioners as well as its ability to induce behavior patterns. This is performed in two different cases by generating random values for criticality and assessing the corresponding cost and schedule overruns. In Case 1, the odd-numbered project risks are assigned random numbers between 0.5 and 1 (high criticality), and the even-numbered ones are assigned random numbers between 0 and 0.5 (low criticality). In Case 2, the same criticality values from Case 1 are used but in a flipped manner in which the odd-numbered risks take the criticality values of the even-numbered risks and vice versa. In each of the two cases, the cost and schedule overruns are calculated. In theory, although the overall average criticality of risks in Case 1 and Case 2 is the same, the difference in the forecasted overruns (schedule and cost) from one case to another is hypothesized by the model being sensitive to each project risk. This provides a better understanding of how the project performance reacts to changes in the evaluated criticalities of the project risks. In addition, a sensitivity analysis is performed to investigate how project performance changes with criticality.

3.6. COLLECTED DATA

This subsection provides all the information that is pertinent to the collected data.

3.6.1. Respondents' Demographics. The model was developed using an entire dataset collected from a total of 63 respondents (out of 196 targeted professionals) who represent the following major professions in the construction industry: owners (28 out of

63), consultants (12 out of 63), and contractors (23 out of 63). This was equivalent to a response rate of 32.14%. The average experience of the respondents is 24.3 years, and approximately 79% of the respondents have more than 10 years of experience. In addition, 13 respondents have less than 10 years of experience, with the majority (50.79%) having more than 20 years of experience. Moreover, most respondents are senior project managers and construction managers for industrial projects. The aforementioned characteristics of the respondents reflect that the responses are considered to provide well-rounded data that could reliably represent the construction industry in the United States.

To investigate whether the collected data from the respondents could be considered a representative sample that could be used as a good basis for the developed model, studying the sufficiency of the targeted sample size and the received response rate is crucial.

3.6.2. Determination of Sufficient Targeted Sample Size. According to Belafi et al. (2018), "From a statistical point of view, sample size should be based on the confidence interval and confidence level needed to achieve reliable results." As such, Dillman's (2011) commonly used statistical equations in the field of construction management research were used. These equations utilize confidence intervals and levels and are used in this section of the dissertation to study the sufficiency of the targeted sample size, as presented in Equation (8). Equation (8) was used in many previous construction management studies to determine the sufficiency of the survey's sample size, including the work of Wei et al. (2016), Choudhry and Zahoor (2016), Rasul et al. (2019), Belafi et al. (2018), and Zeb et al. (2015), among others.

$$N_S = \frac{(N_p)(p)(1-p)}{(N_p-1)\left(\frac{B}{C}\right)^2+(p)(1-p)} \quad (8)$$

where N_s = needed sample size (to be determined); p = population proportion [set at the most conservative value of 0.5 (Dillman, 2011)]; B = chosen sampling error corresponding to a confidence interval; C = Z-statistic associated with the chosen confidence level; and N_p = total population [equal to 352 (CII, 2009) but, to be more conservative, the next value of 400 present in statistical reference tables was used (Table 3.4)].

Table 3.4 Statistical reference table for sample size calculation. Data from Dillman (2011); Bartlett et al. (2001); and Taherdoost (2017).

Population size	Sample size for 95% confidence and 5% sampling error
50	44
75	63
100	80
200	132
300	169
400	196
600	234
800	260
1,000	278

Whereas some previous studies employed a sampling error of 10% as a confidence interval (Choudhry et al., 2012; ElZomor et al., 2018; Assaf et al., 2017), this section of the dissertation uses a better and more conservative sample error of 5% (B = 0.05) "which is considered to be an acceptable margin" (Sunindijo and Kamardeen, 2017). In addition, although some previous research utilized a confidence level of 90% (El-Hoteiby et al., 2017; ElZomor et al., 2018), this section of the dissertation uses the commonly acceptable

confidence level of 95% (C = 1.96) (Dillman, 2011; Choudhry et al., 2012). To this end, applying Equation (8) yields a conservative targeted sample size of 196, or it could be retrieved from Table 3.4 or calculated using online calculators, such as Qualtrics (2019) and Creative Research Systems (2019). Accordingly, this section of the dissertation utilizes a sufficient targeted sample size of 196 respondents, indicating that it could be used as a basis to develop the model.

3.6.3. Sufficiency of the Response Rate. Both statistical techniques and an empirical examination of previous studies in the construction industry were used to investigate the sufficiency of the collected data from the 63 respondents.

3.6.3.1. Statistical verification. According to the common research methods applied in construction as provided by Fellows and Liu (2015), the minimum number of respondents that would result in meaningful findings is calculated using Equation (9). Equation (9) was first introduced by Cochran (1977) and used in many previous construction management research studies, including the work of Sunindijo and Kamardeen (2017), Pereira et al. (2018), and Srour et al. (2017), among others.

$$n = \frac{(t^2)(s^2)}{(e^2)} \tag{9}$$

where n = number of respondents (to be determined); t = Z-statistic corresponding to the chosen significant value α; s = estimate of variance deviation for the scale used for data collection, which is calculated by dividing the inclusive range of the scale by the number of standard deviations that include almost all possible values in the range; and e = number of points on the primary scale multiplied by the acceptable margin of error.

For a commonly used significance of 95% (or α value of 0.05) (Pereira et al., 2018; Kamali and Hewage, 2017), the corresponding value of t is 1.96. For a five-point Likert scale, some references take s to equal 5/6 (Fellows and Liu, 2015; Randiwela and Wijayaratne, 2017), whereas other references take s to equal 5/4 (Hatamleh et al., 2018; Ogaji et al., 2018). As such, both values were considered in the calculations. In addition, using the common value of a 5% margin of error (Pereira et al., 2018; Kamali and Hewage, 2017), the value of e is equal to (5) by (0.05), where 5 is the number of points on the used Likert scale and 0.05 is the margin of error.

Accordingly, for an s of 5/4, the calculated minimum number of respondents using Equation (9) is 96, and for for an s of 5/6, the calculated minimum number of respondents is 43. To this end, from a statistical perspective, the collected data from the 63 respondents is considered sufficient because it lies between 43 and 96; however, empirical verification is still needed to better support the sufficiency of the sample size as provided in the next subsection.

3.6.3.2. Empirical verification. Because the construction industry is known to have a lack of participation in questionnaires (Cheong Yong and Emma Mustaffa, 2012; Wu et al., 2015), many studies highlighted a commonly used or acceptable response rate for survey-based construction research work. According to Fellows and Liu (2015), the expected useable response rate is between 25% and 35% in the construction research because "survey techniques, such as questionnaires, interviews and so on, are highly labour intensive on the part of respondents and, particularly, on the part of the researcher." In addition, surveys conducted via the Internet are likely to have lower response rates (Manfreda et al., 2008). According to Akintoye (2000), the most commonly used response

rate in construction research involving surveys is in the 20%–30% range. Further, this range was stressed by other studies (Liu et al., 2016b; Hwang et al., 2015). Moreover, Ryal-Net and Kaduma (2015) provided that 30% is an acceptable response rate in construction studies. Further, Yates (2014) provided that the average response rate is 27%. In addition, according to Tan et al. (2014), the normal research survey rate in the construction industry is considered to be between 10% and 20%. To this end, the obtained 32.14% response rate is considered acceptable because it falls within the commonly used response rate in survey-based construction research studies.

To provide additional insights into the validity and sufficiency of the obtained response rate based on the 63 respondents, the response rate in this section of the dissertation was compared to similar previous construction management efforts that studied or predicted the performance of construction projects. For instance, Karimi et al. (2018) modeled and elucidated the influence of skilled labor availability on construction cost performance based on data collected on industrial projects in the United States and Canada with a response rate of 30%. Yates (2014) investigated the impact of the utilization of sustainable practices on the performance of industrial construction projects in the design and construction phases with a response rate of 13.5%. Oh et al. (2016) studied the impact of the front-end planning process on construction input that is integrated after the completion of the design stage of projects with the highest response rate of 31.65%, for which most of the respondents had experiences in industrial projects. Sadafi et al. (2012) surveyed experts in industrial building systems to explore the adaptability of industrial components and their related issues to suggest improvements for the construction industry using a valid response rate of 20%. Ling et al. (2008) developed models to predict project

performance based on project management practices with a response rate of 17%. Chen and Manley (2014) measured the performance and governance of collaborative projects based on identified key mechanisms with an overall response rate of 19.0%. Liu et al. (2016a) studied the key factors that affect project success under different project delivery systems based on the contractor's characteristics with a response rate of 32.02%. Cheong Yong and Emma Mustaffa (2012) studied the principal factors that determine the success of construction projects and established their relative importance with a response rate of 31.1%. Tabish and Jha (2018) studied the performance of public construction projects and how it is influenced by the compliance of norms, which are composed from transparency- and audit-related variables with a response rate of 12.92%. To this end, the obtained 32.14% response rate is considered acceptable because it falls within the commonly used response rate in previous similar survey-based construction research that studied or predicted the performance of construction projects in general and industrial ones in specific.

3.7. FITTING OF DISTRIBUTION AND CALCULATION OF WEIGHTS

This subsection provides the obtained results related to fitting parametric and nonparametric distribution functions.

3.7.1. Data Normalization. After calculating Cr over a (1–25) interval for each respondent, 63 normalized criticalities per project risk were calculated using Equation (2). As such, the empirical distributions (the collected data) of the normalized criticalities are presented in Figure 3.5.

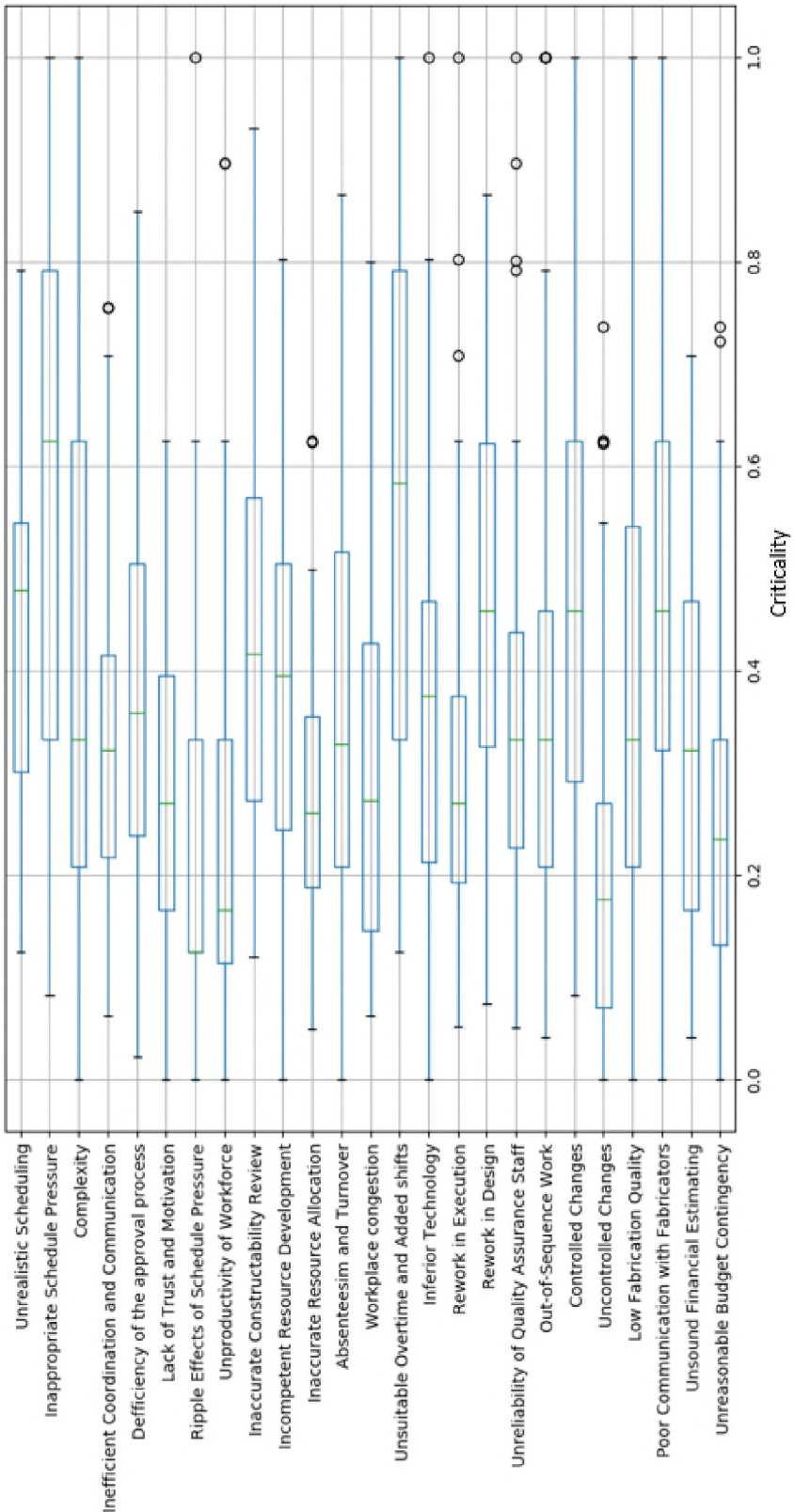

Figure 3.5 Normalized collected data for Cr.

3.7.2. Parametric and Nonparametric Distributions Fitting. Following the methodology depicted previously in Figure 3.1, 22 out of the 25 project risks were fitted using parametric distributions. The three other project risks were fitted using nonparametric functions. The fitted distributions of Cr for each of the risks are presented in Table 3.5.

Table 3.5 Fitted distributions for Cr for each project risk.

Project risk	Distribution for Cr
Unrealistic scheduling	Uniform
Inappropriate schedule pressure	Uniform
Inefficient coordination and communication	Loglogistic
Deficiency of approval process	Kumaraswamy
Lack of trust and motivation	Pert
Unproductivity of workforce	Frechet
Inadequate constructability reviews	Triangular
Incompetent resource development	Dagum
Inaccurate resource allocation	Dagum
Absenteeism and turnover:	Beta general
Workplace congestion	Pearson type V
Unsuitable overtime and added shifts	Uniform
Inferior technology	Dagum
Rework in execution	Loglogistic
Rework in design	Pert
Unreliability of quality assurance staff	Dagum
Controlled change	Dagum
Uncontrolled change	Fatigue life
Low fabrication quality	Fatigue life
Poor communication with fabricators	Extreme value
Unsound financial estimating	Rayleigh
Unreasonable budget contingency	Dagum
Complexity	“nonparametric distribution”
Ripple effects of schedule pressure	“nonparametric distribution”
Out-of-sequence work	“nonparametric distribution”

As observed in Table 3.5, the various project risks have different distributions that reflect their distinct impacts on project performance. In addition, some related project risks had similar distributions but with different distribution parameters. For instance, the unrealistic scheduling and inappropriate schedule pressure project risks, which are related directly to the work schedule, both follow the uniform distribution. Similarly, the incompetent resource development and inaccurate resource allocation project risks, which are related to the management of project resources, follow the Dagum distribution. The Dagum distribution is a continuous function defined over positive real numbers and is useful in many actuarial statistics or risk management (Palisade, 2018). Other distributions include the Kumaraswamy distribution, which is a continuous distribution used for lower and upper bounded variables, that could be used on the [0, 1] interval, and is similar to the Beta distribution but much simpler to implement in simulation studies. Additional distributions included the Pert distribution, which is a continuous function with a curved density that is a special case of the Beta General distribution (Palisade, 2018) and is widely used in risk analysis (PMI, 2013). Accordingly, the fitted distributions present in Table 3.5 are used in risk management to model uncertainties and complexities. The cumulative distribution functions of each of the 25 project risks after truncation are presented in Equation (B1) to (B25) provided in Appendix B (the sequence of these equations as related to each one of the 25 project risks is the same as the sequence of project risks shown in Table 3.5), and the graphs of each of the fitted distributions are presented in Figure 3.6. It is to be noted that the graphs shown in Figure 3.6 aim to act as a quick reference for practitioners that will be using the developed model, so that it would be easier to calculate the 25 different values of F(c) (each corresponding to each one of the 25 project risks).

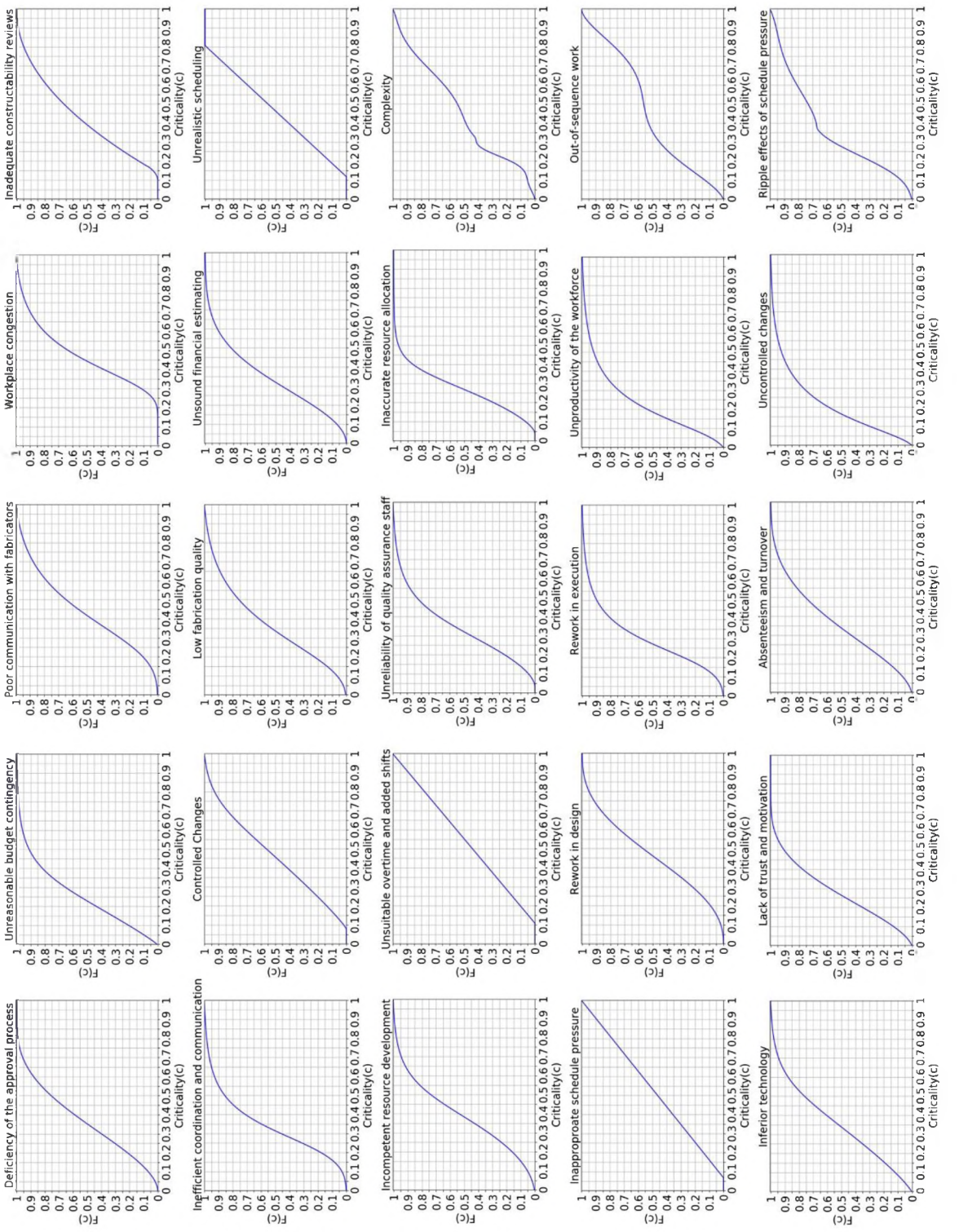

Figure 3.6 Graphs of the fitted distributions for the 25 project risks.

The different behaviors of the cumulative distribution functions reflect that different project risks have different impacts on project performance. The rate of increase (that is, the slope) of the cumulative function is one of the indications of the criticality of the project risk toward project performance. A steep cumulative distribution curve indicates that, for small changes in the criticality value of the project risk, the impact of such a change is higher on project performance in terms of cost and schedule overruns. One other indication for the impact of project risk is the individual weight of each project risk on project performance. The individual weights for the 25 project risks are presented in the next subsection. Additionally, as Table 3.5 indicates, the project risks that could not be fitted using parametric distribution functions are complexity, ripple effects of schedule pressure, and out-of-sequence work. These project risks share common aspects between them: they are very complex in nature, and their impacts are quantified only after their occurrence and, thus, difficult to predict instantly. Because the cumulative distribution functions of these project risks were fitted using nonparametric distribution functions, the curves of these project risks (Figure 3.6) indicate many curvatures that reflect their complex nature. Nevertheless, the fitted cumulative distribution functions are strictly increasing from 0 to 1 over the [0, 1] interval and are continuous over the same interval. This result indicates that the fitted functions possess the properties of cumulative distribution functions and, thus, could be adopted to predict the impact of these project risks on project performance.

3.7.3. Cost and Schedule Overruns Distributions Fitting. Using the data published by CII in the COAA major projects benchmarking summary report (CII, 2009), the relevant data points were extracted from the cost and schedule overruns for industrial

projects in the United States. The retrieved data for the cost overrun corresponds to 352 US projects, and that of the schedule overrun corresponds to 338 US projects. After fitting distributions for the cost and schedule overruns, the triangular distribution yielded more accurate representations of such datasets. The triangular distribution could be defined by three parameters: minimum value, most likely value, and maximum value. Table 3.6 indicates these parameters for the fitted distributions for the cost and schedule overruns.

Table 3.6 Fitted distribution functions for cost and schedule overruns.

Project performance	Distribution	Paraments of distribution		
		Minimum	Most likely	Maximum
Cost overrun	Triangular	-0.32	0.036	0.36
Schedule overrun	Triangular	-0.23	0.0225	0.33

The equations of the fitted distributions for the cost and schedule overruns are presented in Equation (10) and (11) respectively, and the graphs of the cost and schedule overruns are presented in Figure 3.7. As Figure 3.7 indicates, the developed model is able to predict negative as well as positive project performance, reflecting that the negative nomenclature of project risks aims only to provide consistency in how users input their criticality values.

$$F(co) = \begin{cases} 0 & if \quad co < -0.32 \\ 4.13co^2 + 2.64co + 0.422 & if -0.32 \leq co \leq 0.036 \\ -4.54co^2 + 3.27co + 0.411 & if\ 0.036 < co \leq 0.36 \\ 1 & if\ 0.36 < co \end{cases} \tag{10}$$

$$F(so) = \begin{cases} 0 & if \quad so < -0.23 \\ 7.07so^2 + 3.25so + 0.374 & if -0.23 \leq so \leq 0.0225 \\ -5.81so^2 + 3.83so + 0.369 & if\ 0.0225 < so \leq 0.33 \\ 1 & if\ 0.33 < so \end{cases} \quad (11)$$

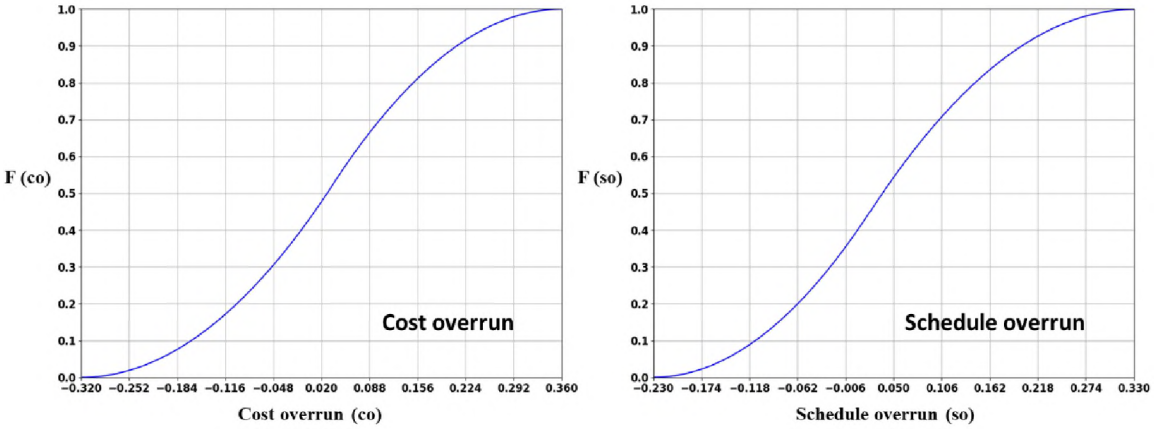

Figure 3.7 Graphs of the fitted distributions for cost and schedule overruns.

3.7.4. Calculation of Project Risks Weights. After fitting the distributions for cost and schedule overruns, the contribution of each project risk toward the project performance was quantified using Equation (7). The results are provided in Table 3.7. In addition to the slope of the fitted cumulative distribution functions, the calculated individual weights are the second indication of the impact of each project risk on project performance. It is recommend to use the weights presented in Table 3.7 to predict the cost and schedule overruns; however, project administrators could assign different weights than those presented in this section of the dissertation (to reflect the possible correlation between the project risks) if enough information is available for the quantification of the individual

contribution of each project risk toward project performance. In such cases, project administrators should ensure that the sum of the assigned individual weights equals 1.

Table 3.7 Weights for various project risks.

Project risk	Weight
Unrealistic scheduling	0.020
Inappropriate schedule pressure	0.051
Complexity	0.020
Inefficient coordination and communication	0.031
Deficiency of approval process	0.038
Lack of trust and motivation	0.057
Ripple effects of schedule pressure	0.034
Unproductivity of workforce	0.054
Inadequate constructability reviews	0.027
Incompetent resource development	0.050
Inaccurate resource allocation	0.017
Absenteeism and turnover	0.069
Workplace congestion	0.038
Unsuitable overtime and added shifts	0.070
Inferior technology	0.031
Rework in execution	0.038
Rework in design	0.037
Unreliability of quality assurance staff	0.038
Out-of-sequence work	0.028
Controlled change	0.047
Uncontrolled change	0.024
Low fabrication quality	0.044
Poor communication with fabricators	0.040
Unsound financial estimating	0.055
Unreasonable budget contingency	0.042

3.8. MODEL VERIFICATION

This subsection provides all details related to the verification of the developed mathematical and risk model.

3.8.1. Extreme Condition Test. The extreme condition test specifies that any created model should be robust in extreme conditions. Therefore, the developed model should behave in a realistic fashion irrespective of how extreme the inputs were imposed on it (Sterman, 2000). To test the model presented, it is important to investigate whether the model behaves appropriately when the inputs (that is, the criticality of project risks) take on the extreme values of 0 and 1, respectively. Two main methods can be used to perform extreme condition tests: direct inspection and simulation (Sterman, 2000). Because the developed model presents explicit equations and graphs for each input, the direct inspection method is used to test the model. Following this method, considering the model's response when all inputs simultaneously take on their extreme values is important (Sterman, 2000). To this end, when the criticalities for all project risks take zero values, Equation (B1) to (B25) take zero values (better visualized in the associated graphs presented in Figure 3.6). Using the weights in Table 3.7, the weighted value of these zero numbers is zero. Substituting the zero value in Equation (10) and (11), the predicted cost overrun will be −0.32, and the predicted schedule overrun will be −0.23 (better visualized in the associated graphs presented in Figure 3.7). The obtained values for the cost and schedule overruns are the minimum values that these variables could take. In a similar manner, when the criticalities for all of the risks take a value of 1, the predicted cost overrun will be the maximum value of 0.36, and the predicted schedule overrun will be the maximum value of 0.33. This result indicates that the predicted cost and schedule overruns are in line with the extreme values of the inputs.

3.8.2. Surprise Behavior Test. Discrepancies between the behavior of the model and expectations indicate that flaws in the developed model are present (Sterman, 2000).

A developed model passes the surprise behavior test when it generates a certain behavior that occurs in the real system (Sterman, 2000). For predicting project performance, the cost and schedule overruns of projects are expected to increase when the criticalities of the project risks increase and vice versa. Equation (B1) to (B25) and their corresponding graphs show that when the criticality value increases, the value of the function increases and vice versa. Checking the increasing behavior of the fitted cumulative parametric functions is not needed because they possess the increasing property by definition. However, investigating the increasing property of the nonparametric fitted distribution functions by investigating the sign(s) of their derivatives is important. As Figure 3.8 indicates, the derivatives of these functions are greater than 0 over the [0, 1] interval, showing that the nonparametric fitted cumulative functions are monotonously increasing. Therefore, the impacts of project risks will increase with an increase in the criticality values. Because the graphs of the cost and schedule overruns are also monotonously increasing (Figure 3.7), higher criticality values will induce higher cost and schedule overruns, reflecting that the developed model generates behaviors expected by the real system.

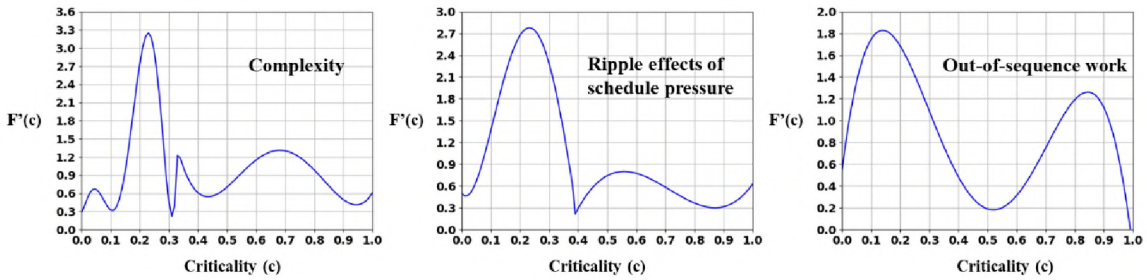

Figure 3.8 Derivatives of the fitted non-parametric functions.

3.9. MODEL APPLICATION

This subsection provides all details related to how the developed model shall be implemented in real projects.

3.9.1. Guidelines for Using the Developed Model in Industry Practice. To provide insights into the use of the developed model, this subsection provides a detailed description for users on how to utilize the formulated framework. First, project administrators will assess on a [0; 1] scale the magnitude of the uncertainty/probability of the project risk, the consequence(s) in the case that a project risk manifests, the amount of time available to react, and the contribution of the project risk failure toward the entire project failure in terms of cost and time as reflected by the qualitative definition of criticality provided previously. The criticality of the project risk is then calculated as the average of the four assessed values.

After calculating the criticality value "c" of each project risk, the project administrators calculate the value of F(c) using Equation (B1) to (B25) or their corresponding graphs, resulting in 25 values for F(c) (one per project risk). Subsequently, project administrators assign individual weights (W_i) for each project risk if they have the information to do so; if not, they are advised to use the weights in Table 3.7.

Subsequently, project administrators calculate $\omega = \sum_{i=1}^{i=25} W_i * F_i(c)$. The predicted cost and schedule overruns are the solutions to the equations F(co) = ω and F(so) = ω; F(co) and F(so) are present in Equation (10) and (11), respectively. Figure 3.9 provides the summarized procedure that project administrators use to predict cost and schedule overruns.

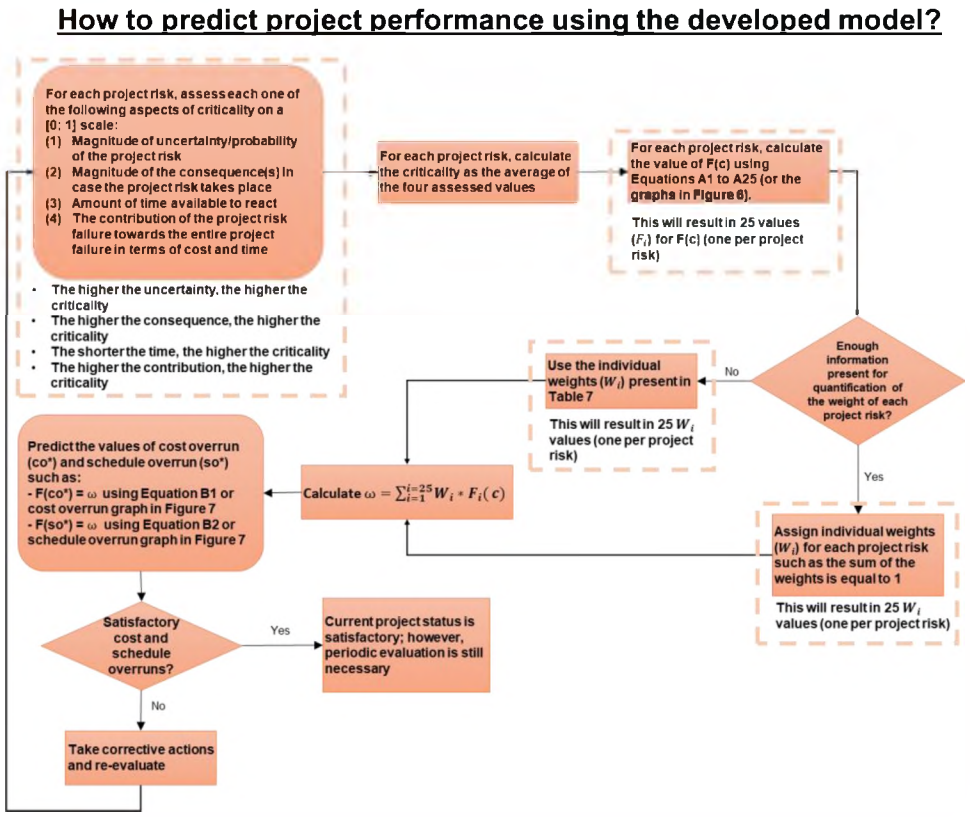

Figure 3.9 Procedure for using the developed model.

Although the developed model includes advanced mathematical formulations that might seem difficult in practice, users are not overwhelmed by the presented distributions or equations because the model could be easily formulated using merely simple tools (such as Excel). Using a simple developed spreadsheet, the user only provides the criticality values of each project risk for the current project under study. The calculation of the cost and schedule overruns is automatically performed by the spreadsheet in a few milliseconds.

3.9.2. Hypothetical Case Study. A hypothetical dataset was used to present the potential of the developed model by demonstrating its applied use by industry practitioners as well as its ability in inducing behavior patterns. Figure 3.10 indicates the prediction of the cost and schedule overruns for the two studied cases.

Project risk	Inputs (considerations for calculating the criticality of each risk)								Calculations						
	Uncertainty		Consequence		Time to react		Contribution to failure		Criticality (c)		F(c)		Weight	F(c)*Weight	
	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2	Case 1	Case 2		Case 1	Case 2
Unrealistic scheduling	0.6	0.5	1	0.4	0.5	0.1	0.9	0.2	0.75	0.3	0.92	0.27	0.020	0.0185	0.0054
Inappropriate schedule pressure	0.5	0.6	0.4	1	0.1	0.5	0.2	0.9	0.3	0.75	0.25	0.73	0.051	0.0127	0.0373
Complexity	1	0.1	0.6	0.4	0.5	0.3	0.8	0.5	0.725	0.325	0.80	0.42	0.020	0.0159	0.0084
Inefficient coordination and communication	0.1	1	0.4	0.6	0.3	0.5	0.5	0.8	0.325	0.725	0.53	0.97	0.031	0.0163	0.0301
Deficiency of the approval process	0.7	0.2	0.6	0.1	0.9	0.3	1	0	0.8	0.15	0.99	0.13	0.038	0.0376	0.0049
Lack of trust and motivation	0.2	0.7	0.1	0.6	0.3	0.9	0	1	0.15	0.8	0.23	1.00	0.057	0.0132	0.0570
Ripple effects of schedule pressure	0.7	0	0.9	0.1	0.8	0.4	0.8	0.4	0.8	0.225	0.92	0.35	0.034	0.0314	0.0120
Unproductivity of the workforce	0	0.7	0.1	0.9	0.4	0.8	0.4	0.8	0.225	0.8	0.60	0.99	0.054	0.0322	0.0533
Inadequate constructability reviews	0.9	0.3	0.7	0.5	0.8	0.4	0.5	0.4	0.725	0.4	0.90	0.51	0.027	0.0243	0.0138
Incompetent resource development	0.3	0.9	0.5	0.7	0.4	0.8	0.4	0.5	0.4	0.725	0.53	0.96	0.050	0.0266	0.0478
Inaccurate resource allocation	0.8	0.2	1	0.1	0.9	0	0.8	0.5	0.875	0.2	1.00	0.29	0.017	0.0170	0.0049
Absenteeism and turnover:	0.2	0.8	0.1	1	0	0.9	0.5	0.8	0.2	0.875	0.24	0.99	0.069	0.0167	0.0685
Workplace congestion	1	0.4	0.5	0.1	1	0.1	0.6	0.4	0.775	0.25	0.94	0.03	0.038	0.0358	0.0013
Unsuitable overtime and added shifts	0.4	1	0.1	0.5	0.1	1	0.4	0.6	0.25	0.775	0.16	0.75	0.070	0.0110	0.0523
Inferior technology	0.5	0.4	0.5	0.5	0.5	0	0.9	0.2	0.6	0.275	0.88	0.39	0.031	0.0271	0.0120
Rework in execution	0.4	0.5	0.5	0.5	0	0.5	0.2	0.9	0.275	0.6	0.51	0.95	0.038	0.0192	0.0360
Rework in design	0.5	0	0.7	0.5	0.5	0.2	0.6	0	0.575	0.175	0.69	0.07	0.037	0.0257	0.0025
Unreliability of quality assurance staff	0	0.5	0.5	0.7	0.2	0.5	0	0.6	0.175	0.575	0.16	0.89	0.038	0.0060	0.0338
Out-of-sequence work	0.6	0.2	0.9	0.1	0.7	0.1	0.7	0.3	0.725	0.175	0.67	0.26	0.028	0.0189	0.0074
Controlled change	0.2	0.6	0.1	0.9	0.1	0.7	0.3	0.7	0.175	0.725	0.16	0.96	0.047	0.0074	0.0449
Uncontrolled change	0.5	0.5	0.9	0.4	0.8	0	0.8	0.5	0.75	0.35	0.99	0.83	0.024	0.0237	0.0198
Low fabrication quality	0.5	0.5	0.4	0.9	0	0.8	0.5	0.8	0.35	0.75	0.52	0.94	0.044	0.0230	0.0414
Poor communication with fabricators	1	0.1	0.8	0.4	0.5	0	0.8	0.1	0.775	0.15	0.93	0.05	0.040	0.0370	0.0020
Unsound financial estimating	0.1	1	0.4	0.8	0	0.5	0.1	0.8	0.15	0.775	0.14	0.98	0.055	0.0077	0.0541
Unreasonable budget contingency	0.5	0.5	0.6	0.6	0.9	0.9	0.6	0.6	0.65	0.65	0.97	0.97	0.042	0.0407	0.0407

sum = ω	0.5457	0.6917
cost overrun	**0.0439**	**0.0997**
sch. overrun	**0.0499**	**0.0992**

Figure 3.10 Prediction of cost overrun and schedule overrun for hypothetical dataset.

Although both Case 1 and Case 2 have an equal average criticality value of 0.5 for all project risks, different cost and schedule overruns were predicted for each case. As Figure 3.10 indicates, the predicted cost overrun for Case 1 is 0.0439 compared with 0.0997 for Case 2. In addition, the predicted schedule overrun is 0.0499 for Case 1 compared with 0.0992 for Case 2. This result indicates that the predicted cost and schedule overruns depend on the specific criticalities of each project risk rather than on the broader project perspective. This dependence demonstrates that the developed model is dynamic in nature and preserves the contribution of the criticality of each project risk toward project performance. Therefore, the developed model leads to better quantification and prediction of project performance in terms of both cost and time. Figure 3.11 indicates an example for the use of the generated graphs through the extraction of the 0.0992 schedule overrun value for Case 2, as presented in Figure 3.10.

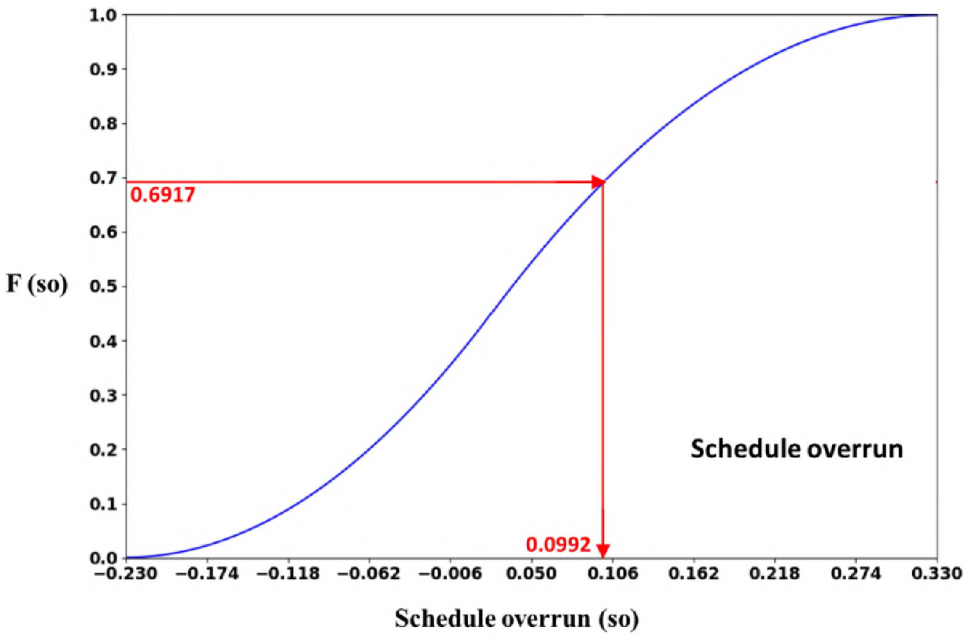

Figure 3.11 Example showing how to use the schedule overrun graph.

Further and to investigate the model's behavior patterns, a similar procedure as depicted in Figure 3.10 was followed; however, the criticality values were maintained constant for all project risks. For example, all risks had a criticality value of 0.1, 0.2, and so on. This analysis should better help present how criticality, alone, impacts project performance. As indicated in Figure 3.12, a sensitivity analysis was performed to reflect changes in project performance as related to criticality.

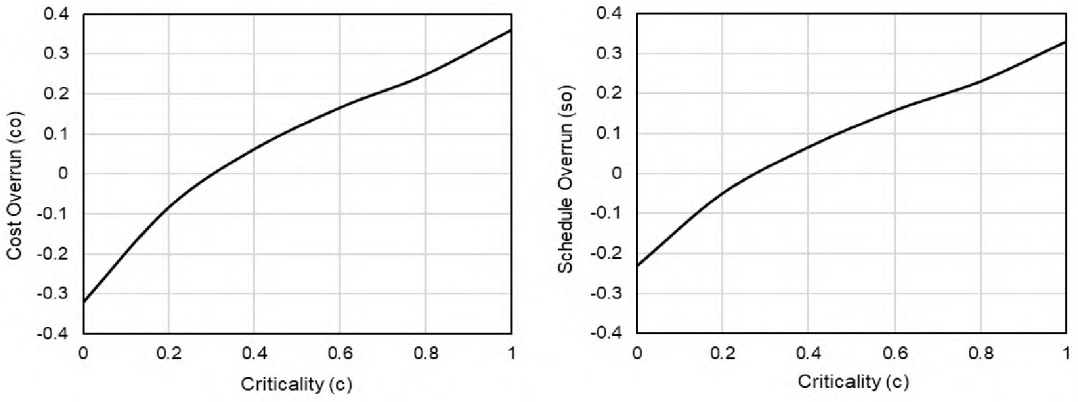

Figure 3.12 Cost and schedule overruns vs. criticality.

Because the curves present in Figure 3.12 are nonlinear, it is evident that the developed model does not treat cost and schedule overruns in a linear fashion. As such, this reflects that the proposed framework is nonlinear in nature, which mimics the real behavior and performance of construction projects and establishes the added value/contribution of the developed model to the construction industry's body of knowledge. Moreover, although both Case 1 and Case 2—as shown in Figure 3.10—have an equal average criticality value of 0.5 for all project risks, different cost and schedule

overruns were predicted for each case. As such, no definite behavior can be established between the performance of projects in terms of time and cost because no correlation could be devised between the graphs presented in Figure 3.12, which is also reflected in the results presented in Figure 3.10. Furthermore, as Figure 3.12 indicates, both cost and schedule overruns increase with increased criticality. To this end and because project performance is anticipated to worsen (that is, higher cost and schedule overruns) for increased criticalities, the findings presented in Figure 3.12 are in line with what is expected to happen in real-life cases. Therefore, the developed model mimics the real behavior of projects and, thus, the predicted project performance indicators are considered good estimates that project administrators could rely on when making decisions and taking corrective actions.

3.9.3. Limitations. Any developed model has limitations and possibilities for improvement. Although the developed model works for all types of construction projects, it yields the best results for industrial projects because the cost and schedule overruns are retrieved from CII's COAA major projects benchmarking summary report that publishes data on industrial projects. Additionally, most of the surveyed experts specialized in industrial projects. One other limitation is the possibility that project risks could unfold differently for each project; therefore, that the weights be project-specific is advised. An additional limitation is that possible correlations between the different project risks are not incorporated. Nevertheless, the incorporation of the correlation between the project risks could be addressed by assigning individual weights by users other than those provided, as previously specified in this section of the dissertation. In addition, the study of these correlations could be performed in future research based on the findings of this section of

the dissertation and on the availability of the needed data. However, this is possible by dint of the current work provided in this section of the dissertation that develops a model to assess and predict project performance in the construction industry; thus, extra capabilities could subsequently be easily added. Moreover, project administrators or users utilizing the developed model are not advised to use it after extreme/out-of-ordinary project performance for which the prediction of cost and schedule overruns will be very complex, uncertain, and unique in nature and, thus, may require more complex methodologies.

3.10. SUMMARY

This section of the dissertation presents a holistic model to predict project performance. The model forecasts two prevailing indicators: cost and schedule overruns. The developed model allows project administrators to predict the performance of construction projects based on 25 project risks that are retrieved from the literature and that have shown to be the most important based on a meta-analysis of the literature. The model presents explicit equations that could be used to forecast the project performance in terms of cost and schedule and a set of graphs that could be adopted as an alternative method for estimating project performance to minimize the computational efforts. In addition, using numerical examples of a hypothetical case study, the proposed model reflected its practicality and reliability in predicting project performance.

3.11. CREDIT

This subsection aims to acknowledge the discussions and inputs provided by Dr. Ibrahim S. Abotaleb, during his journey as a postdoctoral research fellow at the Missouri

University of Science and Technology, which are very appreciated for the development of the research reported in this section (i.e., Section 3) of the dissertation.

3.12. RELATED APPENDIX

Appendix B presents the Equations of the fitted parametric and non-parametric distribution functions for the 25 project risks considered in this section of the dissertation. It is to be noted that the sequence of these equations as related to each one of the 25 project risks is the same as the sequence of project risks shown in Table 3.5.

4. A STATISTICAL AND TIME SERIES MODEL TO STUDY THE IMPACT OF DYNAMIC WORKFORCE AND WORKPLACE VARIABLES ON THE PRODUCTIVITY OF THE CONSTRUCTION INDUSTRY

4.1. OVERVIEW

The construction industry is one of the key contributors to the growth of economies. For instance, the construction industry contributes to about 4.4% of the total gross output in the US (US Bureau of Economic Analysis, 2018b). In addition, the construction industry employs 6%–10% of the workforce (Shohet et al., 2019). Moreover, the construction industry plays an important role in driving the activities of other industries such as manufacturing, mining and agriculture, transportation (Donkor, 2011), and infrastructure (Assaad et al., 2020a), among others. In addition, construction is a complex business (Assaad et al., 2020b), and different construction-related uncertainties can impact productivity (Choy and Ruwanpura, 2005; Sexton et al., 2020).

Construction productivity is a fundamental piece of information for different essential construction activities such as estimating, budgeting, and scheduling (El-Gohary et al., 2017). In addition, construction productivity is an important metric that provides feedback about the industry trends and improvements (Vereen et al., 2016). In fact, construction productivity is considered the industry's predominant determinant of performance (Jarkas, 2016a). Consequently, it is important to understand the fluctuations or movements of construction productivity.

The construction industry is prone to many disruptions (Assaad and El-adaway, 2020a), and the dynamics of the workforce and workplace play an important role in the fluctuations of construction productivity (Chaturvedi et al., 2018; Durdyev et al., 2018).

However, this does not mean that the role of these variables is understood and quantifiable. In other words, although construction productivity is affected by different dynamic workforce and workplace variables, the effects of these variables cannot be understood or quantified unless they are researched. Dynamic variables refer to variables with values changing from one period to another or that are linked to other variables depending on time (Abbas and Mosallamy, 2016). In simple terms, dynamic variables can take different values over time.

Although the construction industry produces a large amount of data on a periodic basis, existing data sets have not been exploited fully (Cao and Goh, 2019). In relation to this, no previous research work exploited publicly available workforce and workplace variables to understand and model the fluctuations in the productivity of the construction industry as a whole. In other words, although the productivity's body of knowledge has productivity has many previous research works that provided important information on construction productivity, including but not limited to Durdyev et al. (2018), Gurmu and Ongkowijoyo (2020), Gupta et al. (2018), and El-Gohary et al. (2017), these efforts focused on labor productivity rather than on the productivity of the construction industry as a whole. Thus, construction productivity in this section of the dissertation is defined as, and refers to, the productivity of the overall construction industry; that is, the monthly generated output (in USD) per worker-hour. Thus, this section of the dissertation takes a high-level perspective by focusing on the productivity of the overall construction industry.

4.2. OBJECTIVE

The goal of this section of the dissertation is to study and model the impacts of dynamic workforce and workplace variables on the productivity of the construction industry. The associated research objectives are to (1) examine the statistical relationships and causalities between numerous workforce and workplace variables on one hand and the productivity of the construction industry on the other hand; (2) associate the changes in construction productivity with its past fluctuations and the past movements in different dynamic workforce and workplace variables; and (3) develop and validate a statistical framework that models and predicts the productivity of the construction industry.

4.3. CURRENT STATE OF LITERATURE, LIMITATIONS OF EXISTING STUDIES, AND BACKGROUND INFORMATION

The construction industry is subject to schedule overruns (Assaad and Abdul-Malak, 2020b), numerous uncertainties (Assaad and El-adaway, 2021b,), risks, and ever-changing conditions as it is ripe for many disruptions (Assaad et al., 2021a). One of the most critical uncertainties or risks is productivity. Thus, this subsection provides the needed background information related to construction productivity. This subsection also includes information related to time series analysis.

4.3.1. Previous Studies Related to Construction Productivity. This subsection details the relevant previous research work on construction productivity. Johari and Jha (2020b) explored the relationships between the work motivation of workers and their productivity based on 116 construction workers from 4 construction sites in India. Florez et al. (2020) proposed a new metric to measure compatibility of personality among workers in a crew and revealed how personality factors affect productivity based on rigorous

methods to analyze correlations for construction experiments. Ghodrati et al. (2018) quantified the effectiveness of a set of implemented management strategies in improving productivity in construction projects in New Zealand. Pan et al. (2019) examined the nature of the constraints on productivity advancement and explored the rationale underpinning the productivity enhancement strategies within the construction industries of Singapore, Hong Kong, and the United Kingdom. Johari and Jha (2020a) established the relationship between construction workers' aptitudes and their productivity based on data collected on 112 workers. Gurmu and Ongkowijoyo (2020) attempted to predict construction labor productivity based on implementation levels of human resource management (HRM) practices by performing a correlation and associations analysis between productivity, HRM practices, company profiles, and project properties based on data collected from 39 contractors. Durdyev et al. (2018) developed a structural equation model of the factors affecting construction labor productivity in the Malaysian construction industry. Gupta et al. (2018) determined different site amenities and worker's welfare factors that impact the workforce productivity in Indian construction projects. El-Gohary et al. (2017) introduced an engineering concept to document, control, predict, and improve contractors' labor productivity for formwork and reinforcing steel fixing crafts, including factors of the project management and administration level and of the activity level. Bonham et al. (2017) applied data mining techniques to quantify the relative influence of design and installation attributes on labor productivity. Kisi et al. (2018) tested and evaluated the validity of the traditional two-prong strategy on a complex and labor-intensive operation and provided a framework for estimating the optimal productivity for the fabrication activity of sheet metal ducts. Zhao and Dungan (2018) reviewed the methods used to quantify lost labor

productivity in the US construction industry, along with relevant cases, to examine the practical considerations in selecting the proper productivity quantification method. Gurmu (2019) developed a tool for scoring materials management practices for building projects and built a tool for predicting productivity based on a questionnaire distributed to construction experts and contractors.

4.3.2. Summary of Existing Literature on Construction Productivity. To provide a general description of the existing knowledge on construction productivity, this subsection summarizes the previous literature at three levels: high level, medium level, and detailed level. The high-level overview provides the broad research areas on construction productivity. The medium-level summary offers more-elaborate information on the existing knowledge on construction productivity. The detailed-level review provides a more thorough summary of the existing literature on construction productivity.

At the high level, Durdyev et al. (2018) demonstrated that the main thread in all contextual interpretations or definitions of productivity is related to efficiency and effectiveness. Efficiency is related to answering the question of how efficiently scarce resources are used throughout the implementation process to attain the intended objectives (Durdyev et al., 2018). Effectiveness is related to answering the question of how effectively the resources are utilized to fulfill the set targets (Durdyev et al., 2018). At the medium level, Yi and Chan (2014) conducted a systematic literature review of labor productivity in the construction industry and found that the existing knowledge could be classified into the following research areas: the effect of variations on construction labor productivity, methods and technology for productivity improvement, factors affecting labor productivity, modeling and evaluation of construction productivity, productivity's trends

and comparisons, and baseline/benchmarking construction labor productivity. At the detailed level, the existing knowledge on labor productivity involves numerous productivity aspects, including loss of labor productivity; the impact of work changes on labor productivity; the effect of engineering aspects on productivity; the influence of labor force skills and experience on productivity; and productivity factors related to management and control practices on construction sites, project financing elements, availability and use of the needed material and equipment, the properties of the project, external components, manpower-related characteristics, technical factors, and many others.

Combining the aforementioned high, medium, and detailed levels with the literature review conducted in the previous subsection, a conclusive summary was made of the existing knowledge on construction productivity. The knowledge is considered to fall under the following main categories: (1) prediction, modeling, and evaluation of labor productivity, because it is an important piece of information used in numerous construction-related aspects; (2) identification of the factors that affect construction labor productivity and its variations, to better understand how to improve the productivity of the construction workforce; (3) development of practical strategies and best practices, to enhance the productivity on construction sites as well as in the construction industry; and (4) examination of the effect of different methods, technologies, and constrains on productivity enhancements, efficiency, and effectiveness in different construction markets and countries. The previous literature on construction productivity includes qualitative, quantitative, and mixed-methods approaches (Kisi et al., 2018) that address the aforementioned diverse aspects of construction productivity.

4.3.3. Knowledge Gap. Previous research studies mainly focused on studying labor productivity in the construction industry. Although previous research works provided important knowledge on labor productivity, no previous research work attempted to study the causalities and relationships between the dynamic workforce and workplace variables and the productivity of the industry as a whole. Consequently, there is a knowledge gap in the literature in terms of the quantification of the impacts of different workforce and workplace variables on the productivity of the entire construction industry. This section of the dissertation proposes a new view of construction productivity by approaching it from the perspective of the entire construction output generated by the US construction industry with respect to total construction employment rather than the individual labor productivity of each construction worker or construction workforce occupation (such as electricians, carpenters, and plumbers, among others). Although the construction industry produces a large amount of data on a periodic basis, existing data sets have not been exploited fully (Cao and Goh, 2019). To this end, this section of the dissertation addresses this critical knowledge gap by developing a statistical framework that can model the causalities and the relationships between different dynamic variables and construction productivity. Consequently, this section of the dissertation takes the previous research directions a step further by rigorously investigating the relationship between dynamic workforce and workplace variables and the overall productivity of the construction industry.

4.3.4. Determination of Workforce and Workplace Variables. Different workforce and workplace variables related to construction productivity are mentioned in the literature. The expression 'workforce and workplace' refers to the variables related to the construction labor force and the workplace's working conditions. This subsection

identifies and describes these variables. Durdyev et al. (2018) considered turnover an influencing factor on construction productivity. Turnover refers to the number of construction workers who leave a construction company. Koch (2017) considered job openings to play a role in construction productivity and hiring practices. Job openings refers to the total number of open job vacancies that need to be filled by a construction worker. Rojas and Aramvareekul (2003) considered construction employment to investigate whether construction labor productivity is declining and compared it with manufacturing labor productivity. Construction employment refers to the total number of construction workers in the construction industry. Sveikauskas et al. (2016) considered average weekly hours worked in the construction industry to measure productivity growth. Average weekly hours refer to the average hours per construction worker for which pay was received. Ozturk et al. (2020) considered job losses to cause changes in productivity. Job losses refers to the total number of jobs lost by construction workers in the construction industry. Sveikauskas et al. (2018) considered the value of construction put in place (VIP) to assess productivity in the construction industry because this measure provides good information about the total construction output. VIP is a measure of the value of construction work installed or erected at construction sites. Vereen (2013) considered job gains to affect labor demand in the US construction industry, which in turn affects productivity. Job gains refers to the total number of jobs gained by construction workers in the construction industry. Also, the construction industry has been noted for its high incident rates and poor safety performance (Abdul Nabi et al., 2020a). In relation to that, Abrey and Smallwood (2014) considered injuries and illnesses as unsatisfactory working conditions that impact productivity in the construction industry. Injuries and illnesses refer

to the total number of nonfatal workplace injuries and illnesses that construction workers experience. Mirhadi (2018) considered hires to affect the efficiency of the construction workforce, and thus also construction productivity. Hires refers to the total number of employees that were hired in the construction industry. Setiani and Abd Majid (2019) considered fatalities and safety to be a factor influencing productivity in construction projects. Fatalities refers to the total number of workplace accidents that led to the death of construction workers. Kuznetsova et al. (2019) considered that there is a relationship between productivity and unemployment. Unemployment is the share of the construction labor force that is jobless. Hendrickson (2005) considered changes in average hourly earnings to affect the productivity measure in the construction industry. Average hourly earnings refer to the average dollar amount that construction workers earn per hour. In addition, the size and contribution of the construction industry to the economy usually is assessed as a percentage of gross domestic product (GDP) (Mahamid, 2013). Higher GDP could reflect better labor well-being, which could result in a better productivity. GDP is the monetary value of all finished goods and services made within the US. Allmon et al., (2000) considered total compensation rates as a factor that affects construction productivity. Total compensation is the cost paid by the construction employer for construction employee compensation per hour worked.

The previous workforce and workplace variables were considered for analysis. These variables were used because they were mentioned in the literature and because they are available from reliable sources such as the US Bureau of Labor Statistics (BLS), the US Census Bureau, and the US Bureau of Economic Analysis. These reliable sources

provide consistent, periodically updated, and well-maintained information and records on these different variables.

4.3.5. Time-Series Analysis and Vector Autoregression. This section of the dissertation studied and modeled the impacts of different workforce and workplace variables on the productivity of the construction industry. Because the collected data were time series showing the fluctuations of the workforce and workplace variables with time, the collected data were multivariate time series. Vector autoregression (VAR) is the most commonly used statistical technique for modeling and predicting multivariate time series (Singh, 2018). Therefore, time series analysis and VAR were deemed to be feasible and suitable for the research objectives and data set of this section of the dissertation. This subsection provides the needed background information on time series analysis and VAR.

4.3.5.1. Overview. Time series represent the changes in the values of dynamic variables over time. In other words, time series are the simplest form of temporal data, and are a sequence of real numbers collected regularly in time (Gunopulos et al., 2001). Studies utilizing time-series analysis are growing at a very fast rate due to its wide applications in a large variety of research fields (Gao et al., 2017). Time-series analysis is a useful tool for better understanding the cause-and-effect relationships between different dynamic variables (Kadilar and Kadilar, 2017). In simple terms, time-series analysis is a statistical method to model and predict the future values of a variable based on previously observed values of the same variable or other relevant dynamic variables (Yang and Liu, 2019). Time-series statistical analysis methods are divided into two main types: univariate time-series analysis, and multivariate time-series analysis. Univariate time-series analysis includes studying a single sequence of values for a particular variable. Multivariate time

series are more common in real life and real-world applications due to the inherent complexities of dealing with two or more dynamic variables (Wang et al., 2017a). Although many methods exist to analyze/predict time-series data, such as moving average, exponential smoothing, and autoregressive moving average, among others (Brownlee, 2018), VAR is the most commonly used statistical technique for modeling and predicting multivariate time series (Singh, 2018). VAR is a statistical modeling technique that expresses the terms in a multivariate time series of K dynamic variables as a linear combination of the previous p values of the K variables (Abdu-Aguye and Gomaa, 2018).

4.3.5.2. Previous studies. Due to the numerous benefits of VAR, such as good forecasting capabilities, simplicity of implementation, and ease of estimation (Anggraeni, 2016), it has been widely used in a variety of applications (Wang and Ding, 2018) such as economic, finance, construction management, and engineering, among others. In relation to previous studies that used time series and temporal data analysis in the construction engineering and management area, Lingard et al. (2017) examined the temporal relationship between the safety performance indicators to uncover time-dependent causal relationships. Cao and Goh (2019) used time-series analysis to identify the leading indicators or predictors of construction accidents, and they developed three different time-series models for predicting accidents' occurrences. Xu and Lin (2016) used VAR to analyze the influencing factors leading to changes in carbon dioxide emissions in the construction industry to develop appropriate energy policy and planning for the iron and steel industry. Faghih and Kashani (2018) relied on time-series analysis to forecast the short- and long-term prices of construction materials based on a set of relevant explanatory variables. Lee et al. (2016) developed a vector error correction model to perform an

empirical analysis of the impact of diversification on construction companies' insolvency. Ilbeigi et al. (2017) used time-series analysis to forecast the asphalt-cement price and examine whether and how time-series forecasting models can predict future prices of asphalt-cement with higher accuracy than the existing approaches. Lin et al. (2018) utilized time-series to find correlations between intellectual capital and business performance to enable corporations to shape policy decisions that benefit business performance. Swei et al. (2017) employed univariate time-series models to project future costs and prices of concrete and asphalt based on a probabilistic approach. Vereen (2013) developed a VAR model to forecast the labor demand and found that 5.3–6.3 million skilled workers will be in demand by 2022. Because the construction industry is a project-based industry, with projects taking several months and years to complete, the monthly productivity of the construction industry generally is considered to be dependent on its previous lagged values. Therefore, time series and VAR were employed to model and predict the productivity of the construction industry.

4.4. METHODOLOGY

As shown in Figure 4.1, the followed methodology is composed of different steps. Details on each one of these steps are provided in the next subsections. It is worth mentioning that the used data and the followed methodology in this section of the dissertation are slightly different than these reported in the work of Assaad and El-adaway (2021c). Therefore, slightly different results are obtained in this dissertation as compared to the work of Assaad and El-adaway (2021c).

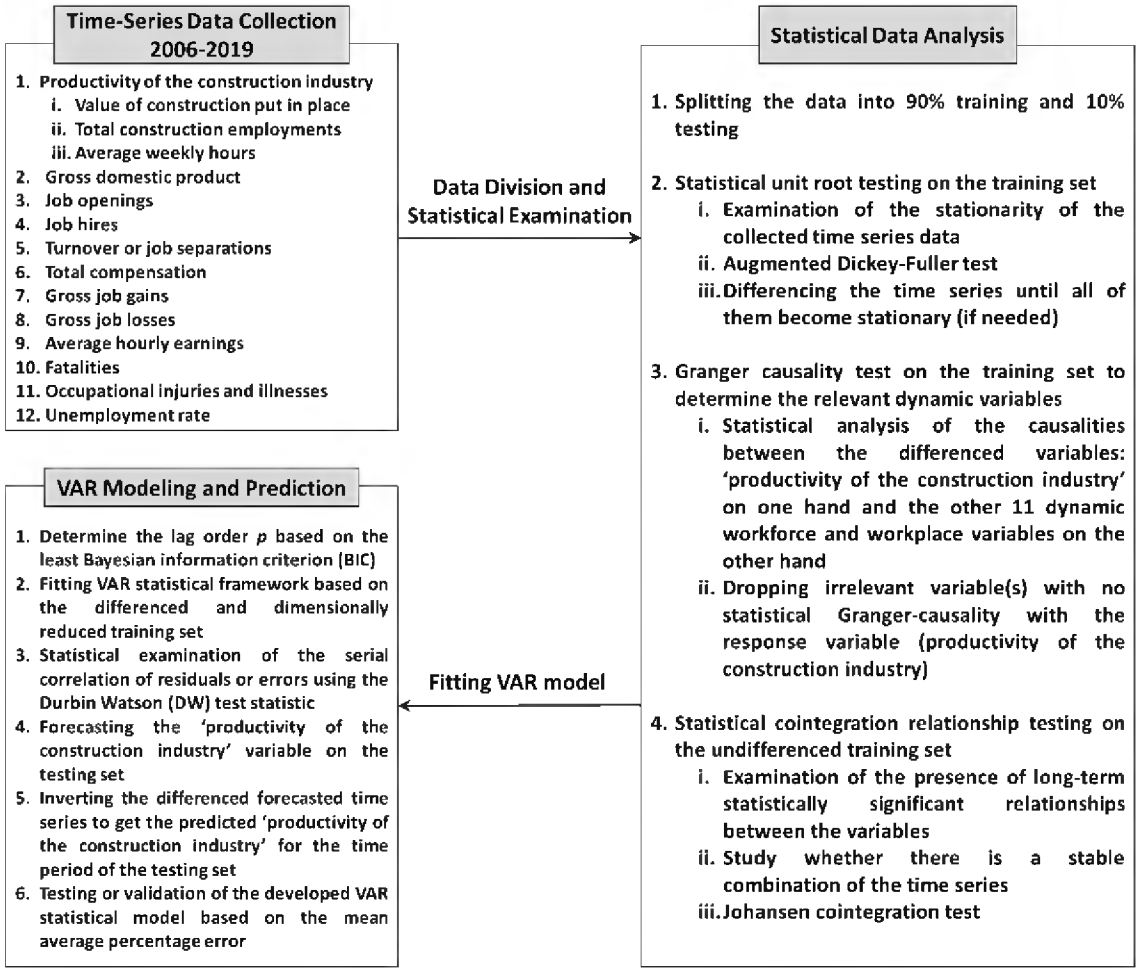

Figure 4.1 Research methodology (VAR: vector autoregression).

4.4.1. Data Collection and Description. The collected data included the following 14 dynamic workforce- and workplace-related variables: VIP, total construction employment, average weekly hours, GDP, job openings, job hires, turnover or job separations, total compensation [includes wages and salaries, and total benefits (insurance and retirement)], gross job gains, gross job losses, average hourly earnings, work-related fatalities, work-related occupational injuries and illnesses, and unemployment rate. The VIP data were collected from the US Census Bureau. The data for total construction

employment, average weekly hours, total compensation, unemployment rate, job openings, job hires, turnover, gross job gains, gross job losses, average hourly earnings, fatalities, and occupational injuries and illnesses were collected from the US BLS. GDP was collected from the Federal Reserve Economic Data (FRED) at the Federal Reserve Bank of St. Louis (data from US Bureau of Economic Analysis). All collected data were converted to a monthly time-step by interpolation and the addition of a variability or noise distribution whenever needed. Table 4.1 summarizes the variables used in this section of the dissertation and their associated sources.

Table 4.1 Used variables and their sources.

Variable	Source
Total compensation	US Bureau of Labor Statistics
Average hourly earnings	US Bureau of Labor Statistics
Unemployment	US Bureau of Labor Statistics
Fatalities	US Bureau of Labor Statistics
GDP	US Bureau of Economic Analysis
Hires	US Bureau of Labor Statistics
Occupational injuries and illnesses	US Bureau of Labor Statistics
Gross job gains	US Bureau of Labor Statistics
Gross job losses	US Bureau of Labor Statistics
Job openings	US Bureau of Labor Statistics
Turnover	US Bureau of Labor Statistics
Total construction employments	US Bureau of Labor Statistics
Value of construction put in place	US Census Bureau
Average weekly hours	US Bureau of Labor Statistics
Construction productivity[a]	US Bureau of Labor Statistics and US Census Bureau

[a]Calculated using Equation (12)

According to Sveikauskas et al. (2018), productivity can be calculated as the ratio of total construction output to total hours worked. The productivity of the construction

industry, which is the response variable, was obtained from the collected data using Equation (12).

$$Productivity = \frac{A}{B \times C \times D} \tag{12}$$

where A is the total value of construction put in place ($/month); B is total construction employments (employee/month); C is total average weekly hours (hours/weeks); and D = 4 (weeks/month).

Three of the collected variables were used to calculate the productivity of the construction industry: VIP, total construction employment, and average weekly hours. Therefore, the final set of the variables used in the statistical analysis included the calculated productivity of the construction industry using Equation (12) and the other 11 dynamic workforce and workplace variables. Although these variables have different scales of measurement, the statistical causalities and relationships between them can be examined. For instance, McGowan (2019) found that a 1% increase in the labor unemployment rate decreases construction value by $12 billion. Also, Shahandashti and Ashuri (2016) used time-series analysis and vector correction models to predict the national highway construction cost index (NHCCI) based on many potential leading indicators/variables such as total employment in the construction industry, average weekly hours, building permits, housing starts, and unemployment rate, among others; these variables have different scales of measurement.

A study period from 2006 to 2019, inclusive, was selected for the collected data/variables. This range was selected for the study period because the variable average weekly hours was not recorded before 2006. Because this variable was used to calculate the output variable total productivity of the construction industry [Equation (12)], no values for the model's output variable could be obtained before 2006. This means that no model could be developed before the 2006 year. In addition, the variable 'average hourly earnings' was not recoded before 2006. Henceforth, construction productivity refers to the productivity of the construction industry as a whole.

4.4.2. Statistical Analysis. The main statistical methods used were unit root testing, Granger causality testing, and cointegration testing.

4.4.2.1. Data division. The first step for the statistical analysis is to divide the data into training and testing sets before conducting any further analysis so that no data leakage occurs in the conducted statistical tests and the developed VAR model. That said, the data was divided into 90% training set and 10% testing set because the total number of data points (i.e., samples) is 168 which is relatively small. In other words, to make sure that enough data points are present for training the VAR model, a 90%-10% division scheme was applied as compared to other division schemes (such as 70%-30%, 80%-20%, or other division schemes). To this end, the statistical tests and the VAR model were conducted on the training set from January 2006 till July 2018 inclusive (which is 151 months because it is around 90% of the total 168 data points), and the VAR model was tested and validated on the testing set from August 2018 to December 2019 inclusive (which is 17 months because it is around 10% of the total 168 data points).

4.4.2.2. Unit root test. The minimum number of times *d* that a time series needs to be differenced (i.e., subtractions between consecutive observations) for transformation to a stationary time series is the called the order of integration (Faghih and Kashani, 2018; Shahandashti and Ashuri, 2013). In this case, the associated time series is said to be integrated of order *d* and denoted by I(*d*). In relation to that, Faghih and Kashani (2018) provided that time series should be successively tested for integration of order 0, I(0), the first order of integration, I(1), the second order of integration, I(2), and so on.

Since identifying the order of integration shall precede other statistical tests (Granger causality and cointegration tests) as only variables with the same order of integration can be further used (Shahandashti and Ashuri, 2013), the first test conducted was to check the stationarity of the time series on the training set (before conducting any other statistical test or analysis). This was stressed by Shahandashti and Ashuri (2013) by stating that a unit root test shall be conducted to identify the order of integration of the variables before implementing any further statistical tests. In addition, it is critical to identify whether the variables are stationary before developing the VAR model because the model can be applied only to stationary time series (Shahandashti and Ashuri, 2016). Stationary time series are those in which the means, variances, and autocorrelation structures do not change over time (Faghih and Kashani, 2018). Stationarity can be assessed using a unit root test (Lee et al., 2019a). If the time-series data are not stationary, they should be made stationary before developing the VAR model by differencing, which eliminates trends' changes by subtracting an earlier value from a later value. This section of the dissertation implemented the widely used augmented Dickey–Fuller (ADF) test (Dickey and Fuller, 1979). The formulation of the ADF test is shown in Equation (13).

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta y_{t-i} + u_t \quad (13)$$

where t = time index; y_t = time-series value corresponding to time t; Δy_t = lagged first differences (that is, $y_t - y_{t-1}$); α = intercept constant (drift term); β = coefficient of time trend; and δ coefficient to test if the data need to be differentiated to make them stationary.

The null hypothesis H_0 of the ADF test is that the time-series data have a unit root; that is, they are nonstationary. The ADF test generates a p-value that should be compared with a significance level of 0.05 to determine whether the null hypothesis is rejected. Differencing should be applied to all time series until all of them become stationary. The order of integration represents the number of differencings required to make a nonstationary time series stationary. Once all the time series are rendered stationary, the Granger causality test can be applied to study the causalities between the different variables and determine the explanatory variables causing fluctuations in the response variable (Abediniangerabi et al., 2017).

4.4.2.3. Determination of relevant variables using Granger causality testing. Before proceeding with developing a VAR model, it is quite common to conduct the Granger causality test, which is developed by Granger (1969). The Granger causality test is a statistical technique used in multivariate time-series analysis to examine whether the lagged values of one variable helps to predict another variable (Swei, 2020). In other words, the Granger causality test is used to identify the leading indicators or the relevant variables that affect the response variable (Shiha et al., 2020). In simple terms, if a multivariate time series comprises two variables y_{1t} and y_{2t}, then Equation (14) is used to investigate the one lagged causality between these two variables.

$$F(y_{1_t}, y_{2_t}) = a_1 * y_{1_{t-1}} + a_2 * y_{2_{t-1}} + b \quad (14)$$

The null hypothesis H_0 of the Granger test is that there is no causality between y_{1t} and y_{2t}, that is, $a_2=0$ [Equation (14)]. In other words, the null hypothesis states that y_{2t} does not Granger-cause y_{1t}. Failing to reject H_0 means that y_{2t} does not Granger-cause y_{1t}; therefore, there is no causality between them. In this case, the variable y_{2t} should not be considered in the set of variables used to develop the VAR statistical model. For a lag order *p*, the general statistical formulation of the Granger causality test for two variables can be represented by Equation (15) (Lütkepohl et al., 2004).

$$\begin{bmatrix} y_{1_t} \\ y_{2_t} \end{bmatrix} = \sum_{i=1}^{p} \begin{bmatrix} \alpha_{11_i} & \alpha_{12_i} \\ \alpha_{21_i} & \alpha_{22_i} \end{bmatrix} \begin{bmatrix} y_{1_{t-1}} \\ y_{2_t-1} \end{bmatrix} + u_t \quad (15)$$

where $i = 1, 2, \ldots, p$; and $\alpha_{12i} = 0$ if y_{2t} does not Granger-cause y_{1t}.

Since the Granger causality test is effective only when the time series are stationary (Shahandashti and Ashuri, 2016; Papana et al., 2014), the Granger causality test was conducted on the differenced, stationary, training set to establish the causalities between the response variable productivity of the construction industry and the other 11 potentially related variables with a significance level of 0.05. It is worth mentioning that a maximum lag of 24 months which is equivalent to 2 years was used for the Granger causality test. In relation to that, and according to Shahandashti and Ashuri (2013), a variable is considered explanatory or relevant to predict the output variable if the null hypothesis is rejected in at least one of the specified lag lengths for the Granger causality test. Therefore, only the

minimum p-values among the different considered lag lengths are reported in this section of the dissertation.

4.4.2.4. Cointegration relationship testing. After conducting the Granger causality test and determining the relevant predictor/input variables, the training set was dimensionally reduced so that to include only the relevant predictor variables. After that, the cointegration between variables was examined. Since the main aim behind the Granger causality test implemented in this section of the dissertation is to focus on one-way causality as to identify the predictor variables that statistically causes the output variable 'productivity of the construction industry'—as it was used in previous works such as Faghih and Kashani (2018), the cointegration test was implemented after the Granger causality test. In other words, it is not mandatory to test the cointegration before testing the causality but rather cointegration testing could be performed after Granger causality test as performed by many studies such as the work of Wong and Ng (2010) and Faghih and Kashani (2018) as examples. In relation to that, Faghih and Kashani (2018) stated that, to implement the cointegration test, a model needs to be created using the variables that the Granger causality test identified as indicators of the future values of the explained variable. This approach was used by many existing research studies such as Shahandashti and Ashuri (2013) and Shahandashti and Ashuri (2016), just to name few.

This section of the dissertation uses the Johansen's cointegration test due to its wide use (Mahadevan and Asafu-Adjaye, 2007). This cointegration test identifies the number of cointegrating relationships r in a group of variables, and the corresponding null hypothesis is that the number of cointegrating relationships r is less than or equal to a specific value that varies from 0 to the number of considered variables minus 1. For example, if $r \leq 0$ is

rejected for a set of variables, then it can be concluded that there is at least one cointegrating relationship in the considered set if variables. According to Faghih and Kashani (2018), to implement the cointegration test, an unrestricted VAR model needs to be created using the variables that the Granger causality test identified and the lag length is determined in such a way that the model selection criterion is minimized (Faghih and Kashani, 2018). In relation to that, the Akaike information criterion (AIC) is commonly used for the cointegration testing as implemented by Shahandashti and Ashuri (2016) and Ashuri et al. (2012), just to name few. Thus the, the AIC was used to identify the lag lengths for the cointegration testing based on the minimum AIC. It is to be noted that the Johansen cointegration test is implemented by assuming or considering that the data has no deterministic terms.

While the Johansen's cointegration test is a procedure for testing cointegration of several, say k, I(1) time series (Johansen, 1991), it can also be used for I(2) time series (Johansen, 1995). Moreover, it is to be noted that although a good deal of progress has been made when it comes to modelling with a mixture of, for example, both I(2) and I(1) data, there are still lots of important questions waiting to be addressed (Giles, 2012) as there is no unified, common, or standard way to deal with variables with a mixture of integration orders (as it is the case in this section of the dissertation where some of the variables are I(0), other variables are I(1), and others are I(2) as detailed later). However, one simple and recommended way in such cases is to make sure that all the variables should be integrated of the same order (Kestel, 2013), which was performed in this section of the dissertation. In other words, Faghih and Kashani (2018) stated that a prior condition for a cointegration test is that all of the variables under evaluation should be integrated in the same order.

Furthermore, it is worth mentioning that the Johansen's cointegration test should be performed on the level form of the variables and not on their differenced forms (Adeleye, 2018).

The purpose of the cointegration test is to study the long-term relationships between the variables (Liu et al., 2019a). The long-term relationship is investigated by the cointegration test, which identifies whether two or more time series are integrated together in a way that they cannot deviate from equilibrium in the long term (Corporate Finance Institute, 2020). A long-term relationship exists between variables when they share a similar trend by being associated together over time rather than being correlated at a specific or instantaneous time. In simple terms, even if the variables deviate from each other in the short-term, they tend to return to the trend in the long-term if they are cointegrated (Wei, 2016). Erica (2020) provides more details on the cointegration test and the long-term relationship between variables. In addition, the cointegration test analyzes whether there is a stable combination of the different time-series data (Lee et al., 2019a). When two or more time series are cointegrated, this indicates they have a long-term statistically significant relationship. The null hypothesis H_0 of the Johansen test is that there is no cointegration between the variables (Elliott and Pesavento, 2009). The trace statistic is calculated for the Johansen cointegration test, and it is compared with the critical trace statistic value at 95% (Rachev et al., 2007). The trace test rejects the null hypothesis if the trace statistic exceeds the critical value. In other words, if the null hypothesis is rejected, there is a statistically significant relationship between the corresponding two variables. Ultimately, according to Adeleye (2018), if there is cointegration then this (1) implies that the series in question are related and therefore can be combined in a linear fashion; (2)

means that even if there are shocks in the short run, which may affect movement in the individual series, the series would converge with time (in the long run); (3) reflects that both long-run and short-run models shall be estimated; and (3) provides that the estimation will require the use of VAR modeling as well as vector error correction model (VECM) analysis. On the other hand, and according to Adeleye (2018), if there is no cointegration then only the VAR (and not the VECM) shall be estimated (i.e., only short-run model shall be of interest). In relation to that, according to Giles (2011), no matter what is concluded about cointegration, such conclusion will not affect the steps of the analysis but rather it provides a possible cross-check on the results. According to Kestel (2013), if the variables are non-stationary in their level forms and they are not cointegrated, then they first have to be differenced *d* times (i.e., until they become stationary) and then a VAR model is developed in difference, which is the approach followed in this section of the dissertation.

4.4.2.5. Vector autoregression modeling and prediction. The lag order *p* for the final VAR model can be iteratively determined by fitting increasing orders of the VAR model on the differenced, stationary, dimensionally reduced training set and by choosing the lag order *p* that gives a model with the lowest Bayesian information criterion (BIC) (Swei et al., 2017; Ayhan and Tokdemir, 2020). When the lag order *p* is determined, the VAR statistical model is fitted to quantitatively examine the relationships between the different time series.

For two time series y_{1t} and y_{2t}, the associated VAR statistical model for a lag order of 2 is shown in Equation (16).

$$F(y_{1_t}, y_{2_t}) = a_{11} * y_{1_{t-1}} + a_{12} * y_{1_{t-2}} + a_{21} * y_{2_{t-1}} + a_{22} * y_{2_{t-2}} + b \quad (16)$$

If the VAR model is applicable to multivariate time series with *K* variables and a lag order of *p*, the general representation of a VAR model is shown in Equation (17) (Lütkepohl et al., 2004).

$$Y_t = A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + U_t \tag{17}$$

where $Y_t = (y_{1t}, \ldots, y_{Kt})$ is a set of *K* time-series variables; A_p = (k x k) coefficient matrix; and U_t = set of unobservable error terms (Lütkepohl et al., 2004).

After fitting the VAR model based on the differenced, stationary, dimensionally reduced time-series training data, the examination of the serial correlation of residuals (errors) should be conducted (Prabhakaran, 2019). Serial correlation is important to check if there is any leftover pattern in the obtained residuals (errors) from the VAR model (Prabhakaran, 2019). Checking for serial correlation ensures that the fitted VAR model is sufficiently able to explain the variances and patterns in the time series (Prabhakaran, 2019). The Durbin–Watson (DW) statistic was calculated using Equation (18) to examine the serial correlation of residuals.

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \tag{18}$$

where DW = Durbin–Watson statistic; and e_t = residuals at time *t*.

The DW statistic will always assume a value between 0 and 4 where a value close to 2 indicates that there is no autocorrelation (Corporate Finance Institute, 2021). When the value is considerably below 2, it indicates a positive autocorrelation, and where the value

is considerably higher than 2 indicates a negative serial correlation (Corporate Finance Institute, 2021). In relation to that, values of DW test statistic in the range 1–3 are relatively normal, and any value outside this range is a cause for concern (Field, 2013).

Because the VAR model was fitted on the stationary differenced data series, the final predictions were obtained by inverting (that is, integrating) the predicted differenced data series. The prediction accuracy of the fitted VAR statistical model was determined based on the mean absolute percentage error (MAPE) shown in Equation (19).

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{19}$$

where MAPE = mean absolute percentage error; n = number of observations during prediction period; y_t = actual value; and $\hat{y}_t$ = predicted value using developed VAR model.

The Python programming language was used for data management, statistical analysis, and visualization. The following libraries were used: pandas, which generally is used for data manipulation/management of numerical tables and data structures; numpy, which generally is used to add support for large and multidimensional arrays and matrices and to provide high-level mathematical functions; matplotlib, which generally is used for data plotting and visualization; and statsmodels, which generally is used to conduct statistical tests, statistical data exploration, and estimation of different statistical models.

4.5. STATISTICAL EXAMINATION

This subsection provides all details related to statistical examination of the data. Figure 4.2 shows the collected time series data with monthly time interval.

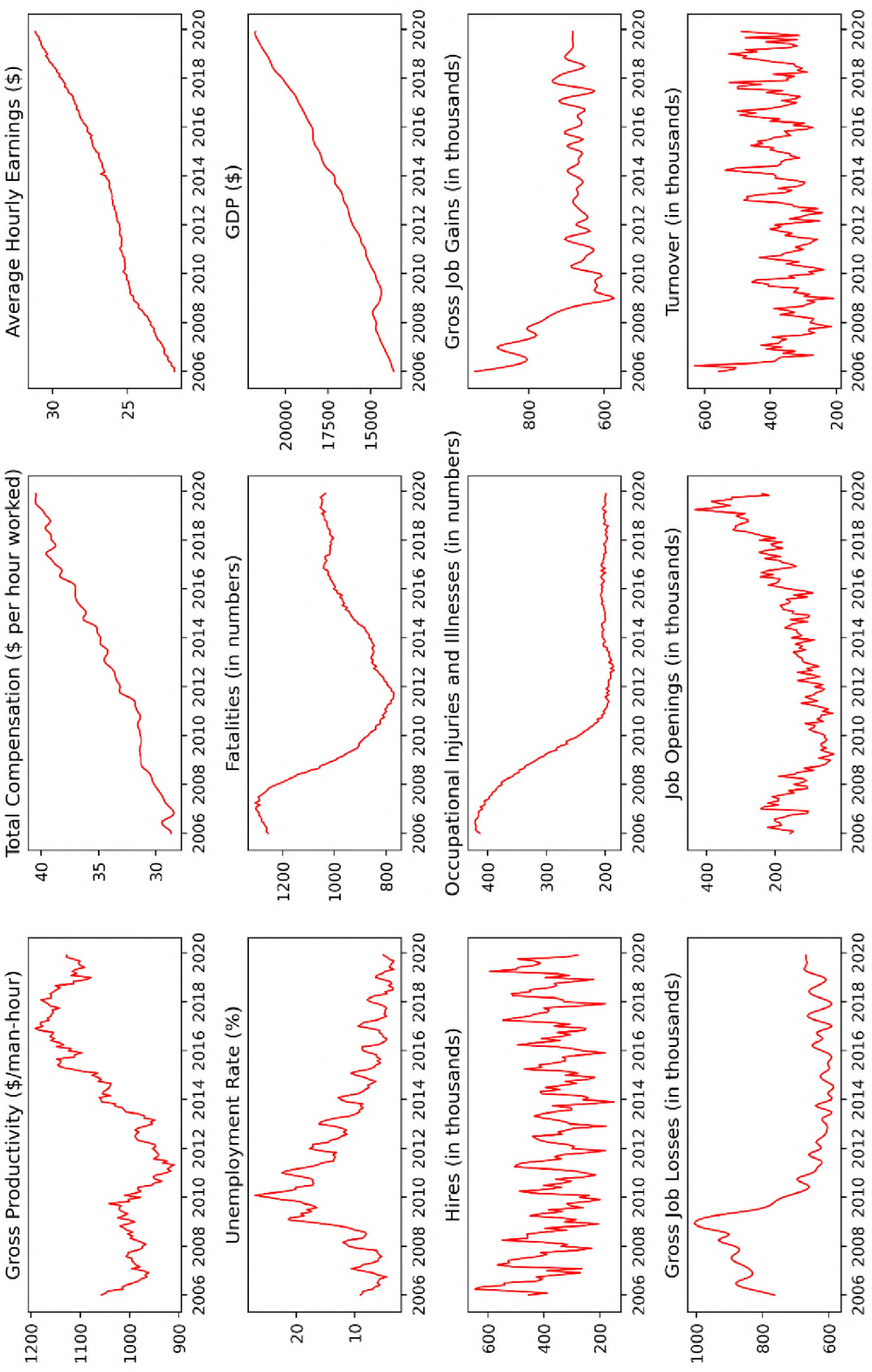

Figure 4.2 Monthly time series data.

After collecting and visualizing the data, the stationarity of the data series was investigated. The obtained results of the unit root test are listed in Table 4.2.

Table 4.2 Obtained results of the stationary test.

Variable	p-value for no differencing	p-value for first differencing	p-value for second differencing
Construction productivity	0.8738	<0.0001[a]	<0.0001[a]
Total compensation	0.9593	0.0056[a]	<0.0001[a]
Average hourly earnings	0.9728	<0.0001[a]	<0.0001[a]
Unemployment	0.0351[a]	0.4669	<0.0001[a]
Fatalities	0.0055[a]	0.4348	<0.0001[a]
GDP	0.9987	0.3297	0.0004[a]
Hires	0.2609	0.0302[a]	<0.0001[a]
Occupational injuries and illnesses	0.0007[a]	0.6766	<0.0001[a]
Gross job gains	0.0051[a]	0.0439[a]	<0.0001[a]
Gross job losses	0.5895	0.0069[a]	<0.0001[a]
Job openings	0.9442	0.0208[a]	<0.0001[a]
Turnover	0.4488	0.001[a]	<0.0001[a]

[a]p-value < 0.05 significance, meaning that the null hypothesis is rejected; therefore, the series is stationary.

Only four of the original collected multivariate time-series data were stationary (Table 4.2, Column 2): unemployment, fatalities, occupational injuries and illness, and gross job gains. Therefore, the data should be differenced so that all time series are stationary before constructing the VAR model. After the first differencing (Table 4.2, Column 3), eight variables were stationary: construction productivity, total compensation, average hourly earnings, hires, gross job gains, gross job losses, job openings, and turnover. Because not all time series were differenced, a second differencing of the data was needed. After the second differencing of the data, all the time series were rendered stationary (Table 4.2, column 4).

Before building the VAR statistical model, the Granger causality test was conducted. The obtained results are listed in Table 4.3. The null hypothesis was rejected for six variables. In other words, and since the null hypothesis means that the input variable does not Granger-cause the output variable (Shahandashti and Ashuri, 2016), the following variables were considered to Granger-cause the variable productivity of the construction industry because they have a p-value less than or equal to the significance level of 0.05: total compensation; average hourly earnings; unemployment rate; GDP; occupational injuries and illnesses, and job openings. Therefore, only these variables were included in the VAR statistical model developed for modeling and predicting the productivity of the construction industry.

Table 4.3 Results of Granger causality test between productivity of construction industry and other dynamic workforce and workplace variables.

Null hypothesis (H_0)	p-value
Total compensation does not Granger-cause construction productivity	0.0122[a]
Average hourly earnings do not Granger-cause construction productivity	0.0416[a]
Unemployment does not Granger-cause construction productivity	<0.0001[a]
Fatalities do not Granger-cause construction productivity	0.0725
GDP does not Granger-cause construction productivity	0.0086[a]
Hires do not Granger-cause construction productivity	0.0526
Occupational injuries and illnesses do not Granger-cause construction productivity	0.0114[a]
Gross job gains do not Granger-cause construction productivity	0.3037
Gross job losses do not Granger-cause construction productivity	0.0888
Job openings do not Granger-cause construction productivity	0.0055[a]
Turnover does not Granger-cause construction productivity	0.0789

[a]*p*-value < 0.05 significance, meaning that the null hypothesis is rejected; therefore, there is causality with the response variable productivity of the construction industry

The cointegration relationships between the selected/relevant variables and the output variable (all combined) was examined using the Johansen's test statistic, and the obtained results are given in Table 4.4.

Table 4.4 Obtained results for the cointegration test.

Hypothesis	Obtained test statistic	Critical test statistic at 95%	Cointegration?[a]
$r = 0$	325.78	111.7797	True
$r \leq 1$	207.8	83.9383	True
$r \leq 2$	116.64	60.0627	True
$r \leq 3$	63.48	40.1749	True
$r \leq 4$	29.97	24.2761	True
$r \leq 5$	8.66	12.3212	False
$r \leq 6$	0.0	4.1296	False

[a]If the obtained test statistic is greater than the critical test statistic at 95%, then the null hypothesis corresponding to the rank level is rejected.

The rank $r = 0$ refers to the null hypothesis H_0 that there is no cointegration between the variables. Since the obtained test statistic for $r = 0$ was greater than the critical test statistic at 95% (Row 2 in Table 4.4), then the null hypothesis H_0 is rejected, which reflects that there is cointegration between at least 1 of the variables. Similar logic and analysis could be applied to the cases where $r \leq 1$, $r \leq 2$, $r \leq 3$, and $r \leq 4$ which reflect that there are at least 2 cointegrating relationships (i.e., $r > 1$), at least 3 cointegrating relationships (i.e., $r > 2$), at least 4 cointegrating relationships (i.e., $r > 3$), and at least 5 cointegrating relationships (i.e., $r > 4$), respectively. For the rank cases where $r \leq 5$ and $r \leq 6$, since the obtained test statistic was less than the corresponding critical test statistic at 95%, then the associated null hypotheses cannot be rejected. This reflects that there are at most 5 cointegrating relationships (i.e., $r \leq 5$) and at most 6 cointegrating relationships (i.e., $r \leq 6$),

respectively. It could be concluded from the results in Table 4.4 that not *all* variables are cointegrated. Hence, only the VAR (without VECM) shall be estimated (Adeleye, 2018).

4.6. VECTOR AUTOREGRESSION MODEL AND PREDICTION

The VAR statistical framework was developed between the response variable productivity of the construction industry, its lagged values, and the following dynamic workforce and workplace variables determined by the Granger causality test: total compensation; average hourly earnings; unemployment rate; GDP; occupational injuries and illnesses, and job openings. The VAR model was fitted on the stationary, differenced, dimensionally reduced, training data series. The lag order p for the VAR model was determined by fitting increasing orders of VAR models and choosing the lag order p with the lowest BIC. The obtained BIC results for the different lag orders are listed in Table 4.5.

Table 4.5 BIC results for different VAR lag orders.

Lag order *p*	BIC
0	14.7
1	12.81
2	11.27[a]
3	12.04
4	12.91
5	13.41
6	14.44
7	15.31
8	16.41
9	17.32
10	17.55
11	18.38
12	19.04
13	19.7

[a]Lowest BIC; therefore, a lag order of 2 was selected for the VAR model.

The minimum BIC corresponded to a lag order p of 2 (Table 4.5). Therefore, a lag order of 2 was selected for developing the VAR statistical model for the construction productivity variable. A VAR model was fitted, and the obtained results are listed in Table 4.6.

Table 4.6 Fitted VAR statistical model for the construction productivity.

Lagged variable	Coefficient[a]
L1.Gross Productivity	-0.80534
L1.Total Compensation	24.19133
L1.Average Hourly Earnings	-21.2219
L1.Unemployment Rate	-0.44377
L1.GDP	0.0068
L1.Occupational Injuries and Illnesses	-0.37586
L1.Job Openings	0.067265
L2.Gross Productivity	-0.25914
L2.Total Compensation	-19.4218
L2.Average Hourly Earnings	-6.44732
L2.Unemployment Rate	-1.08619
L2.GDP	-0.0214
L2.Occupational Injuries and Illnesses	-0.51691
L2.Job Openings	0.045543
Constant	0.151559

[a]Coefficients were obtained for the differenced (with an order of 2) data series, and thus they do not reflect the direct impact of the different lagged variables on the value of the construction productivity itself.

Table 4.6 gives the developed VAR statistical model. The VAR model was developed by obtaining the coefficient associated with each lagged variable. For instance, the coefficient for the one-lagged differenced productivity variable is -0.80534, the coefficient for the one-lagged differenced total compensation variable was 24.19133, and so on (see Table 4.6). The highest coefficient in absolute value was that related to the one-lagged differenced total compensation variable, which means that this lagged variable

contributed the most to the fluctuations in construction productivity. The lowest coefficient in absolute value was that related to the one-lagged differenced GDP variable, which indicates that this lagged variable contributed the least to the fluctuations in construction productivity.

The serial correlation of the errors was checked using DW statistic. This is important to examine if there is any leftover pattern in the residuals (errors) obtained from the VAR model. The DW statistic obtained was 2.2, which is between 1 and 3; therefore, the developed VAR model is sufficiently able to explain the variances and patterns in the time series (Prabhakaran, 2019; Field, 2013). Consequently, the developed VAR model was tested and validated by predicting the construction productivity on the testing set. The final predictions were generated after inverting the predicted differenced data series. To assess and validate the developed VAR statistical model, the MAPE was calculated to evaluate the model's predictions. Figure 4.3 shows the actual and predicted values for the productivity of the construction industry using the developed VAR model.

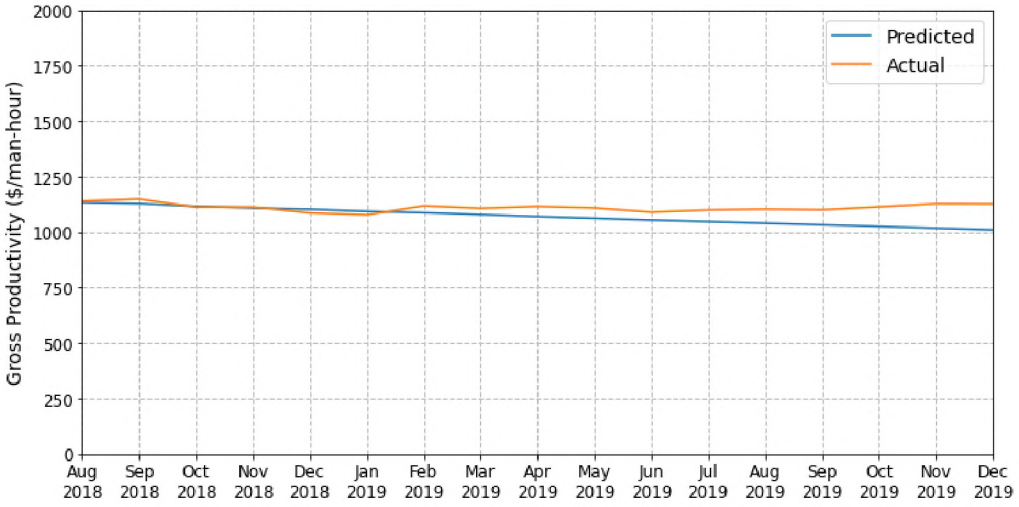

Figure 4.3 Prediction of the construction productivity using the fitted VAR model.

The developed VAR statistical model for the productivity of the construction industry had a MAPE of 4%, which is considered to be acceptable because it is less than 10%, which is the generally accepted MAPE for a robust prediction model (Fan et al., 2010). Hence, the developed VAR statistical model is perceived to be a good framework that can be used to model and predict the productivity of the entire construction industry.

4.7. ANALYSIS OF FINDINGS

The findings of the unit root test indicated that when the collected times-series data were differenced twice, all the time-series variables became stationary. This means that there was no stochastic trend in the differenced time series, which often is referred to as random walk with drift (Shifera, 2019). Therefore, the differenced construction productivity variable and the dynamic workforce and workplace variables did not possess a unit root, and thus they did not have a pattern that is unpredictable (Sodiq Olawale, 2019). According to Glen (2016b), this is a desirable property because the existence of a unit root can create serious issues, including spurious regressions (a high coefficient of determination even if the data are uncorrelated) and errant behavior or invalid inferences (due to assumptions for analysis not being valid). The findings of the Granger test reflected on the structures of the causal relationships between the different variables and the response variable construction productivity. Specifically, the findings highlighted that six of the variables (6 of the 11 dynamic workforce and workplace variables) are useful for predicting construction productivity. Therefore, it is inferred that the fluctuations in construction productivity are associated with corresponding changes in the 6 predictor variables, and hence have a robust pattern of Granger causality. In other words, construction productivity

is affected by 6 of the dynamic workforce- and workplace-related variables considered in this section of the dissertation. Thus, the obtained results justified the rationale behind conducting this research and have statistically proved the concept of productivity for the entire construction industry.

The obtained findings from the VAR model provided insights on the particular relationships between the variables. More specifically, the dynamic nature of the variables was determined by investigating different orders for the leads and lags in the time series data. It was found that a lag order of 2 best modeled the fluctuations in construction productivity. This indicates that the changes in the workforce- and workplace-related variables for two periods or steps have a relevant impact on the construction productivity. Because it is difficult to interpret the large number of coefficients in a VAR model (Stanslaus, 2017), this section of the dissertation did not focus on the individual interpretation of the obtained coefficients, however the coefficients could still be of value to better understand the nature of the impact of each dynamic variable on the construction productivity. To check if the developed VAR model is sufficiently able to explain the variances and patterns in the time series, the serial correlation of the error terms or residuals was investigated using the DW statistic. Serial correlation occurs when the error terms of a time series are transferred from one period to another; that is, when the error in a period is correlated with the error in a subsequent period (Glen, 2016a). For example, an overestimate of the construction productivity for one month would result in an overestimate of the construction productivity for the subsequent month. This would lead to biases, and thus serial correlation of error terms is not desired because it can lead to myriad problems, including but not limited to inefficient estimation of the coefficients, underestimation of

the error variance, and underestimation of the variance of the coefficients, among many other issues (NCSS, 2006). The findings indicate that the obtained DW statistic was 2.2, which is a desirable value because it lies between 1 and 3 (Field, 2013). Thus, the findings concluded that the developed VAR statistical framework is sufficiently able to explain the variances and patterns in the time series. Finally, the prediction accuracy of the developed VAR model was acceptable, with a MAPE of 4%.

4.8. SUMMARY

This section of the dissertation identified a critical gap in the body of knowledge in terms of construction productivity. Specifically, previous research works concentrated on labor productivity, without studying whether the productivity could be assessed for the entire construction industry. This section of the dissertation examined and quantified the impacts of different workforce and workplace variables on the productivity of the entire construction industry. Empirical multivariate time-series data was collected for the period from 2006 to 2019, inclusive, for construction productivity and for 11 dynamic workforce and workplace variables: job openings, job hires, turnover or job separations, total compensation, gross job gains, gross job losses, average hourly earnings, fatalities, occupational injuries and illnesses, gross domestic product, and unemployment rate. Causalities and statistically significant relationships were examined. Relying on statistical test for the collected data series, a VAR framework was developed based on six dynamic workforce and workplace variables to model the temporal variations in the construction productivity. The developed VAR model was validated by predicting the construction productivity on the unseen testing set with an acceptable MAPE of 4%. The outcomes of

this section of the dissertation add to the body of knowledge by providing a better understanding of the impact of different dynamic workforce and workplace variables on the construction productivity and by offering a new concept called gross construction productivity.

4.9. RELATED APPENDIX

Appendix C presents the used data and the Python code for the developed statistical and time series model used to study the impact of dynamic workforce and workplace variables on the productivity of the construction industry.

5. A HYBRID UNSUPERVISED COMPUTATIONAL MODEL FOR DETERMINING THE CRITICAL COMBINATIONS OF SAFETY FATALITY CAUSES

5.1. OVERVIEW

The construction industry is considered one of the key contributors to the growth of economies. For instance, around 4.4% ($1,608.4 billion) of the total gross output is attributed to the construction sector in the United States (US Bureau of Economic Analysis, 2018a). In addition, the construction industry currently employs 6%–10% of the world's workforce (Shohet et al., 2019), and it is responsible for 30%–40% of the occupational fatalities (Sunindijo and Zou, 2012). Many challenges have been present for decades in the construction industry and are still not sufficiently addressed by practitioners and scholars, one of such challenges include safety performance. Although many efforts were directed to enhance the safety performance in the construction industry, safety fatalities and injuries are still considered to be a plague (Zhang and Fang, 2013). In fact, the construction industry is considered to be the most dangerous sector in many nations (Shin et al., 2014) because it is ripe for many disruptions (Assaad et al., 2020d). For instance, in the United Kingdom, one third of all workplace fatalities occur on construction sites; which is equivalent to a fatal injury rate of more than four times the average rate of all sectors (Health and Safety Executive Construction Division, 2009). In Korea, the construction industry possesses the highest percentage of fatalities among the different industries (Yi et al., 2012).

Although the Occupational Safety and Health Administration (OSHA) has played a key role in trying to decrease the safety accidents on construction sites, such reduction is not well perceived among the different types of accidents. For instance, while the injuries

resulting in days away from work were substantially decreased from 219,000 in 1994 to 80,000 in 2015 (CPWR, 2018), the construction fatalities rose by 26% in the previous years compared to its lowest level of 781 deaths in 2011 (CPWR, 2018). This rise in fatality accidents is attributed to the fact that the construction site is considered one of the most hazardous and dangerous workplaces (Lee et al., 2017). As such, there is a critical need to further enhance the safety fatality performance in the construction industry.

Despite the improvements in the construction safety that were made in the last decades (Hinze et al., 2013), safety accidents and injuries still happen on construction sites, and the construction safety is believed to have attained a plateau (Bhattacharjee et al., 2011). In fact, the construction industry has poor safety performance (Abdul Nabi and El-adaway, 2020b) and high incident rates compared with other industries. According to the recent 2017 statistics in the United States, the construction industry possesses the highest number of fatal work injuries of 971 among all other high-risk sectors including transportation and warehousing, agriculture, professional and business services, and manufacturing, among others (BLS, 2018). In fact, the fatality rate is equivalent to around 20% of the total 4,836 fatal injuries at workplaces in the US, more than any other industry (CPWR, 2018).

Despite that a number of techniques have been developed to improve project management and the decision-making process in the construction industry, more advanced management methods should be applied (Yang et al., 2019) to solve the critical fatality problem that the construction industry is facing. However, such methods were hard to achieve due to the unique nature and characteristics of the construction site's operations being hazardous, dangerous, complex, uncertain, and highly labor intensive (Tixier et al.,

2016; Lee et al., 2017). That said, to build the needed advanced methods, there is a need to understand the inherent complexities, abstractness, and uncertainties related to the safety fatality causes or factors and their interdependencies and combinations (Liu et al., 2019b).

5.2. OBJECTIVE

The purpose of this section of the dissertation is to determine the critical combinations between the causes leading to fatalities in the construction industry. The associated objectives are to (1) categorize the different construction fatality causes based on their interconnectivities, (2) identify the key causes leading to construction fatal accidents, and (3) determine and quantify the critical combinations and associations between the different key fatality causes in the construction sector. It is worth mentioning that this section of the dissertation distinguishes between two different levels of safety causation: direct causes and root causes. Direct causes are the most immediate causes leading to the fatality recorded directly by OSHA compliance officers and found readily in case files such as fall, struck by, caught in between, and electrocution. On the other hand, root causes are defined by OSHA (2016) as "fundamental, underlying, system-related reasons why an incident occurred that identify one or more correctable system failures." In relation to that, according to Goldberg (2003), focusing solely on the direct or immediate causes of safety accidents, rather than the root or underlying causes, does not preclude the recurrence of the incidents. That said, since many previous studies were conducted to study the direct safety causes, this section of the dissertation focuses mainly on studying the relationships and associations between the root causes themselves and how they combine together to lead to fatal construction accidents.

5.3. CURRENT STATE OF LITERATURE ON SAFETY AND BACKGROUND INFORMATION

Many previous research efforts were conducted to enhance the safety performance in the construction industry. However, the construction safety research topic includes many different aspects, thus, it is important to provide a review of only relevant previous research efforts to properly identify the knowledge gap and research need that this section of the dissertation addresses. More specifically, this subsection presents a review of the literature on accident analysis in the construction industry and other closely related industries.

5.3.1. Previous Safety Research Work on Accident Analysis in the Construction Industry. As far as the construction industry is concerned, many research studies on accident analysis were conducted. In relation to that, Ayhan and Tokdemir (2020) used latent class clustering analysis and artificial neural networks to conduct an accident analysis for safety attributes in the construction industry based on data collected from incident reports from construction sites located in the Euro-Asia region. Chiang et al. (2017) offered an analysis of the fatal accidents in terms of the time of their occurrence and how they occur in Hong Kong construction trades and found that more fatal accidents occurred in repair, maintenance, alteration, and addition works. Choi et al. (2019b) compared the fatal occupational injuries in the construction industry between the US, South Korea, and China and showed that these industries have high fatal injuries and that the top common accident types are falling from a high level and struck by. Meng et al. (2018) studied the fatal accidents that occurred in the Chinese construction industry and investigated the effect of climate factors, the time period distribution, and the provincial distribution. Karimi and Taghaddos (2019) examined the influence of education and experience level of construction craft labors in fatal injury prevention in construction

projects in Tehran; they showed that in the majority of accidents, higher educational attainment in craft workers resulted in a significant lower risk of fatal injuries. Al-Bayati and York (2018) analyzed the data from the Fatality Assessment and Control Evaluation program to study the fatal injuries among Hispanic workers in the US construction industry and found that there is a difference in accident characteristics between Hispanic workers and all workers. Wong et al. (2016) developed a framework for extracting structured information from construction accident reports as to find the general characteristics and managerial causes behind the falls from height fatalities accidents in Hong Kong. Zhao and Shi (2019) examined construction fatality reports and investigated the sociotechnical systems of fatal electrical injuries in the construction industry by using a triangulation approach; in relation to that, they identified the typical systems, revealed their weaknesses, and provided remedial recommendations. Shao et al. (2019) explored the fatal accident patterns in China's building construction activities and found that more fatal accidents occur in certain months, on particular days, and during certain time periods. Eteifa and El-adaway (2017) studied the root causes of construction fatalities and found that the lack of job-specific training is the most central cause of struck by and caught in between accidents. Dong et al. (2017) analyzed the construction Fatality Assessment and Control Evaluation database to study the fatal falls and personal fall arrest systems used in the construction industry.

5.3.2. Previous Safety Research Work on Accident Analysis in Roads and Transportation Systems. As far as roads and transportation systems are concerned, many research studies on accident analysis were conducted. In relation to that, Sayed et al. (1995) identified accident-prone locations using fuzzy pattern recognition and cluster analysis

based on an assessment of factors that contributed to accidents. Li et al. (2016) analyzed the traffic accidents on highways using the latent class clustering method to identify the key factors leading to severe traffic accidents based on data collected from China's State Administration of Work Safety. Yang et al. (2014) relied on the orderly clustering approach to provide proper division of highways in China based on four accident indicators (accidents, deaths, injuries, and direct economic loss). Jia et al. (2011) applied the principal component analysis method to road traffic safety in Jiamusi city of China. Mousavi et al. (2019) used Jerk-Cluster analysis to identify high crash risk highway segments so that observational data can be used as surrogate measures of safety and as a way of predicting safety problems. Chen et al. (2018) extracted arterial access density impacts on safety performance based on clustering and computational analysis. Rahimi et al. (2019) developed a clustering approach towards the analysis of large truck crashes. Wang and Chen (2013) studied the traffic safety evaluation using the traffic conflict technique and gray clustering at signalized intersections in China. Zhang et al. (2019b) used machine learning methods to identify significant injury severity risk factors in traffic accidents. Chen et al. (2020) used association rules to analyze the factors that influence expressway traffic crashes on the Shaoyang–Xinhuang section of the Shanghai–Kunming expressway in China. Wang et al. (2020a) developed a deep clustering model to estimate the driving style for a better safety performance on road and for a better design of personalized autonomous driving. Poch and Mannering (1996) estimated a negative binomial regression model for the frequency of accidents at urban intersections to reduce traffic accidents. Khasnabis and Ramiz-Al-Assar (1989) presented an exposure-based technique that uses the concept of opportunity for interaction between vehicles to analyze the heavy truck

accidents. Al-Masaeid and Sinha (1994) applied a probabilistic procedure to evaluate the safety effectiveness of pavement markings of undivided rural roads. Zhong et al. (2007) used the ordinal clustering method to propose a new method of freeway section division based on the safety profile and important safety influence factors. Zhou and Irizarry (2016) proposed an integrated framework of modified accident energy release to explore the complexity of the Hangzhou subway collapse in China. Medina et al. (2014) performed a study to enhance the analysis of accidents at railroad grade crossings based on detailed accident information such as location and direction of vehicles and trains, crossing layout and its surroundings, driver demographics, and others. Li et al. (2015) conducted an accident analysis for traffic system stability in China based on different accident-related factors such road environment, traffic facilities, and vehicles. Xie et al. (2020) analyzed the spatial characteristics of traffic accidents and the relationship between the accident rate and the linear parameters of a two-lane highway in China using statistical analysis and the K-means clustering based on data of traffic accidents, traffic flow, and road profile. Jun et al. (2009) used principal component analysis to study the influencing factors leading to rear-end collisions on highways, and the significant impact factors are detected and identified by charting analysis and the clustering method. Lin et al. (2011) used principal component analysis-cluster analysis method to perform a macroscopic evaluation of road traffic safety and the associated safety level classification.

5.3.3. Previous Safety Research Work on Accident Analysis in Other Industries. Other studies were also performed as related to fire accidents, maritime accidents, and mine accidents. In relation to that, Xie et al. (2019) used the fuzzy C-means algorithm and fuzzy maximum support tree clustering algorithm to identify the key factors

of fire risk of oil depots in China. Balahadia et al. (2019) used a data mining approach for profiling fire incident reports of the Bureau of Fire and Protection in Manila, Philippines to help facilitate the assessment of fire incidents by providing a quick, thorough, and scientific analysis of fire data. Yuan et al. (2020) used case statistics and dynamic Bayesian networks for scenario deduction on fire accidents to help decision makers make more targeted emergency disposal measures. Kim and Shin (2020) used classification machine learning algorithms to improve the classification of fire accidents and the analysis of periodicity to provide the ability to forecast critical fire accidents. Ntzeremes et al. (2020) proposed an evacuation simulation model for increasing the efficiency of quantitative risk assessment of fire accidents. Yuan et al. (2019) used interpretative structural modeling and the analytic hierarchy process method to identify the cause factors in the emergency process of fire accidents based on 23 influence factors. Wu et al. (2019) used computer vision methods to create an intelligent fire detection approach to meet the needs of real-time fire detection on the precision and the speed. Zhang et al. (2014) proposed an approach to analyze and predict the combination patterns of human factors for maritime accidents using matrix transformation, the clustering method, and Bootstrapping. Zhao and Shi (2019) used density-based clustering and recurrent neural networks to detect maritime anomaly for an improved situational awareness of vessel traffic supervisors and a reduction in maritime accidents. Wuellner et al. (2019) applied principal component analysis and K-mean clustering for the environmental conditions of historical accident data to efficiently generate testing sceneries for maritime systems. Wang and Sun (2011) performed an accident causation chain analysis of ship collisions using Bayesian networks to avoid maritime accidents. Guo et al. (2020) used the Grey-Buffet Operator-Markov chain method

to forecast the civil aviation unsafe events rate. Wu (2020) performed an in-depth study of coal mine safety risk using a data analysis approach to develop a system that can dynamically display the safety production situation of coal mines and predict the risk of coal mine accidents. Wang et al. (2019a) used neural networks for automatic construction of coal mine accident ontology. Mulenga (2020) developed a mathematical model using the Gray Markov approach to analyze Zambia's fatal mining accidents.

5.3.4. Knowledge Gap and Research Need. According to the extensive review of the existing body of literature on safety accidents performed in the previous subsections, it could be concluded that many previous research works have provided important knowledge on safety accidents using different methodologies. Nevertheless, some limitations and gaps still exist in the body of literature.

First, many of these studies were conducted to study the factors leading to safety accidents in general rather than the factors leading to safety fatalities in specific. In relation to that, Zhang and Fang (2013) stated that safety fatalities and injuries are still considered to be a plague, despite the presence of many efforts that were directed to enhance the safety performance in the construction industry. This is also reflected by the statistics published by CPWR (2018) which showed that while the injuries resulting in days away from work were substantially decreased, the construction fatalities rose in the previous years.

Second, as detailed in the previous subsections, the previous limited studies on safety fatalities focused on (1) identifying the individual factors that lead to fatalities in the construction industry; (2) analyzing the fatal accidents in terms of their time of occurrence and their presence among different construction trades; (3) identifying the construction works that are the most susceptible to fatalities; and (4) studying the impacts of climate,

training, education, experience, race, and ethnicity, among others on the occurrence of fatalities. As such, previous safety fatality studies in the construction industry focused on the individual safety fatality factors rather than on analyzing the critical combinations, associations, and interconnectivities between the different fatality causes. As such, there is a critical research need to further enhance the safety fatality performance in specific in the construction industry.

Third, many of the previous studies, that have used clustering methods and data mining techniques, were applied to different nonconstruction applications such as road accidents, fire accidents, maritime accidents, and mine accidents, among others. Thus, there is a lack of studies that have used clustering methods and data mining techniques to study the fatal safety accidents in the construction industry in specific. As far as the clustering methods used by previous studies are concerned, these methods mainly included traditional approaches such as principal component analysis and k-means clustering, among others, which have many limitations in terms of their reliance on the individual properties of the data points rather than on the strength of the interconnectivities or associations between the different data points. Therefore, more advanced clustering methods are needed to study the associations between different fatal safety causes. In relation to that, this section of the dissertation uses spectral clustering because it has proved to address the limitations of traditional clustering methods (Wang et al., 2019b; Fu et al., 2005; El Mouden et al., 2019; Chen and Cai, 2011). Similarly, as far as the data mining approaches used by previous studies are concerned, these methods do not factor the possibility of having combinations or causations between different factors or causes. More specifically, the traditional way of looking at safety incidents focuses on analyzing the

weakest link in the chain of events by identifying the only one main cause for the accident and what went wrong that allowed the incident to occur (Goldberg, 2003). However, the rigid adherence to this way of thinking can lead to some significant errors in improving safety performance (Goldberg, 2003). That said, other data mining techniques shall be used with the capabilities of studying the accident causations through examining the interconnectivities and associations between different fatality causes. In relation to that, the frequent pattern mining algorithm allows valuable combinations, associations, and relationships to be identified and for complex, interesting, and hidden associations to be discovered that otherwise could not be found (Verma et al., 2014; Tasneem et al., 2019). In relation to that, this section of the dissertation uses the frequent pattern mining method since it has emerged as a significant approach to discover fascinating knowledge concealed in the data (Ragaventhiran and Kavithadevi, 2019), and it overcomes the limits of other techniques as it does not restrict the underlying relationships between the variables (Xu et al., 2018).

As such, there is a knowledge gap in the literature in terms of focusing on the individual fatality factors rather than possible combinations and associations between them. That said, to address this critical knowledge gap, there is a pressing research need to implement more advanced data-driven management methods for the analysis of safety accidents (Yang et al., 2019) with a special focus on fatalities. In relation to that, this section of the dissertation uses spectral clustering and frequent pattern mining computational algorithms to study the critical associations and combinations between different fatality causes by relying on a data-driven approach, accident causation principles, and concepts pertaining to graph theory.

5.3.5. Graph Theory. One of the most common and effective methods to study the combinations, associations, and interconnectivities between different causes or factors is graph theory; which is also referred to as network representation (Qiao et al., 2018). Graph theory dates back to the 18th century, and it is a branch of mathematics that deals with the way items, factors, or causes are connected (Wilson, 1972). In simple terms, graph theory is a quantitative mathematical approach to examine network properties, where networks are represented as graphs comprised of individual elements (called nodes) and the relationships between them (shown as edges between the nodes) (Gallen and D'Esposito, 2019). That said, a graph G formed by a set of vertices V and a set of edges E can be represented as G = (V, E) in an abstract mathematical structure (Devi and Murugaboopathi, 2019). Graph theory is very popular and widely used in different domain applications due to its exceptional capability in revealing the important combinations, associations, and interconnectivities between different factors or causes (Abotaleb and El-adaway, 2018). This notable ability is attributed to the fact that graph theory combines remarkable techniques including matrix theory, group theory, combinatorics, and numerical analysis (Trinajstic, 2018). That said, graph theory is considered as a systematic approach for conversion of qualitative factors to quantitative values, and its mathematical modeling abilities give an edge over conventional methods making its applications to be very well renowned (Singh et al., 2019). In relation to that, graph theory was employed in diverse research areas such as computer science, economics, biomathematics, engineering, nuclear physics, biology, psychology, theoretical physics, linguistics, and sociology, among others.

Every graph could be mapped into a matrix representation as to make use of the matrices' computational capabilities in handing complex mathematical operations (Singh et al., 2019). Figure 5.1 shows a demonstrative example of the equivalence between the graph and matrix representations and the exploitation of the computational and mathematical capabilities of the latter.

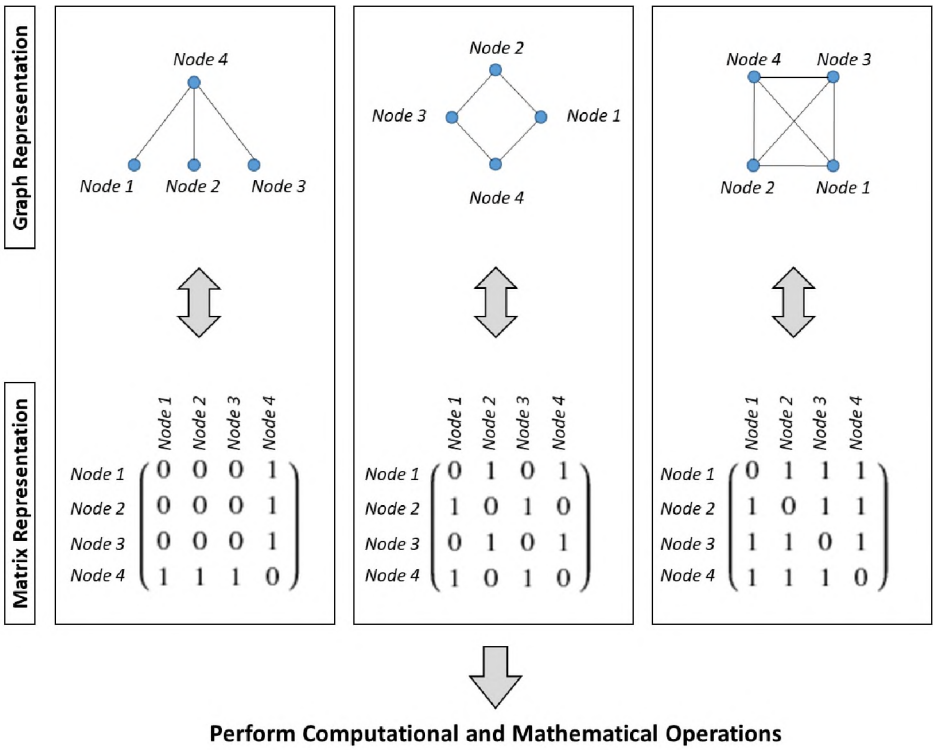

Figure 5.1 Example showing the equivalency between graph and matrix representations.

For the matrix representation in Figure 5.1, a value of 1 is entered if there is an edge (or connection) between two nodes and a value of 0 is entered in the absence of such connection. Graph theory also has the ability to model the combinations, associations, and

interconnectivities between the factors or causes on two levels: global (meaning on the entire network level) and local (meaning on the sub-networks' levels) (Akbarian and Erfanian, 2019). Hence, graph theory was utilized to determine the critical combinations and associations between the different causes of construction fatalities.

5.4. METHODOLOGY

A data-driven methodology comprised of three main steps was followed as summarized in Figure 5.2. All needed details pertaining to each one of the three steps are presented in the following subsections.

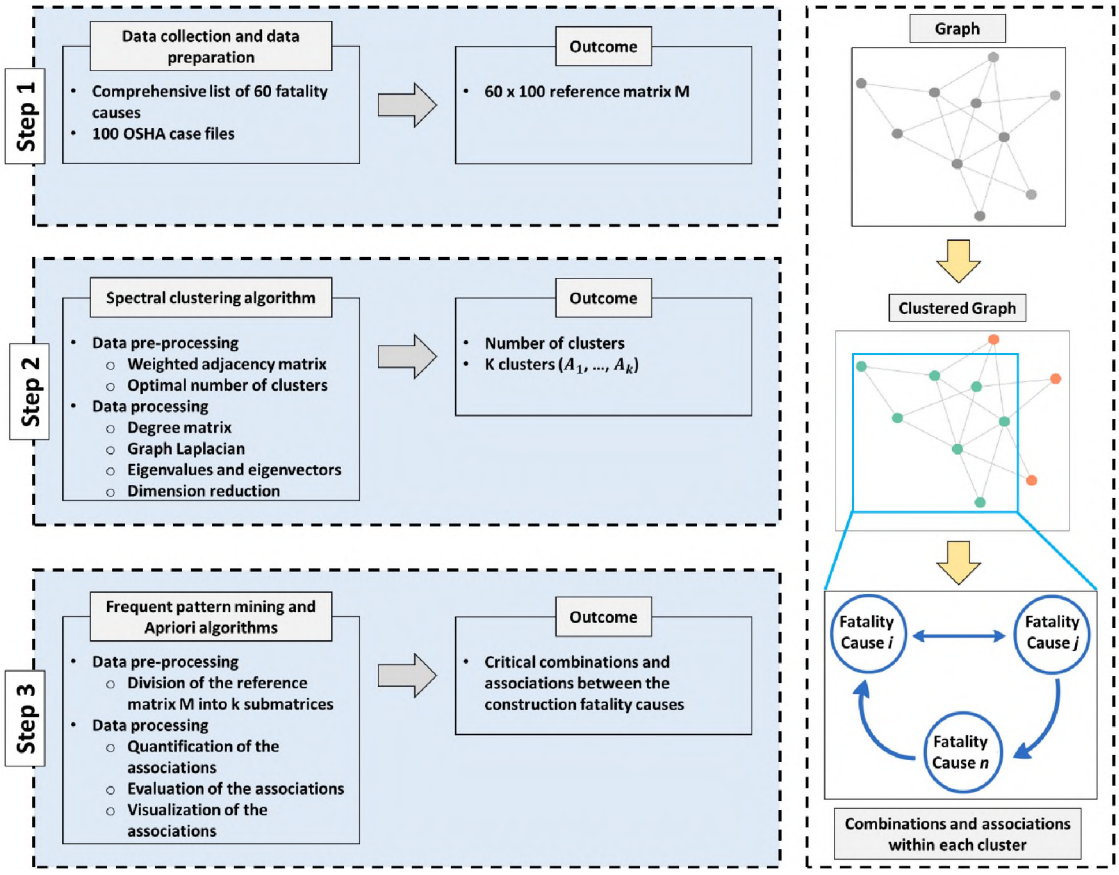

Figure 5.2 Research methodology.

5.4.1. Step 1: Data Collection and Data Preparation. This subsection provides all details related to the data collection and data preparation efforts.

5.4.1.1. Fatality causes and case files. The first step to determine the critical combinations between the causes leading to construction fatalities is to determine such causes. That said, since new research work should build on previous research efforts to convey prospective findings that add to the body of knowledge, this section of the dissertation used the data present in Eteifa and El-adaway (2017) where a list of 66 causes was identified. It is worth mentioning that some changes were performed to Eteifa and El-adaway's (2017) list of 66 causes in this section of the dissertation by removing the redundant and uncontrollable causes. In relation to that, a final list of 60 causes was obtained as shown in Table 5.1. The distribution of the key direct causes (rather than root causes) for the investigated fatalities is shown in Figure 5.3.

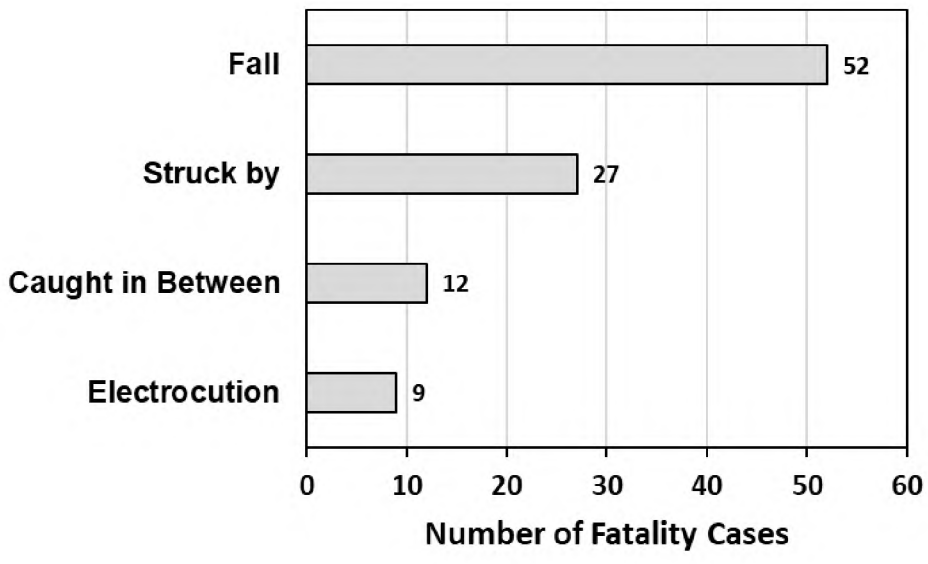

Figure 5.3 Direct causes of fatalities.

Table 5.1 Causes of fatality accidents; adapted from Eteifa and El-adaway (2017).

Cause number	Construction fatality cause
C1	Damaged, defective, or malfunctioning equipment
C2	Lack of necessary Equipment
C3	Poor Equipment handling and not following proper operation procedures or manufacturer's specs
C4	Use of non-suitable equipment
C5	Equipment/spoil on excavation edge
C6	Vibration in excavation
C7	Equipment tipping
C8	Poor Labeling
C9	Improper use of PPE
C10	No PPE
C11	No fall arrest system, guardrails or safety nets.
C12	No cable insulation
C13	Not wearing seatbelt
C14	Safety element failure
C15	No/damaged Cave in protection
C16	No protection from traffic
C17	Inappropriate decking
C18	Lack of employee knowledge
C19	Lack of coordination of site activities
C20	Poor tool handling
C21	Inappropriate tools used
C22	Poor Material Handling
C23	Poor storage
C24	High exposure to chemical
C25	Wet Material
C26	Employer gross negligence
C27	Employer allowed employee to work in an unsafe environment
C28	Lack of knowledge by employer about site conditions
C29	Lack of clear employer instructions
C30	Failure to properly locate Utilities
C31	Lack of preventive action
C32	No first aid personnel
C33	Lack of supervision and/or absence of competent or necessary personnel
C34	Vehicle observer error

Table 5.1 Causes of fatality accidents; adapted from Eteifa and El-adaway (2017). (Continued).

C35	Operation carried out by noncompetent individual
C36	Employee misconduct
C37	Willfully exposing self to hazardous situation
C38	Misjudgment of hazardous situation
C39	Lack of specific on the job training
C40	Lack of general health and safety training
C41	Jobsite inspection related
C42	Lack of Inspection for equipment and tools
C43	Poor Assembling of equipment/scaffold/decking/formwork
C44	Error in design
C45	Failure of structural element
C46	Collapse of structure
C47	Over excavating
C48	No safe access to site/scaffold/trench
C49	Poor housekeeping
C50	No safe exit to site
C51	Working surface condition not suited to task
C52	Site obstruction
C53	No safe walkways
C54	No site survey
C55	Inappropriate lighting
C56	Not following proper work procedures
C57	No hazard identification/ communication program
C58	No testing procedure for equipment
C59	No effective emergency plan
C60	Lack of safe working procedures

The study conducted by Eteifa and El-adaway (2017) was selected due to its feasibility to the scope and purpose of this section of the dissertation. More specifically, this existing study has investigated the fatalities in the construction industry rather than safety accidents in general (such as occupational injuries). Also, this existing study has

focused on the US construction industry in specific and has provided a comprehensive list that includes the most pertinent causes contributing to construction fatalities in the US. In addition, the study by Eteifa and El-adaway (2017) used social network analysis to study the relationships between the fatal accident root causes and the commonly quoted direct causes and found that lack of job-specific training is the most central cause of struck by and caught in between accidents and that the absence of fall arrest systems, lack of jobsite training, and lack of personal protective equipment are the most interrelated root causes.

The data analyzed in relation to the 60 construction fatality causes was sampled using the stratified random sampling technique from the most recent complete available completion OSHA national fatal accident case files from the Construction Industry Research and Policy Center (CIRPC). Those files are prepared by OSHA compliance officers after the occurrence of any fatal accident; a total of 100 fatality case files were secured. Each one of these case files included an inspection summary, site walk around, OSHA citations, inspection narratives, safety narratives, checklists, and interviews for personnel relevant to the fatal accident. Based on the data present in the case files, the data was prepared in a 60×100 matrix format where each row corresponds to each one of the 60 fatality causes, and each one of the columns corresponds to each one of the 100 case files. If a fatality cause was mentioned in a case file, a value of 1 is entered into its corresponding cell; otherwise, a value of 0 is entered. The process of the transformation of the 100 case files to a digital matrix format was performed by following multiple steps. First, the 100 case files were collected and downloaded from the CIRPC database. Second, each one of these case files was inspected by examining the fatality-related information reported in each file. In relation to that, the collected case files included the following information: (1)

quoted direct cause of the fatal accident, (2) the root cause of the fatal accident, (3) number and codes of quoted OSHA citations linked to the fatal accident, (4) the state in which the fatality occurred, (5) the function of the worksite, (6) whether the victim was conducting their regular assigned task, and (7) the role of the victim in the facility. Third, this information was reviewed to ensure that the collected data is organized and consistent among all 100 case files and that there is no missing information. Fourth, to be able to perform the needed quantitative analysis for this information, the collected data was transformed into a digital format through a simple and straight forward manual mapping between each one of the causes that are present in Table 5.1 and the associated case file. Since the data was prepared by OSHA compliance officers, the fatality causes were very clearly specified in each one of the case files and thus there was no subjective interpretation of these files. In relation to that, no subjectivity was introduced in the manual transformation of the information into a digital matrix format. Also, since the same fatality cause cannot be repeated in the same case file more than once, a binary matrix format was enough. That said, if one of the 60 fatality causes in Table 5.1 was mentioned in any one of the 100 case files, a value of 1 was manually entered into its corresponding cell; otherwise, a value of 0 was entered. Ultimately, this has resulted in a binary 60×100 matrix, which is referred to as the reference matrix (M) in this section of the dissertation. It is worth mentioning that similar manual transformation was performed by many previous research studies such as the works conducted by Assaad and El-Adaway (2020d) and Abdul Nabi and El-adaway (2020), just to name few.

5.4.1.2. Sufficiency of the sample size. It is worth mentioning that the obtained 100 cases are not 100 combinations of safety fatality factors. In fact, multiple fatality

causes have been reported in each one of these 100 case files. More specifically, 462 occurrences were reported in the entire 100 case files. This total number of 462 occurrences is acceptable compared to previous similar research works that used data mining and association analysis for 309 occurrences in construction projects (Liao and Perng, 2008).

On the other hand, to test the sufficiency of the 100 cases considered in this section of the dissertation, the sample sizes of past safety-related studies have been reviewed. In relation to that, Liao et al. (2015) studied the influence of person-organizational fit on construction safety climate based on data from 80 actors. Also, Wang et al. (2017b) studied the human safety risks and their interactions from a broad project-stakeholder perspective based on data provided from 75 stakeholders. Lingard et al. (2019) examined the relationship between intragroup communication related to work health and safety performance and the workgroup safety climate based on data collected from 39 workgroups. In addition, Esmaeili and Hallowell (2012) determined the influential factors by studying the diffusion patterns of safety innovations in the construction industry based on data collected from 58 companies. Pirzadeh and Lingard (2017) studied the dynamics of construction decision making and the impact on work health and safety by examining the interconnectivities in network information for 42 sequential design decisions. Alsamadani et al. (2013) analyzed the interconnectivities between safety performance, language proficiency, and communication patterns based on 14 small construction crews in the Denver Metropolitan region in the United States. Moreover, Allison and Kaminsky (2017) studied the connections between communication patterns and crew safety performance based on data collected from 8 construction crews. Montella et al. (2011)

identified crash contributory factors and studied the interdependences and associations between them based on 15 urban projects.

It is concluded that the sample size of 100 case files is sufficient based on a comparison with the sample sizes that were used by previous relevant safety research works.

5.4.2. Step 2: Spectral Clustering Algorithm. This subsection provides all details related to the implemented spectral clustering algorithm.

5.4.2.1. Overview. Spectral clustering (SC) technique was used to cluster the 60 fatality causes based on the strength of the interconnectivities between them. SC is a clustering technique based on algebraic and spectral graph theory (Jia et al., 2014). SC is considered a powerful clustering method using partitioning of a graph (Shinnou and Sasaki, 2008), or equivalently, clustering of data based on its matrix representation and its spectral analysis (El Mouden et al., 2019). In other words, SC divides the graph nodes into groups so that connectivity is maximized between nodes in the same cluster and the connectivity is minimized between nodes in different clusters (Brunskill et al., 2007). This would result in an optimal clustering of the data as SC converges on the global optimal solution (Wang et al., 2020b); thus, it is considered an innovative clustering technique for graph matrix partitioning (Janani and Vijayarani, 2019). SC has aroused extensive attention in recent research studies due to its solid theoretical foundation as well as its good practical clustering performance (Jia et al., 2014). That said, SC is considered one of the most popular and most commonly used modern clustering approaches (Liu et al., 2019b). The unique capabilities of SC made it receive much attention as a competitive clustering

algorithm emerging in recent years (Wang et al., 2019b). Based on the aforementioned information, SC was employed in this section of the dissertation.

5.4.2.2. Data preprocessing. The two needed inputs for the SC algorithm are the weighted adjacency matrix and the number of clusters (Von Luxburg, 2007). As such, this subsection details how these two inputs are determined during data preprocessing.

Weighted adjacency matrix: before processing the data, the data needs to be preprocessed as to construct a weighted n by n adjacency matrix (W) (Xu et al., 2019). W is calculated by multiplying the reference matrix M by its transpose and replacing the diagonal values by zeros (Abotaleb and El-adaway, 2018) as provided by Equation (20).

$$W_{nxn}(i,j) = \begin{cases} M_{nxm} \cdot M_{mxn}^{T} & for\ i \neq j \\ 0 & for\ i = j \end{cases} \tag{20}$$

where W is the adjacency matrix with elements w_{ij} such that $W = (w_{ij}) \in \mathbb{R}^{nxn}$, M is the reference matrix, M^T is the transpose of the reference matrix, n is the number of fatality causes (i.e. 60), m is the number of case files (i.e. 100). Figure 5.4 shows a demonstrative example of how the weighted adjacency matrix is calculated.

Once the weighted adjacency matrix W is obtained, the fatality-causation interdependencies could be represented as a weighted undirected graph $G = (V, E)$ with the vertex set $V=\{v_1,v_2,\ldots,v_n\}$ such that vertex v_i represents the fatality cause C_i, and $E \subset V \times V$ is the edge set where a nonnegative weight $w_{ij}=w_{ji} \geq 0$ [calculated using Equation (20)] is associated for each edge $(i, j) \in E$ between two vertices v_i and v_j.

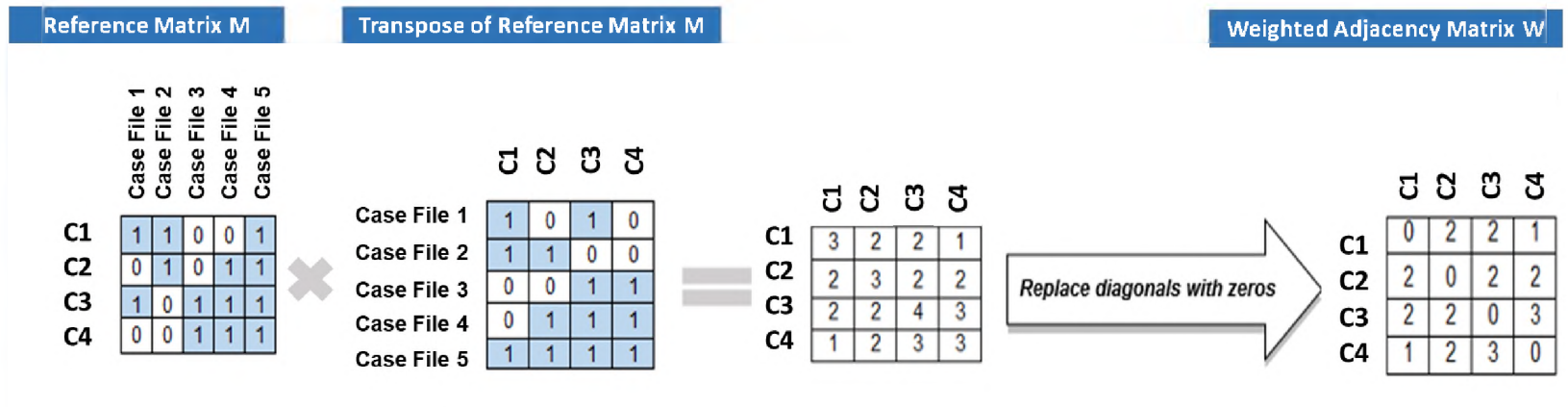

Figure 5.4 Example showing the calculations of the weighted adjacency matrix.

Determination of the optimal number of clusters: after obtaining the matrix W, the optimal number of clusters k shall be determined. In relation to that, the silhouette score/width method (Rousseeuw, 1987) was used since it is the most popular technique for determining the optimal number of clusters (Chang and Chi, 2019). The optimal value k of clusters is identified based on the highest silhouette score (Oskouie et al., 2017) with $k>2$. The silhouette score is calculated using Equation (21).

$$S = \frac{b-a}{\max(a,b)} \tag{21}$$

where S is the silhouette score, a is the mean intracluster distance, and b is the distance between a sample and the nearest cluster that the sample is not a part of.

5.4.2.3. Data processing. Once the matrix W and the optimal number of clusters (k) are obtained, the SC algorithm could be applied according to the steps presented in the next subsections and that are mentioned in Von Luxburg (2007) and Ng et al. (2002).

Determination of the optimal number of clusters: The degree of a vertex $v_i \in V$ is obtained using Equation (22).

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{22}$$

In relation to that, the $n \times n$ degree matrix D is defined as the diagonal matrix with the degrees $d_1, d_2, \ldots, d_n$ on the diagonal, and it is obtained using Equation (23). In other words, the (i, i)-elements of the matrix D are calculated as the sum of W's ith row.

$$D = diag(d_i) \tag{23}$$

where $d_i = (\mathbf{W1})_i$ and $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ is the indicator vector.

Decomposition of the graph Laplacian matrix: the main tools for SC are graph Laplacian matrices. First, the $n \times n$ unnormalized graph Laplacian L matrix is obtained using Equation (24).

$$L = D - W, \qquad i.e., L(W) = diag(W\mathbf{1}) - W \tag{24}$$

where D is the degree matrix, W is the weighted adjacency matrix, and $\mathbf{1}$ the indicator vector. Based on the obtained L matrix, the $n \times n$ normalized symmetrical graph Laplacian L_{sym} is obtained using Equation (25).

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \tag{25}$$

where D is the degree matrix, L is the unnormalized graph Laplacian matrix, I is the identity matrix, and W is the weighted adjacency matrix.

Calculation of the eigenvalues and eigenvectors: once L_{sym} is obtained, its eigenvalues (λ) and eigenvectors (e) are calculated based on Equation (26).

$$L_{sym}.e = \lambda.\mathrm{e} \qquad \text{with } \left|L_{sym} - \lambda \mathrm{I}\right| = 0 \tag{26}$$

where λ is a scalar, e is a vector, and I is the identity matrix.

It is to be noted that for each vertex v_i, there exists one eigenvalue and one eigenvector (i.e., n eigenvalues in total and n eigenvectors in total for all v_i).

Dimension reduction and determination of clusters: the dimensionality of the data is reduced by constructing the matrix $U \in \mathbb{R}^{nxk}$ through stacking the largest k eigenvectors in columns. Next, the matrix $T \in \mathbb{R}^{nxk}$ is obtained by normalizing each of U's rows to have unit length (a norm of 1) using Equation (27).

$$t_{ij} = \frac{u_{ij}}{(\sum_k u_{ik}^2)^{1/2}} \tag{27}$$

Let $y_i \in \mathbb{R}^k$ be the vector corresponding to the ith row of matrix T for i=1,…,n. By treating each y_i as a point in $\mathbb{R}^k$, these points are clustered into k clusters (A_1, …, A_k) via k-means. The end output is k clusters that group the 60 fatality causes based on the strength of their interconnectivities.

5.4.2.4. Coding and software libraries. It is worth mentioning that the data preprocessing and processing steps for the SC algorithm were performed using Python. Table 5.2 shows the utilized libraries along with their associated use for this research.

Table 5.2 Used python libraries for the spectral clustering algorithm.

Library	Use
Pandas	Data manipulation and analysis for data structures and numerical tables
Numpy	Add support for large and multi-dimensional arrays and matrices as well as provide high-level mathematical functions to operate on the arrays and matrices
Matplotlib	Data plotting (two dimensional)
Seaborn	Data visualization (matrix format)
Sklearn	Machine learning and clustering

5.4.3. Step 3: Frequent Pattern Mining and Apriori Algorithm. After classifying each of the 60 fatality causes into their respective clusters, a determination and quantification of critical combinations and associations between the different fatality causes within each cluster was performed using the frequent pattern mining (FPM) method and the Apriori algorithm as detailed in the next subsections.

5.4.3.1. Overview. FPM and Apriori algorithms are considered one of the key approaches to study the combinations and associations between different factors or causes (Joshi et al., 2018; Susymary and Lawrance, 2017). FPM is also referred to as association rule mining/analysis or market basket analysis (Patel, 2018). FPM is considered a very important field for mining of the association rule to extract meaningful information from large data sets (Singh, 2019b). In fact, FPM is an eminent association method to quantify the combinations between variables, factors, or causes in a data set (Hosseini et al., 2018). The primary purpose of FPM is to discover obscure patterns/combinations hidden in a mass of data (Cheng et al., 2010). That said, FPM is considered an important knowledge discovery technique in data mining (Htet, 2019).

On the other hand, the first algorithm for frequent pattern mining is Apriori (Fournier-Viger, et al., 2019). The Apriori algorithm is one of the most popular FPM algorithm for numerous reasons summarized by Rahman et al. (2019), including (1) it is effective to find frequent combinations and patterns; (2) it is very easy to understand and implement; and (3) it uses comparatively less memory. Based on all aforementioned information, FPM and Apriori algorithm were used.

5.4.3.2. Data preprocessing. After clustering the 60 causes into k clusters using the SC algorithm, the data on the 60 fatality causes was preprocessed to be analyzed using

FPM and Apriori algorithm. That said, the reference matrix M was divided into k submatrices W_1, W_2, …, W_k where each submatrix W_i includes the set of fatality causes pertaining to Cluster A_i (obtained from the SC algorithm in Step 2). That said, FPM and Apriori algorithm are applied for each submatrix W_i. Figure 5.5 shows a demonstrative example on the data preprocessing before the application of the FPM and the Apriori algorithm.

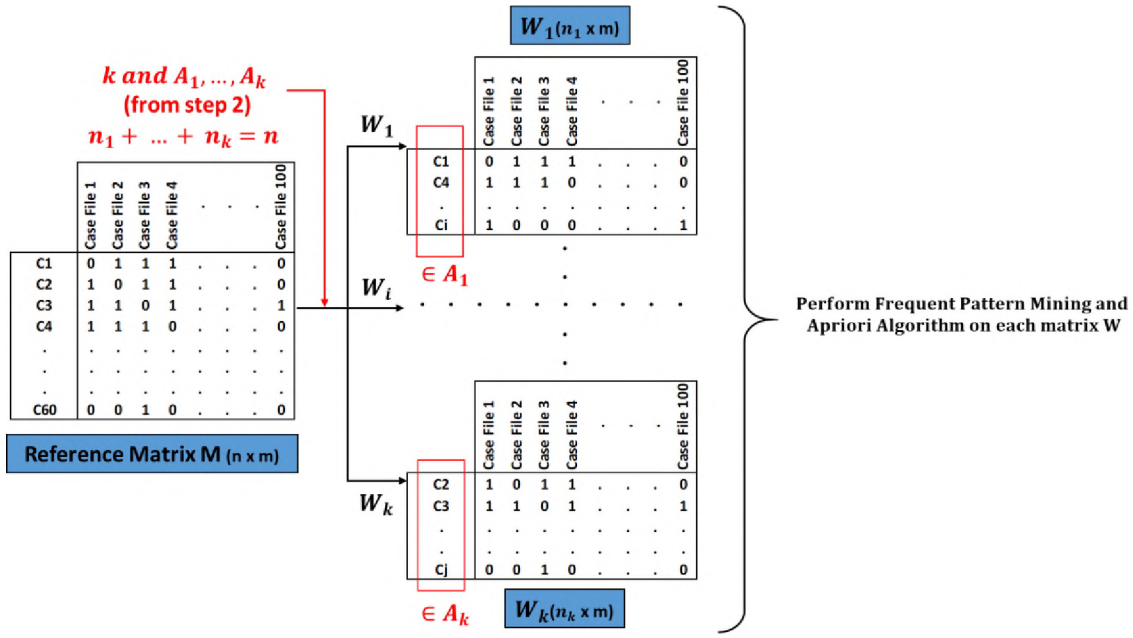

Figure 5.5 Demonstrative example for data pre-processing for the frequent pattern mining and Apriori algorithms.

5.4.3.3. Data processing. This subsection provides all details related to the data processing efforts for the FPM and Apriori algorithm.

Quantification of the associations: after preprocessing of the data, the Apriori algorithm was used for performing the FPM. Let $C = \{C_i, \ldots, C_j\}$ be the set of fatality

causes (called itemset; if an itemset has q items it is called a q-itemset) represented by Cluster A_j. Let $C_F = \{f_1, f_2, \ldots, f_n\}$ be the associated set of case files that represent the set C, where each f_i is a set of fatality causes present in case file i such that $f_i \subseteq C$. For instance, if a case file includes the fatality causes C_1, C_4, and C_{10}, then the corresponding case file can be expressed as the set of fatality causes $\{C_1, C_4, C_{10}\}$. The combinations or associations between the different fatality causes could then be expressed as $C_i \sim C_j$, where $C_i, C_j \subseteq C$ and $C_i \cap C_j = \emptyset$. The frequent fatality causes are calculated using the measure support for each fatality cause C_i as provided in Equation (28).

$$Support\ (C_i) = \frac{Number\ of\ case\ files\ that\ contain\ the\ fatality\ cause\ C_i}{Number\ of\ all\ case\ files\ in\ the\ cluster} \tag{28}$$

On the other hand, three measures are used to evaluate/quantity the associations between the fatality causes: (1) support, (2) confidence, and (3) lift (Shi et al., 2019). The support measure reflects how frequently the combination is applicable in the data set, the confidence measure represents the predictability of the combination, and the lift measure determines the associations between the simultaneous occurrences between the causes (Liu et al., 2018; Mateos, et al., 2019). These three measures are calculated using Equation (29) to (31), respectively. It is worth mentioning that the difference between Equation (28) and Equation (29) is that the former calculates the support for each individual fatality cause, whereas the latter calculates the support for a pair or combination of fatality causes.

$$Support\ (Ci \sim Cj) = \frac{Number\ of\ case\ files\ that\ contain\ Ci\ and\ Cj}{Number\ of\ all\ case\ files\ in\ the\ cluster} \tag{29}$$

$$Confidence\ (Ci \sim Cj) = \frac{Number\ of\ case\ files\ that\ contain\ Ci\ and\ Cj}{Number\ of\ case\ files\ that\ contain\ Ci} \quad (30)$$

$$Lift\ (Ci \sim Cj) = \frac{Support\ (Ci \sim Cj)}{(Number\ of\ case\ files\ that\ contain\ Ci) \times (Number\ of\ case\ files\ that\ contain\ Cj)} \quad (31)$$

Evaluation of the associations: the support and confidence measures are used to control the number of combinations and associations to provide meaningful results (Kadimisetty, 2018). In relation to that, threshold values for the support and confidence measures are required (Liao and Perng, 2008). In other words, the identified combinations and associations between the construction fatality causes shall satisfy the conditions present in Equation (32) and Equation (33) (Hahsler and Chelluboina, 2011).

$$Support\ (Ci \sim Cj) \geq \sigma \quad (32)$$

$$Confidence\ (Ci \sim Cj) \geq \delta \quad (33)$$

where σ and δ are minimum thresholds for support and confidence measures, respectively.

To this end, a support threshold $\sigma = 0.01$ and a confidence threshold $\delta = 0.75$ were used since they are the recommended values to provide useful and noteworthy combinations and associations (Verma et al., 2014; Hosseini et al., 2018). Figure 5.6 shows the simplified steps performed for the identification and evaluation of the combinations

and associations between the construction fatality causes based on the threshold values for the support and confidence measures. It is worth mentioning that the steps present in Figure 5.6 are performed in an iterative manner for each one of the *k* clusters.

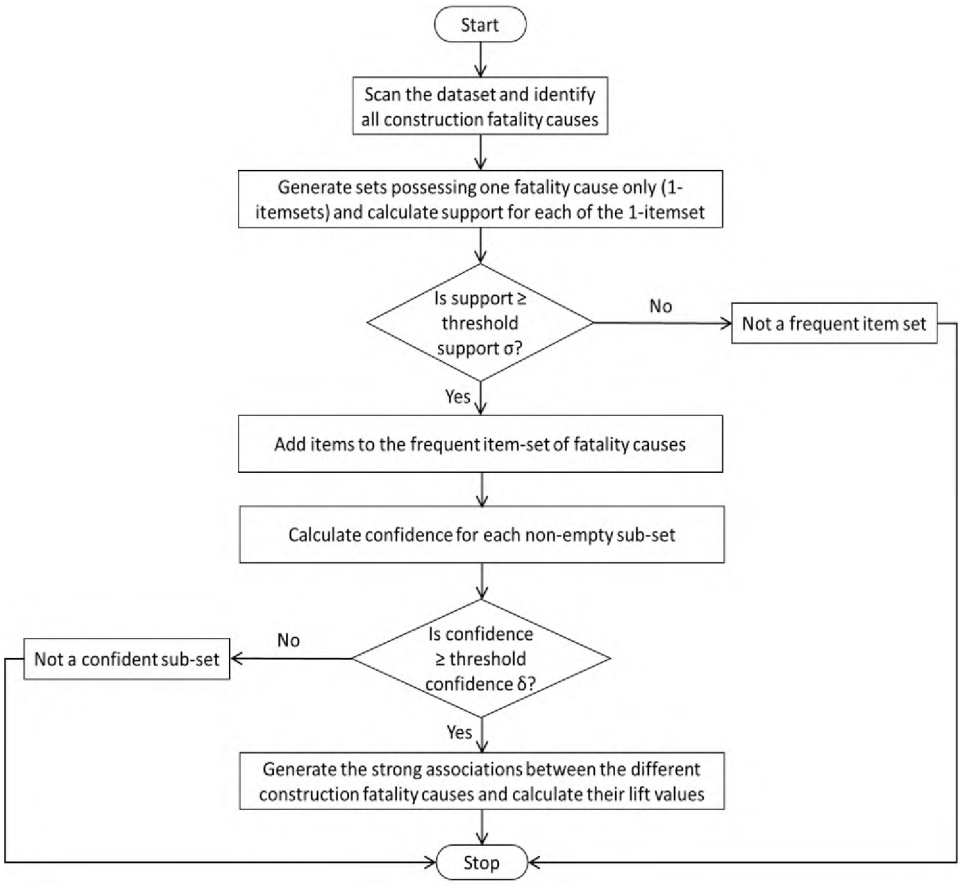

Figure 5.6 Iterative steps to identify and evaluate the associations between fatality causes.

Visualization of the associations: all obtained associations were visualized using two-dimensional (2D) scatter plot. In addition, the parallel coordinates plot was used to visualize the multidimensionality of the obtained combinations between the fatality causes.

5.4.3.4. Coding and software libraries. It is worth mentioning that all the coding and associated analyses required for the FPM and Apriori algorithm were performed in R using RStudio. RStudio is an integrated development environment (IDE) for R which is a programming language for statistical computing and graphics. Table 5.3 shows the utilized software packages along with their associated use for this research.

Table 5.3 Used R packages for the frequent pattern mining and Apriori algorithm.

Package	Use
Arules	Provide the infrastructure for representing, manipulating and analyzing the data and patterns. Also provides C implementations of the association mining algorithm Apriori
ArulesViz	Extend the package 'arules' with numerous visualization techniques for combinations, associations, and item-sets. It also includes several interactive visualizations.
Shiny	Build interactive features

5.5. RESULTS AND ANALYSIS

This subsection presents the obtained results for the SC, FPM, and Apriori algorithms.

5.5.1. Spectral Clustering Algorithm. After obtaining the weighted adjacency matrix W, the optimal number of clusters was determined based on the silhouette score. That said, the silhouette score for different values of $k > 2$ (number of clusters) was the highest when k is 5. Thus, the optimal number of clusters to be used for the SC algorithm is 5. Consequently, the SC algorithm was implemented where the obtained results are shown in Figure 5.7. Accordingly, the identification and quantification of the combinations or associations between the different construction fatality causes within each one of the identified five clusters was performed using the FPM and Apriori algorithms.

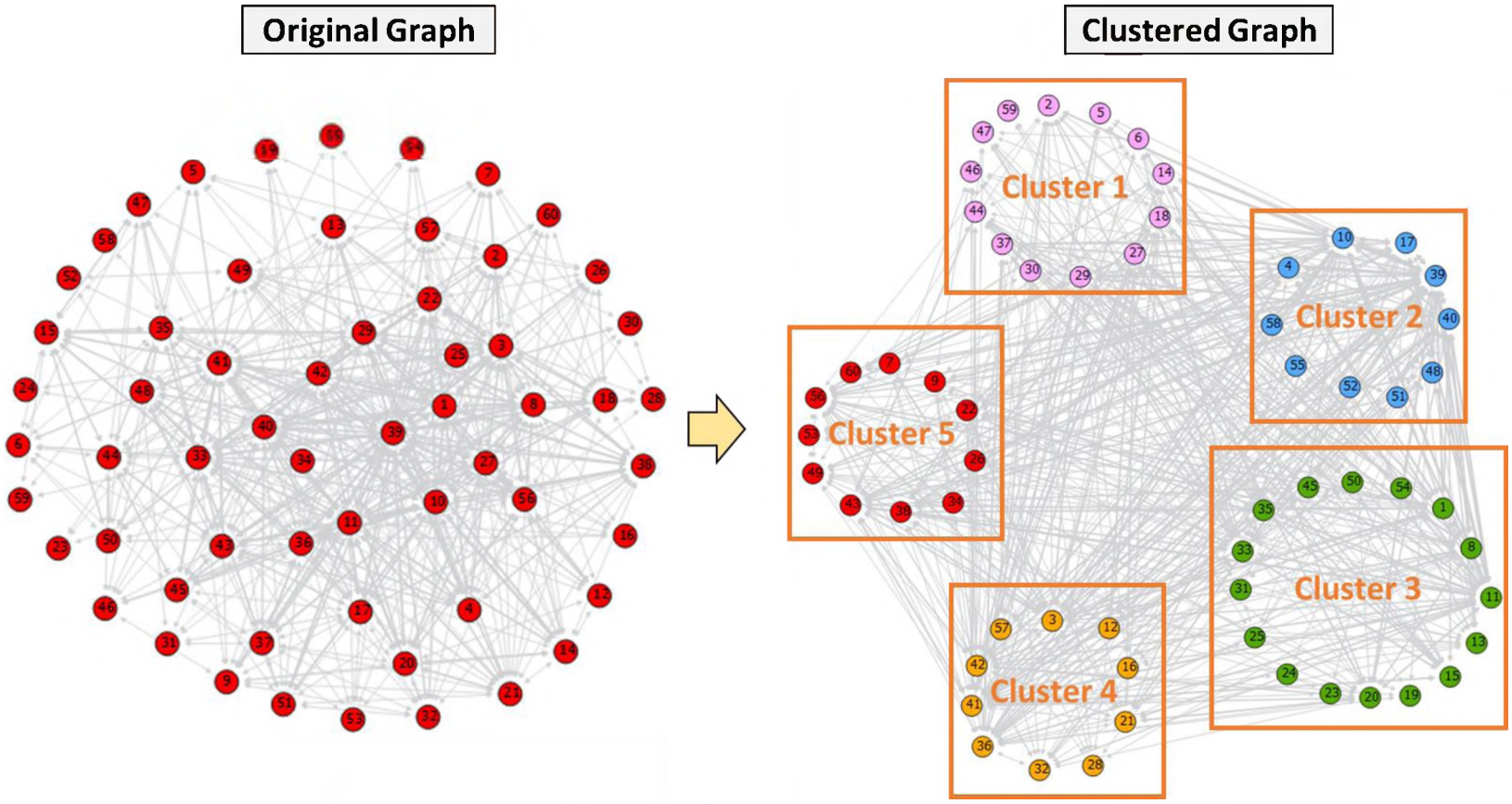

Figure 5.7 Obtained results for the spectral clustering algorithm.

5.5.2. Frequent Pattern Mining and Apriori Algorithm. This subsection presents the obtained combinations or associations and discusses the most critical ones within each cluster.

5.5.2.1. Frequent items and associations within cluster 1. The obtained support measures for the individual fatality causes within Cluster 1 are shown in Figure 5.8.

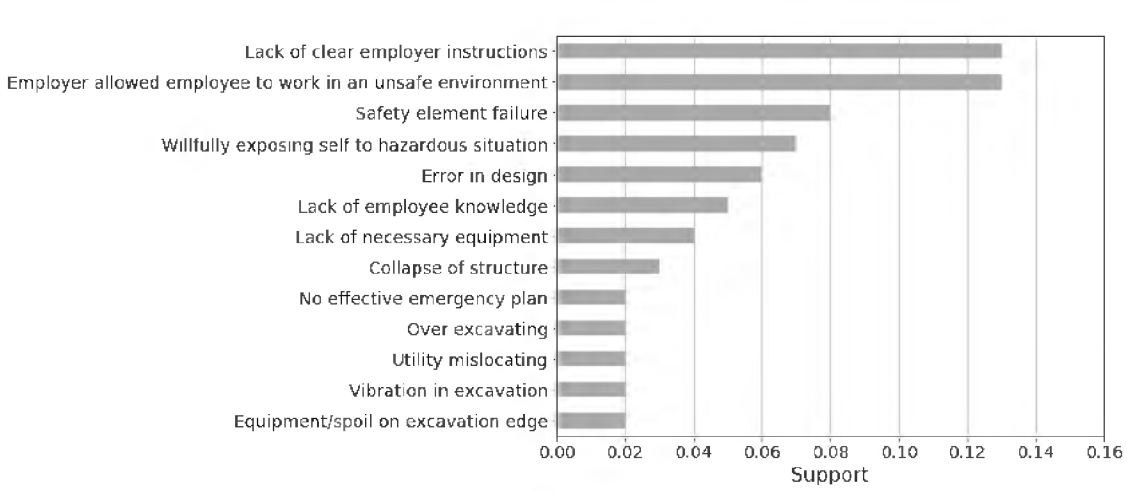

Figure 5.8 Support measure for the individual fatality causes in cluster 1.

As it could be seen from Figure 5.8, the most frequent three fatality causes contributing to construction fatality accidents within Cluster 1 are (1) lack of clear employer instructions, (2) employer allowed employee to work in an unsafe environment, and (3) safety element failure. Figure 5.9 shows the obtained noteworthy combinations within Cluster 1. A total of six associations were obtained [Figure 5.9(a)] between six fatality causes [Figure 5.9(b)]: (1) overexcavating, (2) vibration in excavation, (3) error in design, (4) employer allowed employee to work in an unsafe environment, (5) lack of employee knowledge, and (6) lack of necessary equipment.

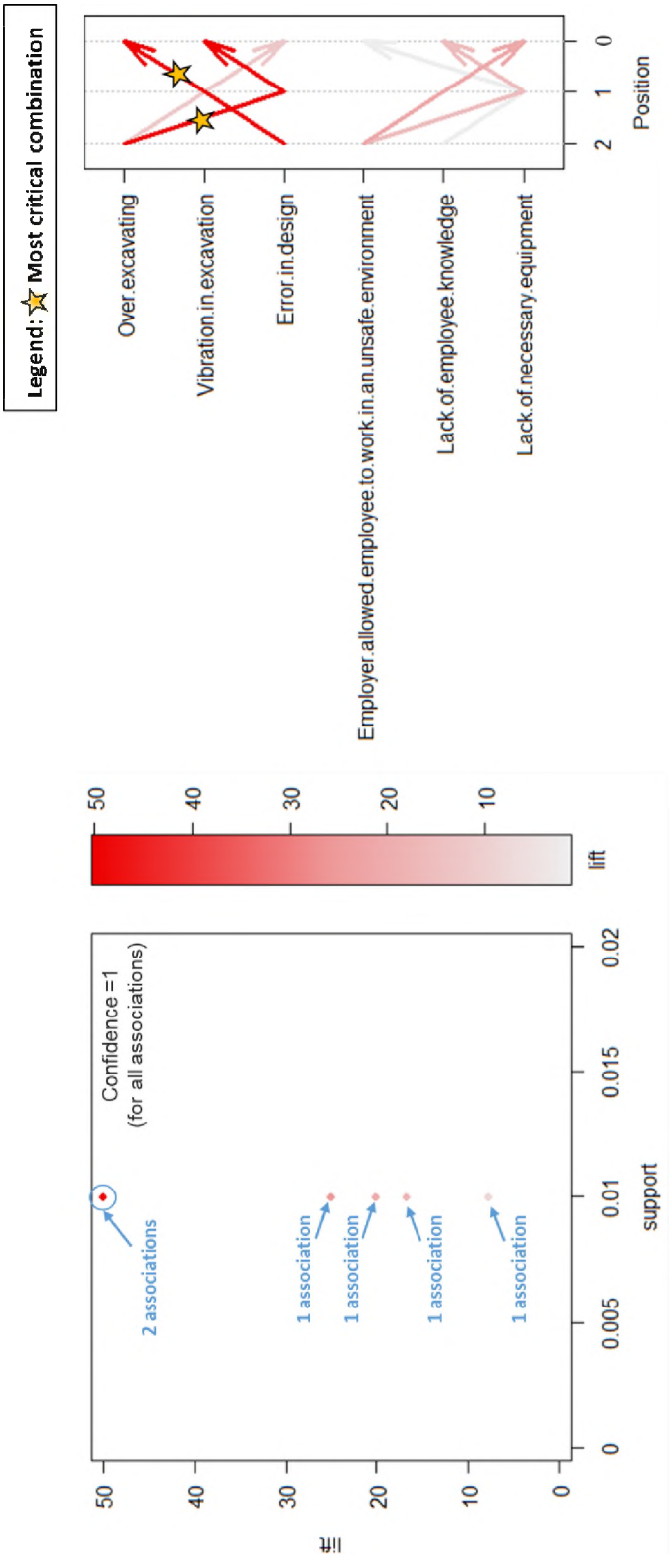

Figure 5.9 Identified remarkable combinations or associations within cluster 1.

Figure 5.9(b) shows that there are two most critical combinations, but these two combinations exist between the following same three fatality causes (but with different ordering): (1) overexcavating, (2) vibration in excavation, and (3) error in design. This association has a support measure of 0.01, a confidence measure of 1, and a lift measure of 50. The interpretation of these measures reflects that there is a high chance of occurrence of this combination of fatality causes (confidence is equal to 1). In addition, a lift value of 50 > 1 reflects that the obtained association is of value (Yao et al., 2019) in the sense that the combinations of these three fatality causes appear more often than expected (IBM, 2019). That said, this identified combination would lead to many fatal accidents on construction sites. In other words, the presence of these fatality causes on construction sites would serve as a warning sign that a fatal accident is very likely to occur.

5.5.2.2. Frequent items and associations within cluster 2. The obtained support measures for the individual fatality causes within Cluster 2 are shown in Figure 5.10.

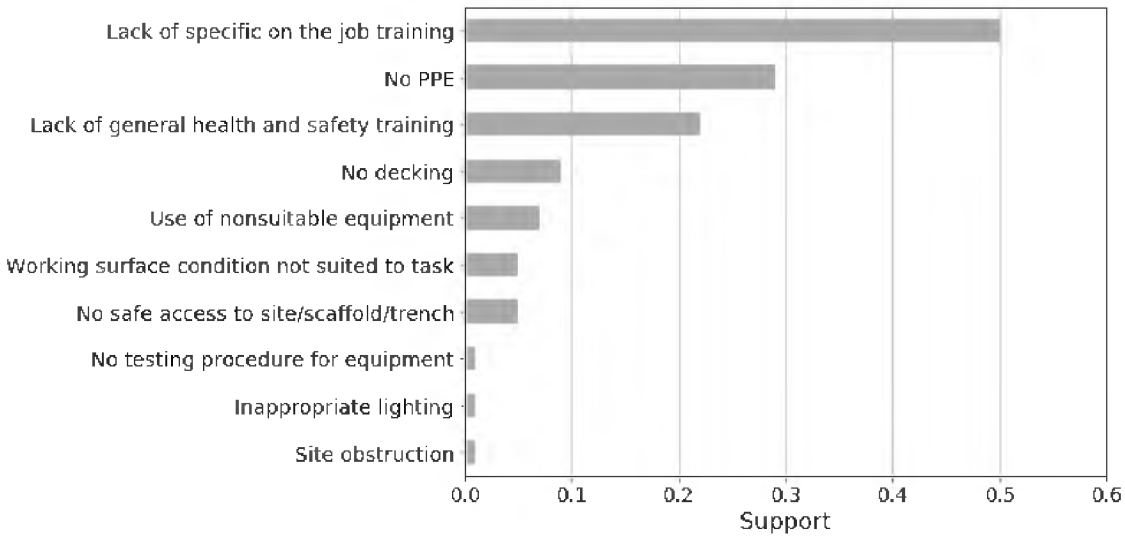

Figure 5.10 Support measure for the individual fatality causes in cluster 2.

As it could be seen from Figure 5.10, the most frequent three fatality causes contributing to construction fatality accidents within Cluster 2 are (1) lack of specific on the job training, (2) no PPE (protective personal equipment), and (3) lack of general health and safety training. This highlights the importance of providing the necessary PPE, the critical role of specific job training on construction sites, and the importance of implementing adequate training on general health and safety considerations.

Figure 5.11 shows the obtained noteworthy combinations or associations within Cluster 2. A total of 15 associations were obtained [Figure 5.11(a)] between 7 fatality causes [Figure 5.11(b)]: (1) use of nonsuitable equipment, (2) no safe access to site/scaffold/trench, (3) lack of general health and safety training, (4) no decking, (5) lack of specific on the job training, (6) no PPE, and (7) working surface condition being not suited for the task. The distribution of these associations within the combinations of the fatality causes is reflected in Figure 5.11(b). One notable and critical combination exists within Cluster 2 (highest lift compared to others), and it is between three fatality causes: (1) lack of general health and safety training, (2) no decking, and (3) no safe access to site/scaffold/trench. This combination has a support of 0.01, a confidence measure of 1, and a lift measure of 20. The interpretation of these measures reflects that there is a high chance of occurrence of this combination of fatality causes (confidence is equal to 1). In addition, a lift value of $20 > 1$ reflects that the obtained association is of value (Yao et al., 2019) in the sense that the combinations of these three fatality causes appear more often than expected (IBM, 2019). That said, this combination reflects that the presence of these three causes of fatality on construction sites would serve as a warning sign that a fatal accident is very likely to occur.

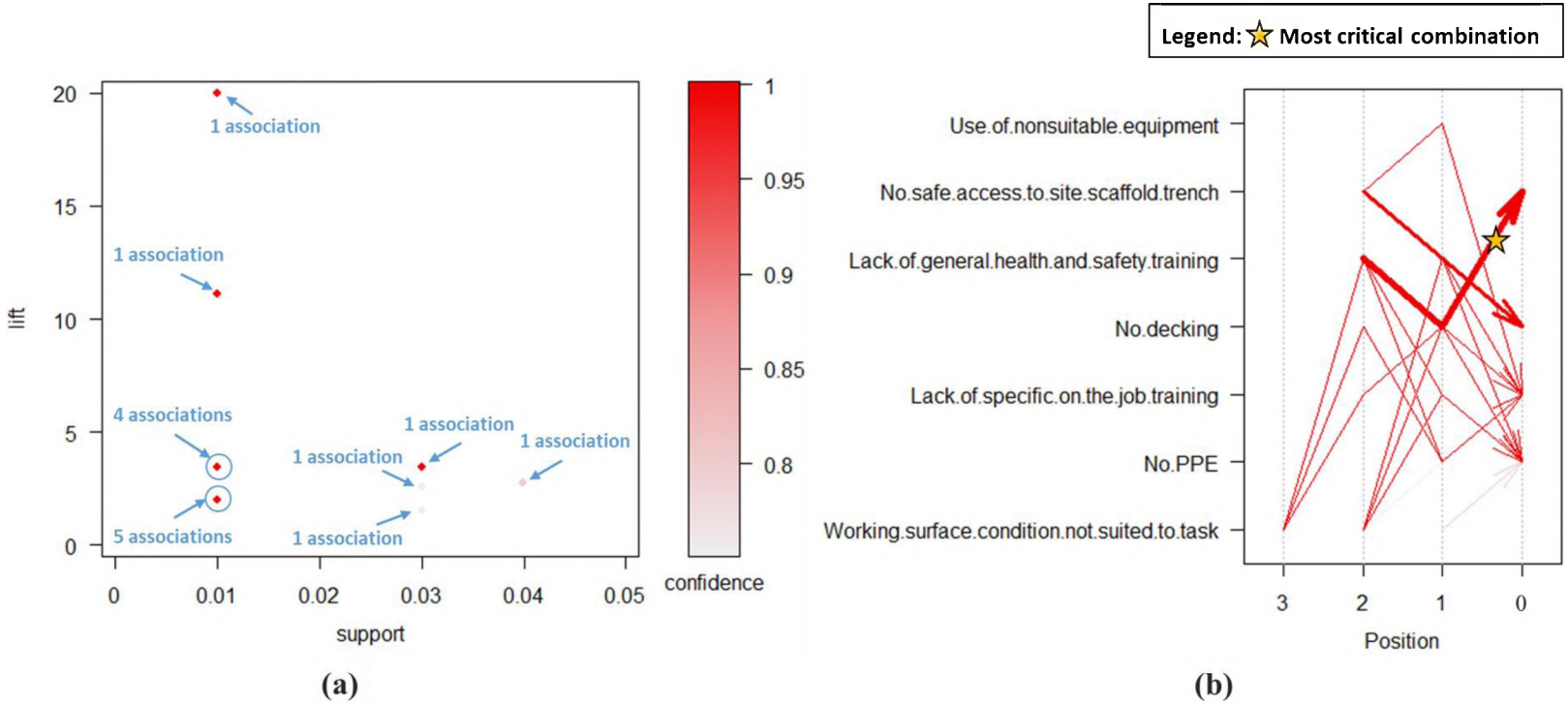

Figure 5.11 Identified remarkable combinations or associations within cluster 2.

5.5.2.3. Frequent items and associations within cluster 3. The obtained support measures for the individual fatality causes within Cluster 3 are shown in Figure 5.12.

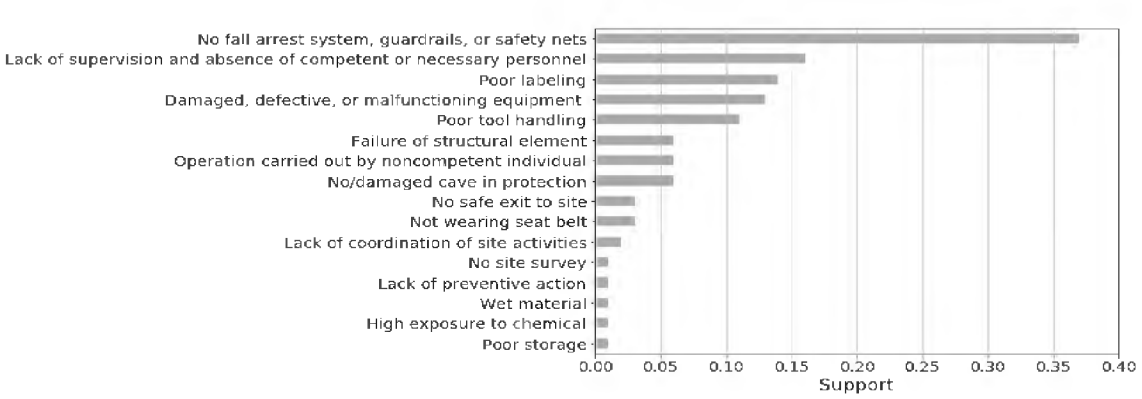

Figure 5.12 Support measure for the individual fatality causes in cluster 3.

As it could be seen from Figure 5.12, the most frequent three fatality causes within Cluster 3 are (1) no fall arrest system, guardrails, or safety nets; (2) lack of supervision and absence of competent or necessary personnel; and (3) poor labeling. This highlights the importance of providing the necessary fall arrest system, guardrails, or safety nets for the workers onsite, the critical role of having good supervision and component personnel, and the importance of labeling to the overall safety of the construction site. Figure 5.13 shows the obtained noteworthy combinations within Cluster 3 with a total of 14 obtained associations [Figure 5.13(a)] between the following 12 fatality causes [Figure 5.13(b)]: (1) no fall arrest system, guardrail, or safety nets; (2) wet material; (3) lack of preventive action; (4) lack/absence of supervision and competent/necessary personnel; (5) not wearing seat belt; (6) damaged/defective or malfunctioning equipment; (7) no/damaged cave in protection; (8) no safe exit to site; (9) operation carried out by noncompetent individual; (10) absence of site survey; (11) failure of structure element; and (12) and poor labeling.

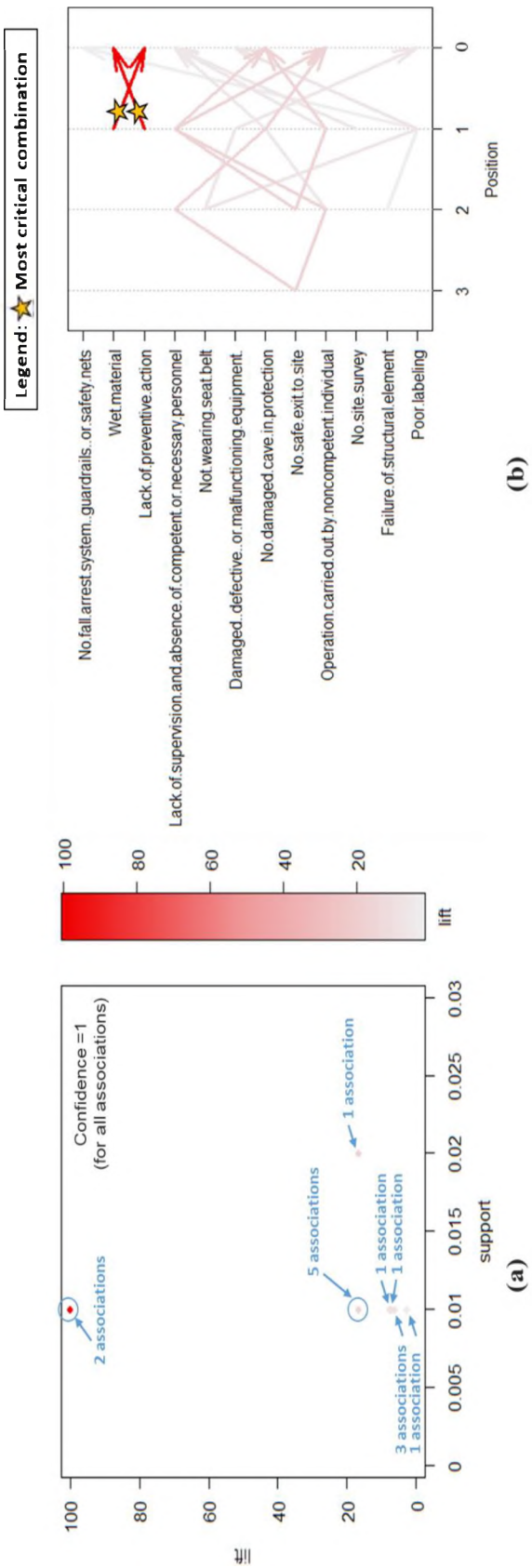

Figure 5.13 Identified remarkable combinations or associations within cluster 3.

Figure 5.13(b) shows that there are two most critical combinations, but these two combinations exist between the following same two fatality causes (but with different ordering): lack of preventive action and wet material. This combination has a support measure of 0.01, a confidence measure of 1, and a lift measure of 100. The interpretation of these measures reflects that there is a high chance of occurrence of this combination of fatality causes (confidence is equal to 1). In addition, a lift value of 100 > 1 reflects that the obtained association is of value (Yao et al., 2019) in the sense that the combinations of these two fatality causes appear more often than expected (IBM, 2019). That said, this combination reflects that the presence of these two causes of fatality on construction sites would serve as a warning sign that a fatal accident is very likely to occur.

5.5.2.4. Frequent items and associations within cluster 4. The obtained support measures for the individual fatality causes within Cluster 4 are shown in Figure 5.14. As it could be seen from Figure 5.14, the most frequent three fatality causes contributing to construction fatality accidents within Cluster 4 are (1) no job site inspection; (2) poor equipment handling and not following proper operation procedures or manufacturer's specifications; and (3) employee misconduct. Figure 5.15 shows the obtained noteworthy combinations or associations within Cluster 4. A total of two associations were obtained [Figure 5.15(a)] between three fatality causes [Figure 5.15(b)]: (1) lack of inspection for equipment and tools, (2) no jobsite inspection, and (3) inappropriate tools used. One notable and critical combination exists between these fatality causes with a support measure of 0.01, a confidence measure of 1, and a lift measure of 16.67. The interpretation of these measures reflects that there is a high chance of occurrence of this combination of fatality causes (confidence is equal to 1).

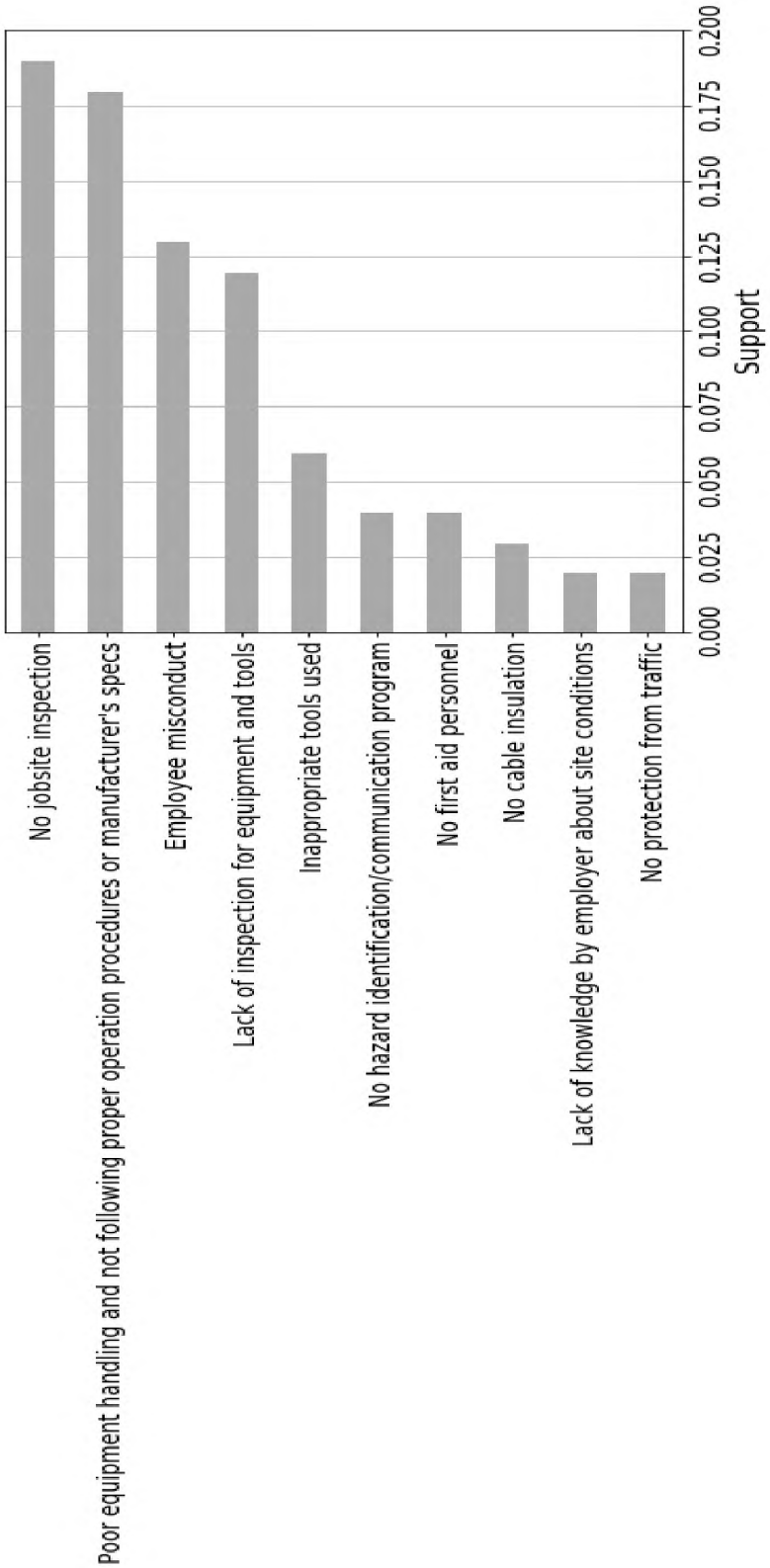

Figure 5.14 Support measure for the individual fatality causes in cluster 4.

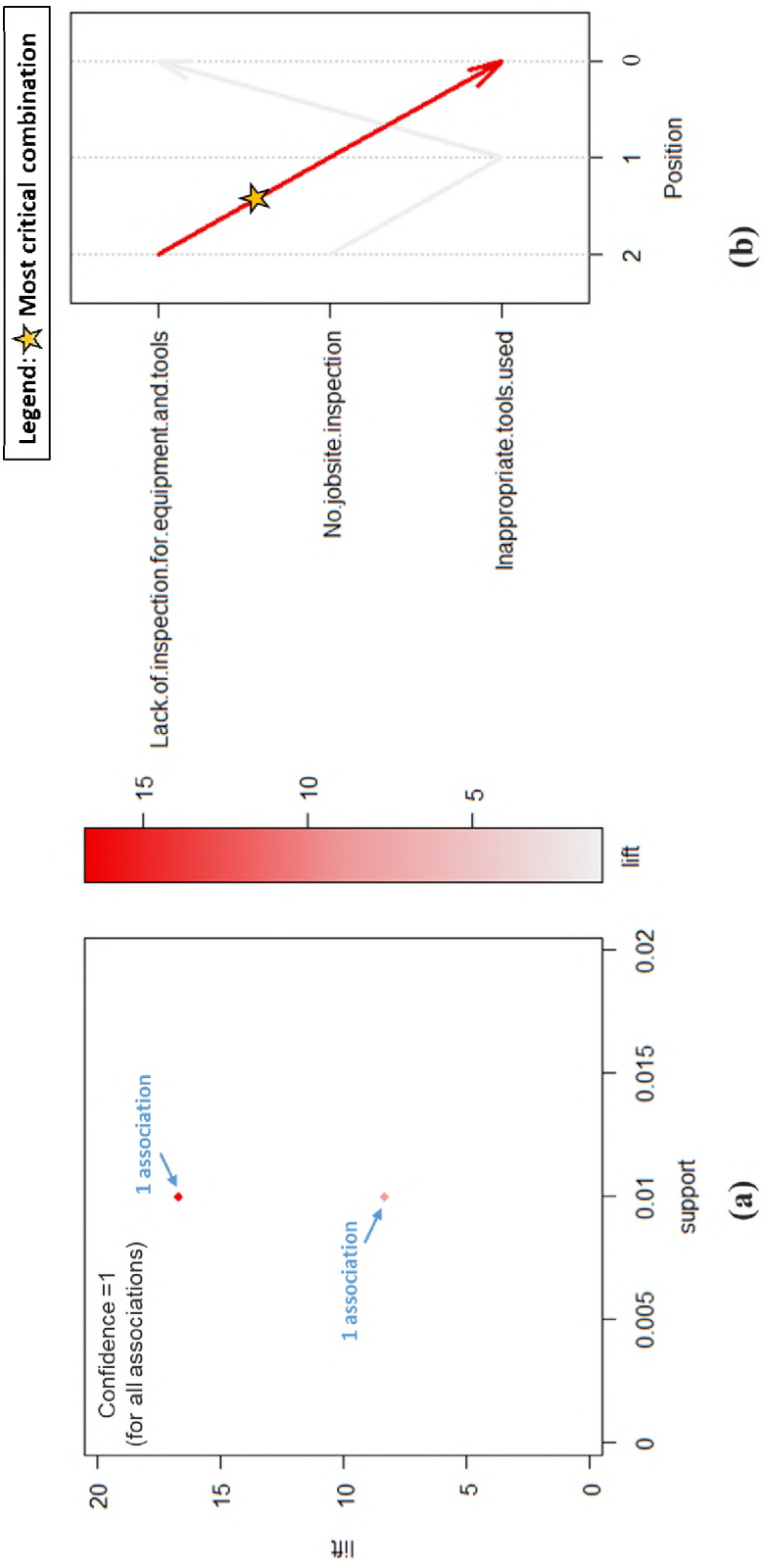

Figure 5.15 Identified remarkable combinations or associations within cluster 4.

In addition, a lift value of 16.67 > 1 reflects that the obtained association is of value (Yao et al., 2019) in the sense that the combinations of these three fatality causes appear more often than expected (IBM, 2019). That said, this combination reflects that the presence of these three causes of fatality on construction sites would serve as a warning sign that a fatal accident is very likely to occur.

5.5.2.5. Frequent items and associations within cluster 5. The obtained support measures for the individual fatality causes within Cluster 5 are shown in Figure 5.16.

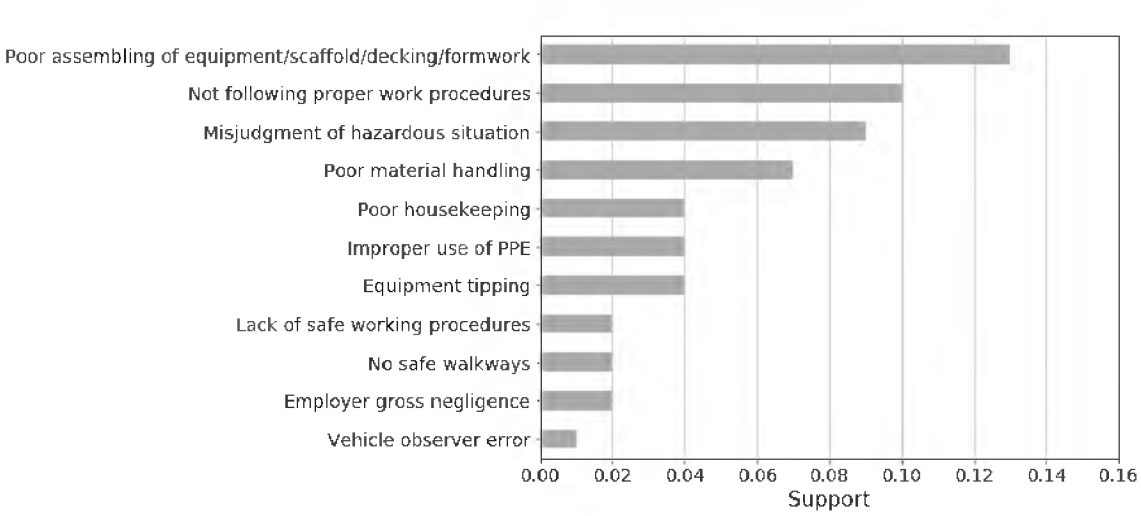

Figure 5.16 Support measure for the individual fatality causes in cluster 5.

As it could be seen from Figure 5.16, the most frequent three fatality causes contributing to construction fatality accidents within Cluster 5 are (1) poor assembling of equipment/scaffold/decking/formwork, (2) not following proper work procedures, and (3) misjudgment of hazardous situations. Also, Figure 5.17(a) shows the obtained noteworthy critical combination within Cluster 5.

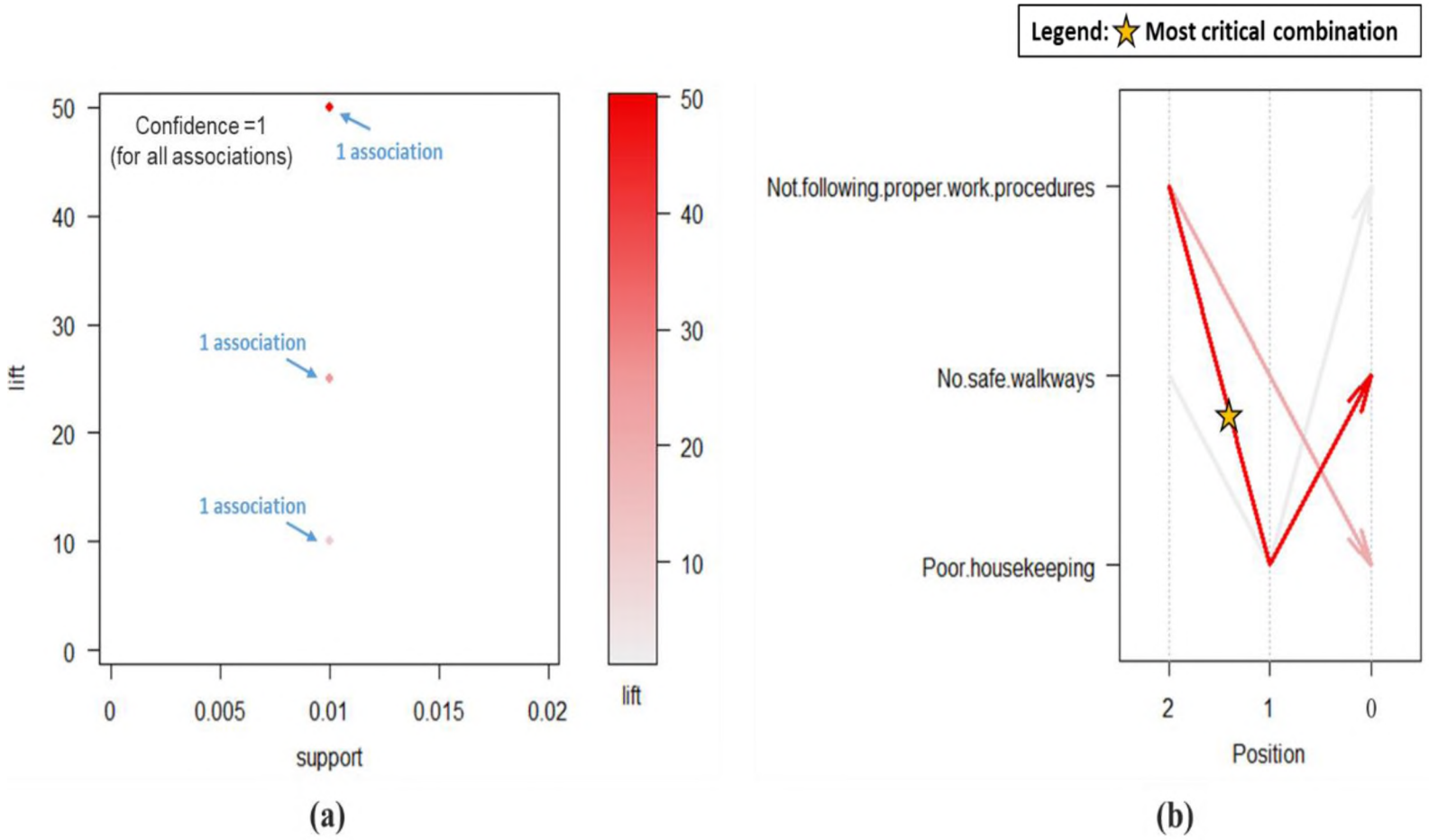

Figure 5.17 Identified remarkable combinations or associations within cluster 5.

A total of three associations were obtained [Figure 5.17(a)] between three fatality causes [Figure 5.17(b)]: (1) not following proper work procedures, (2) no safe walkways, and (3) poor housekeeping. One notable and critical combination exists between these fatality causes as shown in Figure 5.17(b). This combination has a support measure of 0.01, a confidence measure of 1, and a lift measure of 50. The interpretation of these measures reflects that there is a high chance of occurrence of this combination of fatality causes (confidence is equal to 1). In addition, a lift value of $50 > 1$ reflects that the obtained association is of value (Yao et al., 2019) in the sense that the combinations of these three fatality causes appear more often than expected (IBM, 2019). That said, this combination reflects that the presence of these three causes of fatality on construction sites would serve as a warning sign that a fatal accident is very likely to occur.

5.6. DISCUSSION

To reflect the validity of this research, this subsection discusses and interprets the obtained results and highlights the contributions of this section of the dissertation.

The traditional analysis of safety accidents involves the weakest link approach which claims that there is only one main cause for a given incident and that addressing that one cause will preclude a recurrence of the incident. However, the exclusive consideration of the single direct cause of safety incidents directs the main focus on the point of failure by assuming that this is also the best and most effective point of control to prevent accidents. Nevertheless, this exclusive focus could make it nearly impossible to seek out and deal with the interrelated root causes. In relation to that, Goldberg (2003) states that "It is not that hard to find what is apparently the weak link in almost any given incident

situation…In fact, [it] is…[t]oo easy…in the sense that once we do find the weak link, we tend to stop looking for any other sources of the problem…[while] there are very likely still more factors that have gone undiscovered. That being the case, there is very little likelihood of effective controls being designed to positively preclude a recurrence of the incident…It is vitally important to move past the notion that there is only one cause for an incident, or the almost congruent notion that only one thing needs to be corrected to preclude a recurrence." Since safety incidents are of a multivariable nature, the conventional approach to safety management is very much at odds with the contemporary philosophy on multiple causation factors in safety incidents (RiskWise, 2020).

On the combinations level, this section of the dissertation showed that construction fatalities would most likely happen when the following five critical combinations of factors exist on construction sites: (1) overexcavating, vibration in excavation, and error in design; (2) lack of general health and safety training, absence of decking, and lack safe access to site/scaffold/trench; (3) lack of preventive action and wet material; (4) lack of inspection for equipment and tools, no jobsite inspection, and inappropriate tools used; and (5) not following proper work procedures, no safe walkways, and poor housekeeping. Therefore, these results reflect the relationships between different fatality causes and how they combine together to form fatality accidents. This research is of particular importance with the modern thinking in the safety profession, which is directed towards the reliance on multiple causation factors in virtually all incidents (RiskWise, 2020). Unless and until safety practitioners are reasonably sure of the critical combinations of factors leading to the incident, they cannot make an effective decision on how to control the root causes behind the problem. That said, this section of the dissertation provides a better

understanding of the different combinations of factors that result in construction fatalities and offers insights on possible accident causation relationships in the construction industry. In addition, this section of the dissertation is of particular importance to the safety management practice in avoiding fatalities on real construction sites since it provides an approach that can make safety professionals avoid falling prey to the undesired phenomenon of satisfaction of search. This phenomenon occurs when finding a single cause to the incident leads to suspending the search, which if it had been continued would have resulted in finding other combination of causes to the incident (Carr, 2017).

On the individual level, the results showed the important role that the employer could play in minimizing fatalities on construction sites by providing the needed and adequate instructions and by putting robust safety requirements that do not compromise the safety of any worker on the jobsite with regard to performing any unsafe construction activity. In fact, despite that the employers' continuous failure to comply with the requirements of one or more of the occupational safety and health standards has been reported for decades (Moran, 1977), such instances still occur in today's construction activities. One of the main reasons behind that is attributed to the so called Proof of Employee Endangerment which states that "only where employees of a cited employer are affected by noncompliance with an occupational safety and health standard can such employer be in violation" (OSHRC, 2020) and to the so called Proof of Employer Knowledge which provides that in order to sustain a contested citation and alleged to be a violation, it should be proved that the cited employer had knowledge of the condition. Therefore, to prevent fatalities in the construction industry, employers shall implement reasonable precautionary actions to protect their employees from any reasonably

foreseeable dangers. Examples of such precautionary actions include having an adequate safety program with clear safety instructions to employees, well-established and defined work rules or instructions on safe procedures, diligent supervision, and the imposition of sanctions for noncompliance (Incident Prevention, 2016).

Second, the results reflected the importance of having proper job training in reducing fatalities on construction jobs. In fact, the poor construction safety performance is attributed to the general lack of proper training practices in the industry in relation to occupational safety and health (Evia, 2011), especially with regard to practices that are related to the specific characteristics of each construction activity. Therefore, to reduce construction fatalities, it is not enough that workers have the general and basic safety training; rather, employees shall also acquire the specific safety knowledge that is needed to perform the associated construction task. Third, the results reflected the importance of having proper fall arrest systems, guardrails, and safety nets when performing the construction activities. In fact, the lack of fall protection has been the most frequently cited violations by OSHA in the US construction industry (OSHA, 2017). Therefore, to reduce construction fatalities, an analysis of the work zone shall be conducted by (1) performing a hazard study to identify possible areas of risk (such as the height of the work to be performed and the number of workers working in the same area); (2) implementing proper fall-arrest systems such as anchorage, body support, connectors, and descent and rescue plan; and (3) properly calculating the fall clearance to have an adequate arrest of the fall before the worker hits the group or any other object.

Fourth, the results reflected the importance of having proper jobsite inspection to minimize or reduce the number of fatality accidents on construction jobsites. In relation to

that, since everyday use can take a toll on safety equipment, machines, and systems, it is very critical to have periodic inspections by (1) examining all equipment and tools before each use, (2) not using any damaged or malfunctioning equipment—at least until it is fully inspected and determined to be safe for further use, and (3) having a specialized team that is responsible for performing regular checks to critically identify and rectify hazards and document abatement. Thus, it is very crucial to have a clear safety inspection purpose, to set expectations, and to follow a well-defined inspection strategy. Fifth, the results showed that poor assembling of construction equipment and systems could lead to fatalities. Therefore, to prevent construction fatalities, it is important that the critical and complex construction equipment and systems be assembled by certified and well-trained workers that are specialized in performing such task. In addition, it is crucial to have proper manuals or training programs that details the associated erection, assembly, or installation of construction equipment, scaffold, decking, and formwork, among others.

5.7. SUMMARY

It is imperative to have a better understanding of the combinations and associations between the different fatality causes in construction sites. 'To this end, this section of the dissertation thoroughly studied 100 case files involving 60 causes using a data-driven approach involving spectral clustering, frequent pattern mining, and Apriori computational algorithms. Accordingly, the 60 fatality causes were determined to be categorized into five clusters with the following combinations being the most critical within each cluster: (1) Cluster 1—over-excavating, vibration in excavation, and error in design; (2) Cluster 2—lack of general health and safety training, absence of decking, and lack of safe access to

site/scaffold/trench; (3) Cluster 3—lack of preventive action and wet material; (4) Cluster 4—lack of inspection for equipment and tools, no jobsite inspection, and inappropriate tools used; and (5) Cluster 5—not following proper work procedures, no safe walkways, and poor housekeeping. All the aforementioned combinations across the different clusters are of importance because the existence of each combination would indicate that a fatality on the job site is very likely to occur.

As such, this section of the dissertation helped in identifying and quantifying the critical combinations and associations between the different fatality causes in construction sites. To this end, the findings provide warning signs in relation to the combinations and associations of the causes that are most probably to cause more fatalities compared to others. This section of the dissertation adds to the body of knowledge by equipping safety managers and supervisors with a proactive approach that allows them to take the needed preventive actions to avoid fatalities on construction sites by identifying, in hindsight, the critical combinations and associations of fatality causes. Ultimately, the outcomes would enhance the safety performance in the construction industry and prevent construction fatalities by exercising more effective and efficient practices to proactively prevent the occurrence of the identified safety fatality combinations on construction sites.

5.8. RELATED APPENDIX

Appendix D presents the used data and the Python and R codes for the developed hybrid unsupervised computational model for determining the critical combinations of safety fatality causes.

6. STUDYING THE IMPACT OF OFFSITE CONSTRUCTION TECHNOLOGY ON THE WORKFORCE AND LABOR CHARACTERISTICS

6.1. OVERVIEW

Despite that the construction sector is a substantial contributor and driver of the US economy, it has experienced a slow increase in its overall productivity when compared to other industries (Larsson et al., 2014). For instance, the productivity in the manufacturing sector has doubled during the last decades while the construction productivity has remained flat during the same period of time (Changali et al., 2015). In addition, the construction sector is facing considerable challenges due to shortages in skilled labor which became more pronounced after the Great Recession of 2007-2009 (CLMA, 2013). This is reflected by the fact that around 2.3 million construction employees were displaced after the Great recession, which constitutes around 30% of the workforce (Jones, 2018; Sundukovskiy, 2018).

The shortage of skilled labor has led many construction projects to suffer substantial cost and schedule overruns. To this end, companies in the construction industry have shifted their efforts to the use of manufacturing techniques and methods, namely offsite construction, to address the challenges of poor productivity, shortage of skilled labor, and unsatisfactory project performance (Barbosa et al., 2017; Larsson et al., 2014). In fact, offsite construction is considered as the construction equivalent of assembly-line production which is seen in the manufacturing sector (Kwiatek, 2018). Offsite construction showed great potential to deal with the industry's epidemic problems (Arashpour, et al., 2017).

Offsite construction — also sometimes referred to as offsite manufacturing, modular construction, or industrialization — is one of the most prominent disruptions currently experienced by the sector (Tuulberg, 2018). Offsite construction could be defined as the technique of exporting a portion of site-based work to off-site, such as at fabrication/modular assembly shops or yards (Choi et al., 2019a). That said, offsite construction includes different activities, being prefabrication of parts, sections, or entire units called modules in a factory-controlled environment (Burson, 2017). As such, the term 'offsite construction' is used to indicate the construction method that relies on prefabrication, preassembly, and modularization, rather than any specific type of two-dimensional (2D) or three-dimensional (3D) units or components.

While multiple research efforts have been conducted to study different characteristics of offsite construction, very little research studies were directed to examine the workforce-related aspects of offsite construction. In fact, a study conducted by Hanna et al. (2017) showed evidence that offsite construction users are planning to invest more effort in improving offsite construction practices within their companies by developing training programs for personnel and establishing protocols for working with key project stakeholders. However, no previous study has been performed to help industry practitioners in prioritizing such training needs and programs for the different workforce occupations involved in the offsite construction-related activities; being offsite or onsite. To this end, this section of the dissertation aims to address this workforce aspect of the offsite construction.

Offsite construction is a widely accepted alternative to conventional site-based construction since it provides many benefits compared to traditional stick-built

construction (Pan and Sidwell, 2011). In fact, offsite construction is an effective driver of growth in the construction industry and has made considerable progress around the world and (Mao et al., 2015). Offsite construction is comprised of three main aspects: (1) producing construction components in a controlled environment or factory; (2) transporting them to the site; and (3) installing the components or modules onsite (Zhang et al., 2020).

The shortage of skilled labor has led many construction projects to suffer substantial cost and schedule overruns. To this end, companies in the construction industry have shifted their efforts to the use of manufacturing techniques and methods, namely offsite construction, to address the challenges of poor productivity, shortage of skilled labor, and unsatisfactory project performance (Barbosa et al., 2017; Larsson et al., 2014). In fact, offsite construction has been used in different types of construction projects and has shown to enhance project performance on multiple metrics including improved predictability, lower labor and soft costs, shorter schedules, better quality, enhanced safety standards, reduced waste management, and lower demand for labor (O'Connor et al., 2013; Baldwin et al., 2009; Jaillon and Poon, 2008; O'Connor et al., 2016). Also, it is estimated that 25% of onsite time is spent creating value, while 75% of time spent offsite creates value (McKinsey & Company, 2019). Thus, many organizations are pushing offsite construction aggressively across all construction sectors and are achieving great results (FMI, 2018).

The American Institute of Architects and the National Institute of Building Sciences (2019) provided that there will be a significant shift towards offsite construction approaches in the coming years; especially if the shortage of labor continue to be exacerbated as large numbers of skilled construction workers retire and are not replaced.

In relation to that, a study conducted by Market Research Future (2017) predicted that the market for offsite/modular construction will flourish with a compound annual growth rate of 5.95% and a global market value of $154.8 million by the end of the forecast period (i.e. 2026). Another study conducted by Frost & Sullivan (2019) established that the global market of modular/offsite construction will grow with a compound annual growth rate of 6.3% as a result of the uptick in construction activities and significant cost, labor, and time savings associated with offsite construction (Limaye, 2019). Following the Asia-Pacific region, the North America area is the second largest region for offsite/modular construction with a market share of 27.6% and a market value of $28.7 million (Global Modular Construction Market Research Report, 2017).

While academic researchers and construction industry practitioners have invested many efforts to enhance the productivity over the lifecycle of offsite construction, particular attention is still needed to the technical and managerial aspects of offsite construction (Zhang et al., 2020). In fact, the technical and managerial skills of the workforce involved in the offsite construction operations play an important role in better leveraging such smart, modern, and innovative construction method and in enhancing the efficiency and productivity of the offsite construction works and activities.

Despite that multiple research efforts have been conducted to study different characteristics of offsite construction, very little research studies were directed to examine the workforce-related aspects of offsite construction especially as related to its impact on the technical and managerial skills of the workforce. In fact, a study conducted by Hanna et al. (2017) showed evidence that offsite construction users are planning to invest more effort in improving offsite construction practices within their companies by developing

training programs for personnel and establishing protocols for working with key project stakeholders. However, little or no previous study has been performed to help construction firms in prioritizing such training needs and programs for the different engineering, construction, and administrative workforce occupations involved in the offsite construction-related activities. To this end, this section of the dissertation addresses this unstudied workforce aspect of the offsite construction.

6.2. OBJECTIVE

The goal of this section of the dissertation is to investigate the impact of offsite construction on the workforce. In relation to that, the objectives of this section of the dissertation include: (1) identification of the different onsite and offsite construction workforce occupations and labor characteristics; (2) quantification of the impacts of offsite construction on the skill set of the identified onsite and offsite workforce occupations; (3) evaluation of the impacts of offsite construction on the demand for the identified onsite and offsite workforce occupations; (4) assessment of the impacts of offsite construction on the identified labor characteristics; (5) identification of the different engineering, construction, and administrative workforce occupations involved in the offsite construction operations; (6) quantification of the impacts of offsite construction on the technical skills of the identified engineering, construction, and administrative workforce occupations; (7) evaluation of the impacts of offsite construction on the managerial skills of the identified engineering, construction, and administrative workforce occupations; and (8) prioritization of the impacts of offsite construction for the identified engineering, construction, and administrative workforce occupations. To this end, this section of the dissertation helps

companies in better understanding and realizing the implications of offsite construction as related to the key workforce occupations and labor characteristics. This would ultimately help industry practitioners in workforce planning and management, in the prioritization of training needs and programs, and in improving the quality of the workforce involved in the offsite construction operations; being onsite or offsite.

6.3. CURRENT STATE OF LITERATURE AND ASSOCIATED LIMITATIONS

Many research efforts were directed to study multiple aspects of offsite construction. Therefore, this subsection provides the current state of knowledge and discusses the knowledge gap present in the literature.

6.3.1. Existing Offsite Construction Related Research Efforts. Previous work was conducted to perform a comparison of worker safety risks between offsite construction and onsite methods and to provide an empirical and evidence-based explanation for why offsite construction can help reduce safety risks on construction sites (Ahn et al., 2020). Other studies provided a multifaceted productivity comparison of offsite timber manufacturing strategies between the Mainland Europe (EU) and the United Kingdom (UK), and it was found that the labor productivity of the surveyed UK panelized and EU volumetric manufacturers was comparable but the UK volumetric manufacturers' productivity was lower (Duncheva and Bradley, 2019). Other efforts included addressing unbalanced resource distribution and efficiently managing resource-constrained production in offsite construction by examining the use of different labor structures to address bottlenecks in production (Nasirian et al., 2019). Previous studies also developed learning curve models related to workers' mastery of the manufacture of precast

components (a type of offsite construction) and to determine the worker time for precast component production in construction (Tai et al., 2021). In addition, a study was performed to assess the state of prefabrication practice in the electrical construction industry, and it provided that prefabrication users are planning to invest more effort in improving prefabrication practices by developing training programs for personnel and establishing protocols for working with suppliers and vendors (Hanna et al., 2017).

On the other hand, efforts were directed to study the effects of lean construction on the sustainability of modular homebuilding, and recommendations were provided on how to address the barriers to the widespread application of sustainable homebuilding (e.g., higher initial costs largely attributable to the learning curve of workers building with these practical innovations and technologies and the added cost resulting from ill-defined construction processes) (Nahmens and Ikuma, 2012). Moreover, previous work modeled and predicted the likelihood of prefabrication feasibility for electrical construction firms and found that prefabrication feasibility is significantly dependent on (1) four industry-related determinants: regional economic growth, industry competition, labor cost rate, and worker union resistance; and (2) two main internal firm-related determinants: building information modeling (BIM) capability and supply coordination with vendors (Said, 2016). In addition, a delivery framework was presented for multi-story modular buildings to address both the project-based and product-based nature of these buildings and to outline a baseline to extend the industry foundation classes (IFC) data schema and enable it to model specific elements of modular buildings (Ramaji et al., 2017). Furthermore, decision-making factors that affect the use of offsite construction were identified based on a comprehensive review of the literature, and it was found that a total of 50 factors impact

the various modular construction operations in the construction industry (Abdul Nabi and El-adaway, 2020). Since mandatory law provisions take precedence over contractual stipulations by the parties (Assaad and Abdul-Malak 2020a), other research efforts studied the commercial and legal considerations of offsite construction projects and their hybrid transactions and offered guidance to the management, commercial, and legal practitioners on different aspects related to offsite construction transactions (Assaad et al., 2020e). Also, the changes needed in current engineering, procurement, and construction processes were determined to create optimal environment for a broader and more effective use of modularization, and it was found that project teams shall pay close attention to module envelope limitations, team agreement on project drivers, adequate owner-planning resources and processes, timely freeze of scoping and design, and due recognition of possible early completion from modularization (O'Connor et al., 2014).

Furthermore, previous research efforts investigated the combinatorial effects of modularization critical success factors on the cost and schedule performances of industrial modular projects and proposed a conceptual model (Choi et al., 2016). In addition, the feasibility, challenges, and critical success factors of modular integrated construction were investigated, and a feasibility index was proposed based on performance levels and aggregation of global weights of critical success factors (Zhang et al., 2021). Moreover, a methodology was introduced for the dynamic assessment and proactive management of excessive geometric variability issues in modular construction projects (Enshassi et al., 2020). Finally, previous efforts were conducted to address a leveraging opportunity for modularization augmentation by examining how modularization and design standardization relate to one another in the industrial sector by providing insights into the

characteristics of modular projects with standard design, assessment of its impact and benefits, and lessons learned (O'Connor et al., 2015).

6.3.2. Knowledge Gap. Based on the extensive literature review of previous research studies on offsite/modular construction performed in the previous subsection, it could be concluded that multiple research efforts have been directed to examine different characteristics of offsite construction such as critical success factors, opportunities for modularization augmentation, effects of lean construction on the sustainability of modular homebuilding, prefabrication feasibility, industry foundation classes (IFC) data schema in modular buildings, decision-making factors affecting the use of offsite construction, commercial and legal considerations of hybrid offsite construction projects and transactions, creating optimal environment for a broader and more effective use of modularization, proposing a feasibility index, and managing excessive geometric variability issues. However, little research studies were directed to study the workforce-related aspects of offsite construction. More specifically, the previous studies that investigated labor-related characteristics mainly focused on (1) performing a comparison of worker safety risks between offsite construction and onsite methods, (2) providing a productivity comparison of offsite timber manufacturing strategies between different geological regions and areas, (3) addressing unbalanced resource distribution and efficiently managing resource-constrained production in offsite construction; and (4) developing learning curve models for workers to master some of the manufacturing aspects of offsite construction. Thus, no previous study has been performed to study the impact of offsite construction on the onsite and offsite construction workforce, on different labor characteristics, and on the engineering, construction, and administrative workforce. To this

end, there is a knowledge gap in the current body of knowledge as related to helping industry practitioners in prioritizing training needs and programs for the different workforce occupations involved in the offsite construction-related activities as well as in understanding the implications of offsite construction on multiple, rather than specific, workforce-related characteristics. That said, this section of the dissertation addresses this critical research need and knowledge gap.

6.4. METHODOLOGY

To attain the research goal and objectives of this section of the dissertation, an interdependent multi-step methodology was followed as detailed in the following subsections.

6.4.1. Formation of a Panel of Industry Practitioners. A panel of 19 industry practitioners was developed at the beginning of this research as appointed by the Construction Industry Institute (CII) for research team (RT)-371 at Missouri University of Science and Technology, Purdue University, and the University of Arkansas. The established industry panel has actively participated throughout the duration and steps of the research through periodic face-to-face and virtual meetings as well as conference calls to ensure that the outcomes, conclusions, and recommendations of this research are highly practical and beneficial to the industry. The members of the panel come from a range of industrial sectors, with a mix of owner, contractor, and service providers organizations. It is worth mentioning that this approach has been widely followed by CII in almost all its applied research projects and has been proven to offer robust and valid findings. To name a few, Goodrum et al. (2011) established a panel of 18 industry members to provide

thresholds for successful, inconclusive, and unsuccessful technologies and to assess the impact of technology on productivity in the construction industry. Austin et al. (2016) created an industry expert panel from 15 practitioners to examine multiple concepts or practices that are essential for the success of flash track projects. O'Connor and Mock (2019) relied on a panel of 16 industry practitioners to provide industry insights, assist with data collection, and help improve research products to examine and characterize common problematic activities across the industrial sector.

6.4.2. Identification of the Offsite and Onsite Workforce Occupations and the Labor Characteristics. Throughout several meetings and online workshops with the industry panel — and based on the occupational profiles provided by the US Bureau of Labor Statistics (2020b); the list of the crafts, titles, disciplines specified by the NCCER (2020); and a review of the literature — the RT-371 team (1) identified a list of offsite construction workforce occupations; (2) identified a list of onsite construction workforce occupations; and (3) identified a list of labor characteristics. The identified offsite and onsite workforce occupations as well as the identified labor characteristics are shown in Figure 6.1 and 6.2, respectively. It is worth mentioning that since offsite construction involves performing the work in a controlled environment and then shipping or assembling it onsite, the construction workforce — involved in current offsite construction projects — performs hybrid construction activities; that is, both offsite construction tasks and traditional onsite construction activities (Arashpour et al., 2016). Therefore, the industry panel considered both offsite and onsite construction occupations in this research as shown in Figure 6.1.

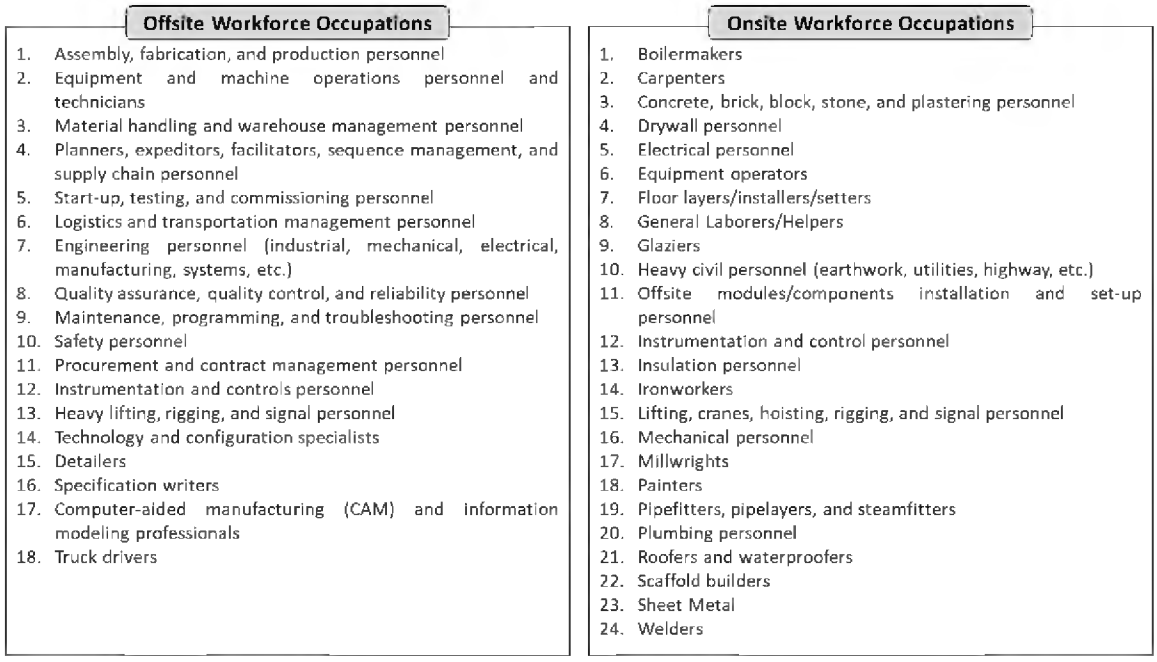

Figure 6.1 Offsite and onsite workforce occupations.

Moreover, the state of the workforce and its demographics are affected by different aspects including gender, diversity, age, education, and workers membership (Carmeli and Weisberg, 2006; Keating, 2019; Goodrum, 2003). As such, these workforce demographic aspects (i.e., gender, diversity, age, education, and workers membership) were considered in this research, with each workforce demographic aspect being divided into numerous sub-elements as shown in Figure 6.2, in addition to the identified workforce attributes that are also shown in Figure 6.2. To ensure the breadth of the identified list of occupations and labor characteristics, the respondents of the prepared survey (the survey details are provided in the next subsections) were provided with the option to add any extra offsite workforce occupations, onsite workforce occupations, or labor characteristics that they see missing or of interest to the industry or to them.

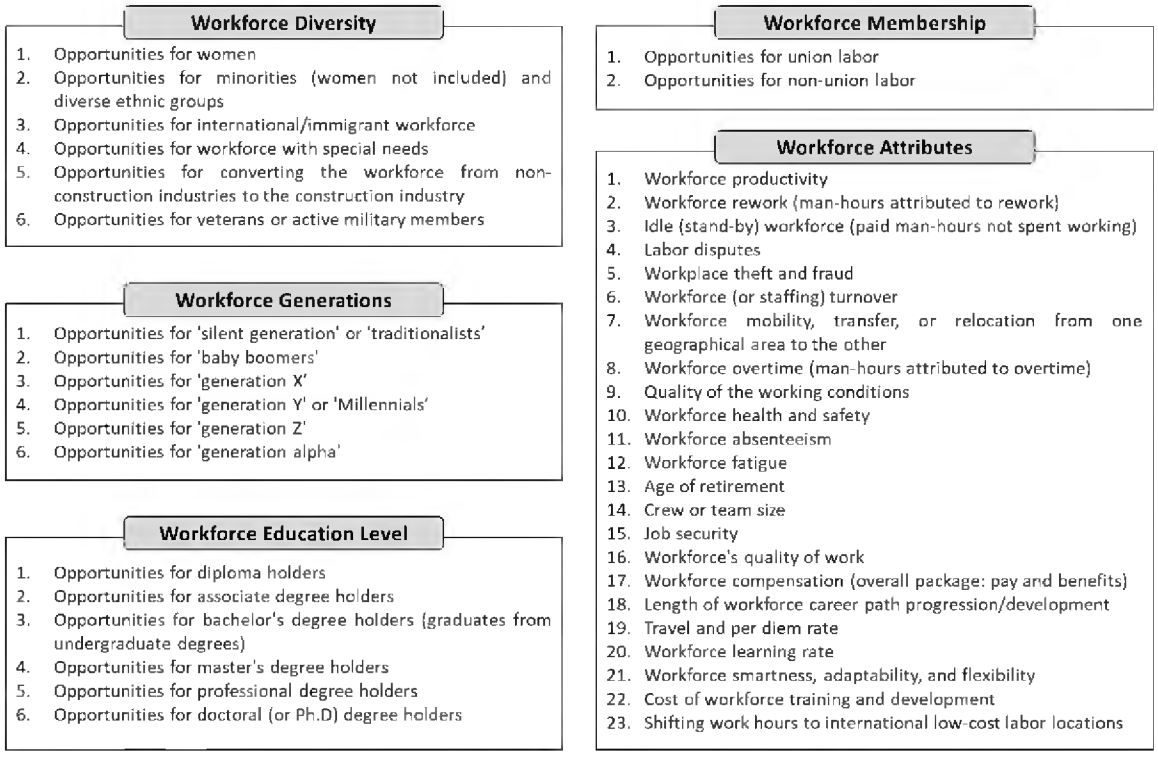

Figure 6.2 Labor characteristics.

6.4.3. Identification of the Engineering, Construction, and Administrative Workforce Occupations. The RT-371 followed the team idea mapping brainstorming method for the identification of the different workforce occupations. This was possible throughout several meetings and online workshops with the industry panel — and with reference to the occupational profiles provided by the US Labor Statistics (2020) as well as the occupations in the industry panel's companies. This allowed the RT-371 team to (1) identify a list of (1) engineering workforce occupations; (2) construction workforce occupations; and (3) administrative workforce occupations.

The team idea mapping method is a type of brain storming technique for group deliberation in a nonjudgmental and friendly environment as it encourages participation by

all the members (Goswami et al., 2020). The team idea mapping method is a powerful whole-brained visual thinking tool that enhances thought organization, planning, communication, and memory, and it is used as a powerful tool for developing collective creativity among the team members (Joseph and Arockiamary, 2014). In addition, the team idea mapping method was chosen in this research due to its practically and benefits compared to other methods — as listed by Markov (2018); Joseph and Arockiamary (2014); Ceauşu (2015); Wilson (2020); Kajabi (2021), including: ensuring that all members participate and no ideas are rejected; allowing a large volume of different ideas; helping in structuring information and better analyzing, comprehending, synthesizing, recalling, and generating new ideas; improving collaboration; making all the members of the team work by association wherein each member puts forward their view; contributing to creative outcomes; offering broader perspective in the variety of ideas; connecting sub-ideas and letting everyone contribute their thoughts on each potential avenue; and giving the group members the increased ability to more competently plan, organize, communicate, remember, innovate, and accomplish faster.

The standard procedures or steps for the team idea mapping method were followed in this research as follows, and as described in Joseph and Arockiamary (2014): (1) presenting a common topic for analysis to the group (i.e., the identification of the engineering, construction, and administrative workforce occupations in this research), (2) encouraging a pool of answers from all the members of the team, (3) creating a common idea map drawing or table using blackboard, and (4) consolidating the answers and driving the solution (i.e., the final list of workforce occupations). During the consolidation phase, participants may discover a common interpretation or understanding of the identified items

under scrutiny as they share the meanings behind their ideas, which enables new ideas to arise by association and to be added to the map as well (Goswami et al., 2020). Once all the ideas are captured, the group can prioritize and/or take the needed actions or final decision (Markov, 2018). To this end, the identified workforce occupations was the result of a face-to-face brainstorming workshop activity between the members of the industry panel, and the final list of the identified workforce occupations is shown in Figure 6.3.

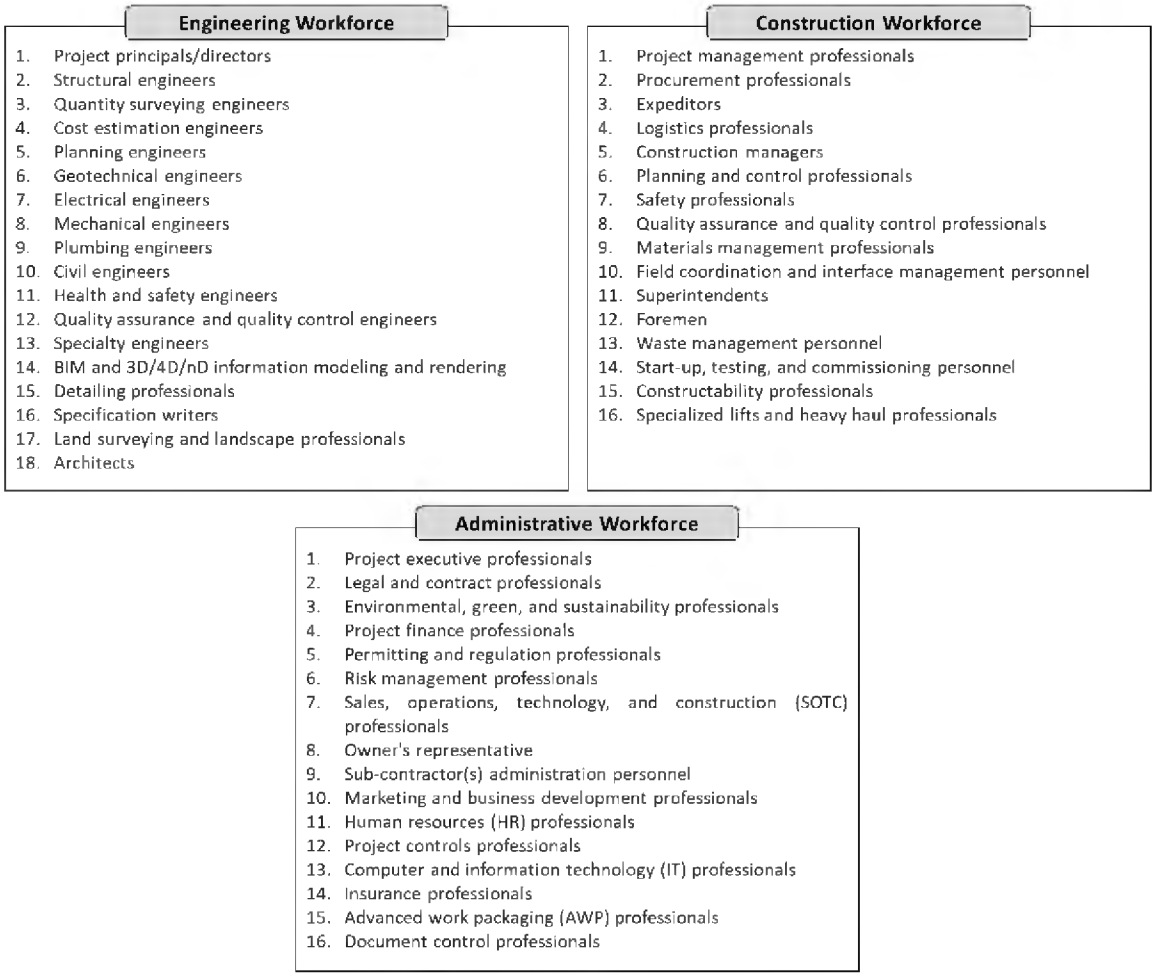

Figure 6.3 Engineering, construction, and administrative workforce occupations.

As shown in Figure 6.3, (1) the engineering workforce includes different occupations that are usually involved in the early phases of the projects (i.e., planning and design), (2) the construction workforce in this section of the dissertation refers to the managerial workforce occupations that are usually involved in the construction phase of the projects (and sometimes in the early phases to provide constructability feedbacks), and (3) the administrative workforce includes different occupations that are usually involved throughout most phases of the projects and that are responsible for administrative aspects. To further ensure the breadth of the identified list of occupations shown in Figure 6.3, the RT-371 team provided the respondents of the prepared survey with the option to add any extra engineering, construction, and administrative workforce occupations that they see missing or of interest to the industry or to them.

6.4.4. Survey Development. A survey was developed to (1) determine the impact of offsite construction on the skillset (reskilling or upskilling) of the (a) offsite construction workforce and the (b) onsite workforce; (2) determine the impact of offsite construction on the demand (shrink or grow) for the (a) offsite construction workforce and the (b) onsite workforce; (3) determine the impact of offsite construction on the labor characteristics; (4) determine the impact of offsite construction on the *technical* skill set of the (a) engineering workforce occupations, (b) construction workforce occupations, and (c) administrative workforce occupations; (5) determine the impact of offsite construction on the *managerial* skill set of the (a) engineering workforce occupations, (b) construction workforce occupations, and (c) administrative workforce occupations.

6.4.4.1. Likert scale used for the onsite workforce occupations, the offsite construction occupations, and the labor characteristics. To ensure reliable results, it

was important that the Likert scale be explicitly defined to the respondents so they would all have the same and consistent understanding of the meaning of the different scale numbers. This also minimizes qualitative bias by the respondents. Thus, the RT-371 team utilized the standard Likert scale that was developed by CII (2013), which has been used in multiple previous research studies. To this end, a double sided 5-point Likert scale (shown in Table 6.1) was used to understand the impact of offsite construction on the onsite and offsite workforce occupations and on the labor characteristics. In relation to that, a positive score reflects the level of upskilling of the skillset of the corresponding workforce occupation, the level of growth of the demand for the corresponding occupation, and the degree of increase for the corresponding labor characteristic, and a negative score reflects the level of reskilling of the skillset of the corresponding workforce occupation, the level of shrinkage of the demand for the corresponding occupation, and the degree of decrease for the corresponding labor characteristic. The used Likert scale is shown in Table 6.1, and the results are reported on both the 5-point scare as well as the equivalent percentage scale.

Table 6.1 Used Likert scale; adapted from CII (2013).

Description of the scale	Associated range
± 1 = Negligible	will change by <5%
± 2 = Minor	will change by 5-10%
± 3 = Moderate	will change by 10-20%
± 4 = Significant	will change by 20-50%
± 5 = Extreme	will change by >50%

Also, the respondents had the option to select 'no change' if they perceive that there will be no change in the associated workforce occupation or characteristic. To further

ensure consistent understanding of the Likert scale by the respondents, 'upskilling' was defined for all respondents as "the process of learning new skills within the same job profile due to the increased shift towards offsite construction; the employee or worker improves its current skill set", 'reskilling' was defined for all respondents as "the process of learning new skills to do a different job due to the increased shift toward offsite construction; the employee or worker might earn a completely new degree or certification", growth was defined for all respondent as "the demand or need for the occupation will increase", and shrinkage was defined for all respondents as "the demand or need for the occupation will decrease".

6.4.4.2. Likert scale used for the engineering, construction, and administrative workforce occupations. To ensure reliable results, it was important that the Likert scale be explicitly defined to the respondents so they would all have the same and consistent understanding of the meaning of the different scale numbers. This also minimizes qualitative bias by the respondents. Thus, the RT-371 team utilized the standard 5-point Likert scale which was developed by CII (2013) and has been used in multiple previous research studies, where 1 = Negligible Impact, 2 = Minor Impact; 3 = Moderate Impact; 4 = Significant Impact; and 5 = Extreme Impact. For the full description of each point scale, interested readers could check CII (2013).

To further ensure consistent understanding of the Likert scale by the respondents, 'technical skills' were defined for all respondents as "the teachable and measurable abilities/knowledge needed to perform specific tasks. They are often gained by qualifications". Also, 'managerial skills' were defined to all respondents as "the attributes

or abilities that individuals possess to fulfill some specific management activities or tasks. This knowledge/ability is usually learned and/or practiced through experience."

6.4.5. Pilot Testing of the Survey. Although the developed survey was reviewed carefully through the entire survey development efforts, the survey was pilot tested to ensure maximum benefits and eliminate any mistakes. Respondents of the pilot study were asked at the end of the survey to provide their comments on the survey as related to potential items or aspects that need to be added, modified, or deleted; questions that need to be added, modified, or deleted; clarifications that need to be performed to ensure consistency of understanding; and any other suggestions to make the survey more beneficial. The pilot survey was completed by 11 industry professionals which is considered satisfactory — as suggested by Connelly (2008) where a pilot study sample is recommended to be 10% of the sample projected for the larger parent study — since a sample size of 100 responses was collected for the larger parent survey in this research as detailed in the next subsections.

6.4.6. Survey Modification based on the Results of the Pilot Testing. Comments received from the respondents in the pilot testing were recorded and the survey was fine-tuned accordingly. The pilot testing enabled some editing of the questions' language and related descriptions, some formatting and aesthetic considerations, the removal of some duplications, and the enhancement and addition of some clarifications and items; however, nothing major was changed.

6.4.7. Survey Distribution. The survey was distributed and developed through Qualtrics which is a cloud-based platform used to create and distribute web-based surveys. Qualtrics is widely used for academic research due to its exceptional abilities of having a

large array of question types, possessing highly customizable survey designs and appearances, and allowing for complex experimental designs and user-tailored survey paths.

The survey was distributed to respondents determined by the industry panel members and research collaborators. The criteria used for the identified industry professionals included industry experts (1) having experience with the US construction industry, (2) having experience with either offsite construction, the craft workforce, and/or the integration and usage of construction technologies; (3) representing one of the central project stakeholders: (a) owners or developers; (b) architects, engineers, or service providers; and (c) contractors, construction managers; or fabricators; and (4) representing one of the key industry sectors: (i) building and commercial; (ii) industrial; and (iii) infrastructure. Overall, the survey was sent to 215 construction professionals. A total of 131 completed the survey (which is equivalent to an overall response rate of 60.93%), but 31 provided incomplete and insufficient responses. As such, 100 responses were included in the analysis (which is equivalent to a useable response rate of 46.51%).

6.4.8. Statistical and Quantitative Analysis. Different statistical and quantitative analyses were performed for the obtained results. This subsection provides all the details related to the performed analyses.

6.4.8.1. Reliability statistical analysis. Two types of reliability are generally of interest: internal consistency "internal reliability" and interrater agreement "external reliability" (Park and Jung, 2003). As such, this subsection provides all details pertaining to the conducted reliability analyses.

Internal consistency using Cronbach's alpha reliability test: To check the internal reliability (internal consistency) among the collected responses, the Cronbach's alpha coefficient was calculated. The coefficient of Cronbach's alpha is used to interpret the reliability of factors retrieved from either dichotomous or multipoint scales (Santos, 1999). A Cronbach's alpha value of 0.75 and above represents a reliable and valid questionnaire (Christmann and Aelstb, 2006) since it reflects that all respondents have the same understanding of the survey questions, and thus making the survey valid and reliable. To this end, the Cronbach's alpha test was applied to examine the reliability of the scales used for determining the impact of offsite construction on the skillset of the workforce occupations, determining the impact of offsite construction on the demand for the workforce occupations, and determining the impact of offsite construction on the labor characteristics.

External reliability using intraclass correlation coefficient: Since measurement error can seriously affect statistical analysis and interpretation (Shrout and Fleiss, 1979), it is important to check the rater consistency or agreement (i.e., inter-rater (external) reliability) among the collected responses. That said, the intraclass correlation coefficient (ICC) was used since it a widely used index in interrater reliability analyses as it reflects not only the degree of correlation but also agreement between measurements (Koo and Li, 2016). In fact, the interrater reliability or agreement reflects the variation between 2 or more raters who measure the same group of subjects (Koo and Li, 2016). The ICC value ranges between 0 and 1, with values closer to 1 representing stronger agreement or reliability. According to guidelines developed by Cicchetti and Sparrow (1981) — which resemble closely those developed by Fleiss (1981), when the ICC is below .40, the level of

agreement is poor; when it is between .40 and .59, the level of agreement is fair; when it is between .60 and .74, the level of agreement is good; and when it is between .75 and 1.00, the level of agreement is excellent.

6.4.8.2. Quantitative calculation of the overall impact of offsite construction on the onsite and offsite workforce. After collecting the data from the respondents, the overall impact of offsite construction on the onsite and offsite workforce occupations was calculated. Each respondent *r* inputs the *skillset* impact (SI_{ri}) and the *demand* impact (DI_{ri}) for each workforce occupation i. The overall impact (OI_i) for each workforce occupation i is computed by multiplying the corresponding *skillset* impact and *demand* impact that are inputted by the respondents. This overall impact aims to help in evaluating and prioritizing the training needs and programs for the different workforce occupations. To this end, the average overall impact was quantified for each one of the offsite and onsite workforce occupations using Equation (34).

$$OI_i = \frac{1}{N}\sum_r SI_{ri} \times DI_{ri} \tag{34}$$

where N is the total number of collected responses (which is 100).

6.4.8.3. Prioritization of impact on the engineering, construction, and administrative workforce using *k*-means clustering. While studying the impact of offsite construction on the technical and managerial skills of the engineering, construction, and administrative workforce occupations is important, it is also of great value to prioritize the impact of offsite construction on the different workforce occupations (e.g., to help in prioritizing training needs, programs, and plans) by taking into consideration, or factoring,

both the technical and managerial impacts. To this end, k-means clustering was used for this purpose. This algorithm was employed because it is perceived to be feasible for the intended outcome and due to its many pronounced benefits including (1) being able to solve prioritization problems (Achimugu et al., 2014), (2) being one of the most commonly and effective methods to classify data because of its simplicity and ability to handle voluminous data sets (Anchalia et al., 2013), (3) being a widely used and simple approach for data clustering (Hassanzadeh and Meybodi, 2012), (4) being able to cluster observations into groups of related observations without any prior knowledge of those relationships (Shafeeq and Hareesha, 2012), and (5) being a versatile algorithm that has been used in diverse applications (Bhimani et al., 2015), and thus it proved its practicality, viability, functionality, convenience, and usefulness in solving and addressing many problems and decisions.

The k-means clustering algorithm is a method of cluster analysis which partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean (Chandhok et al., 2012). The k-means clustering algorithm is generally implemented as follows (Hassanzadeh and Meybodi, 2012): (1) choosing the number of clusters which is referred to as '*k*'; (2) choosing distant and distinct centroids for each of the chosen sets of k clusters; (3) considering each element of the given set and comparing its distance to all the centroids of the k clusters, and based on the calculated distance, the element is added to the cluster whose centroid is nearest to the element; (4) re-calculating the cluster centroids after each assignment or a set of assignments; and (5) continuously updating until no further change is reached, as this is an iterative method.

One of the key needed input parameters for the k-means clustering algorithm is the number of clusters *k*; hence, it needs to be determined. In relation to that, the silhouette score/width method (Rousseeuw, 1987) was used since it is one of the most popular techniques for determining the optimal number of clusters (Chang and Chi, 2019) where the optimal value *k* of clusters is identified based on the highest silhouette score. In relation to that, Oskouie et al. (2017) provided that: "The silhouette method uses a similarity measure that is defined based on the distance…of the data points from their own clusters (intracluster variance) and the comparative distance to other clusters (intercluster variance)..Therefore, the k value that produces the highest average silhouette value represents the most optimal clustering configuration". Once the optimal number of clusters is determined, it is used for each one of the three categories of the workforce occupations: (a) engineering, (b) construction, and (c) administrative. The silhouette score is calculated using Equation (35).

$$S = \frac{b-a}{max\ (a,b)} \tag{35}$$

where S is the silhouette score, a is the mean intra-cluster distance, and b is the distance between a sample and the nearest cluster that the sample is not a part of.

6.5. RESULTS AND ANALYSIS

This subsection provides the results and associated analysis in relation to the collected data.

6.5.1. Respondent Demographics. A total of 100 survey respondents representing major stakeholders in offsite construction projects was obtained as follows: 36% represent owners; 25% represent architects/engineers; and 39% represent builders. In addition, the distribution of the respondents among the major sectors is as follows: 73% industrial sector; 21% building and commercial sector; and 6% infrastructure sector. Moreover, the average industry experience of the surveyed respondents is 28.05 years where most of the respondents (92%) have an industry experience of 10 years and more. On the other hand, the average experience of respondents in offsite construction is 13.92 years, where 64% of the respondents possess at least 10 years of experience in offsite construction. Furthermore, most of the respondents have senior management and high job positions levels. Furthermore, and to ensure that the respondents represent an acceptable range of experience in the industry, the experience of the surveyed industry practitioners in this section of the dissertation was compared with other similar survey-based research studies related to the construction field. For instance, Abdul Nabi and El-adaway (2021) conducted a survey for respondents with an average experience of 24.3 years in the industry and an average experience of 13.6 years with offsite construction to identify the key risks affecting cost and schedule performance of offsite construction projects. Choi et al. (2020) relied on a survey data from professionals with an average experience of 21.75 years in the industry to determine the innovative technologies and management approaches that promote modularization in capital projects and higher levels of design standardization. Thus, compared to previous studies, the collected responses reflect a wide range of experience in the national construction industry as well as in the national offsite construction market.

Ultimately, such wide experience ensures that the collected responses are reliable enough to be considered a good representative of the US offsite construction industry.

6.5.2. Sufficiency of the Response Rate and Sample Size. It is very important to check the sufficiency of the response rate to identify whether the data collected from the respondents could be considered as an adequate representative sample, and thus ensuring a solid basis for the conducted analysis. To this end, the sufficiency of the data collected from the 100 respondents was examined based on both empirical examination of previous research work as well as commonly used statistical techniques.

Since the construction industry is known to have a lack of participation in questionnaires (Cheong Yong and Emma Mustaffa, 2012), many studies have provided a range for the acceptable response rate in survey-based construction research work. In relation to that, according to Fellows and Liu (2015), the useable response rate is between 25% and 35% in the construction research since "survey techniques, such as questionnaires, interviews and so on, are highly labour intensive on the part of respondents and, particularly, on the part of the researcher." In addition, according to Akintoye (2000), the most used response rate in construction research involving surveys is in the 20%–30% range. Moreover, Ryal-Net and Kaduma (2015) provided that 30% is an acceptable response rate in construction studies. Further, according to Tan et al. (2014), the standard research survey rate in the construction industry is between 10% and 20%. Finally, Assaad et al. (2020c) highlighted that the most common response rates in survey-based construction research falls between 10% and 30%. As far as the total number of collected responses is concerned, different studies suggested minimum sample sizes to be used for survey-based research as follows: (1) 15 to 35 respondents (Fowler, 1995); (2) 30 to 50

respondents (Sudman, 1983); and (3) 25 to 75 respondents (Converse and Presser, 1986). To this end, the obtained useable response rate of 46.51% and the total of 100 responses in this section of the dissertation are considered acceptable because they are higher than the commonly used range for the response rate and the number of surveyed industry practitioners in survey-based research studies.

Although the response rate sufficiency is proved empirically, it is still very important to conduct a statistical verification. As such, Equation (36) was used, which was first established by Cochran (1977) and utilized later by many previous construction research studies such as Fellows and Liu (2015) and Srour et al. (2017), to name few. More specifically, Equation (36) calculates the minimum number of respondents needed to ensure meaningful findings and valid generalization of the results. After computing the minimum required responses, the obtained minimum sample n shall be compared with the 100 obtained responses.

$$n = \frac{t^2 s^2}{e^2} \tag{36}$$

where n is the minimum required number of respondents; t is the Z-statistic for a given significant value α; s is the estimated variance deviation for the scale adopted in data collection; and e is equal to an acceptable margin of error multiplied by the number of points on the primary scale. It is to be noted that s is the fraction of the inclusive range of the scale to the number of standard deviations that include almost all possible values in the range.

The commonly used 95% level of significance (Kamali and Hewage, 2017) was used, which corresponds to an α equal to 0.05. Hence, the corresponding value of t is 1.96. Since a 5-point Likert scale is adopted in this survey, s is commonly taken as 5/6 (Randiwela and Wijayaratne, 2017; Fellows and Liu, 2015). Moreover, e is computed by multiplying 5 by 0.05 where 5 is the number of points on the adopted scale and 0.05 is the margin of error. The margin of error is generally decided by the researchers, and it is commonly taken to be 5% (Pereira et al., 2018). As such, the minimum required number of respondents is computed using Equation (36) as 43. As such, it could be concluded that the total of 100 collected responses is considered sufficient because it exceeds 43 which is the required minimum number of respondents calculated using Equation (36).

6.5.3. The Impact of Offsite Construction on the Offsite Workforce. The measure of internal reliability (internal consistency) was assessed using the Cronbach's Alpha and the inter-rater agreement (external reliability) was assessed using the ICC. For the impact of offsite construction on the *skillset* of the offsite workforce occupations, the obtained Cronbach's Alpha is 0.952 with a 95% confidence interval of [0.938, 0.965], and the obtained ICC is 0.909 with a 95% confidence interval of [0.84, 0.96]. As for the impact of offsite construction on the *demand* for the offsite workforce occupations, the obtained Cronbach's Alpha is 0.958 with a 95% confidence interval of [0.945, 0.969], and the obtained ICC is 0.913 with a 95% confidence interval of [0.85, 0.96]. Since all obtained Cronbach's Alpha coefficients are higher than 0.75, it could be concluded that the established survey is valid and reliable (Christmann and Aelstb, 2006). Also, since all obtained ICC are greater than 0.75, it could be concluded that there is an excellent agreement and consistency between the 100 responses obtained from the survey (Cicchetti

and Sparrow, 1981). After checking the reliability and validity of this part of the survey, the average impact on skillset and the average impact on demand was calculated for each offsite workforce occupation. The obtained results for the offsite construction workforce are shown in Table 6.2.

Table 6.2 Quantified impacts on the offsite construction workforce.

Offsite workforce occupations	Impact on skillset			Impact on demand		
	Average score for impact on skillset	Average impact on skillset (in %)	Nature of impact on skillset	Average score for impact on demand	Average impact on demand (in %)	Nature of change in demand
Assembly, fabrication, and production personnel	1.84	9.2%	Upskill	1.79	8.95%	Grow
Equipment and machine operations personnel and technicians	1.55	7.75%	Upskill	1.34	6.7%	Grow
Material handling and warehouse management personnel	1.4	7.0%	Upskill	1.14	5.7%	Grow
Planners, expeditors, facilitators, sequence management, and supply chain personnel	1.82	9.1%	Upskill	1.78	8.9%	Grow
Start-up, testing, and commissioning personnel	1.49	7.45%	Upskill	1.27	6.35%	Grow
Logistics and transportation management personnel	1.67	8.35%	Upskill	1.78	8.9%	Grow

Table 6.2 Quantified impacts on the offsite construction workforce. (Continued).

Engineering personnel (industrial, mechanical, electrical, manufacturing, systems, etc.)	1.76	8.8%	Upskill	1.49	7.45%	Grow
Quality assurance, quality control, and reliability personnel	1.58	7.9%	Upskill	1.57	7.85%	Grow
Maintenance, programming, and troubleshooting personnel	1.39	6.95%	Upskill	1.0	5.0%	Grow
Safety personnel	1.11	5.55%	Upskill	0.73	3.65%	Grow
Procurement and contract management personnel	1.44	7.2%	Upskill	1.21	6.05%	Grow
Instrumentation and controls personnel	1.74	8.7%	Upskill	1.49	7.45%	Grow
Heavy lifting, rigging, and signal personnel	1.45	7.25%	Upskill	1.34	6.7%	Grow
Technology and configuration specialists	1.72	8.6%	Upskill	1.68	8.4%	Grow
Detailers	1.05	5.25%	Upskill	1.08	5.4%	Grow
Specification writers	1.26	6.3%	Upskill	0.92	4.6%	Grow
Computer-aided manufacturing (CAM) and information modeling professionals	2.12	11.2%	Upskill	2.0	10.0%	Grow
Truck drivers	0.76	3.8%	Upskill	0.86	4.3%	Grow

Furthermore, the overall impact of offsite construction on each one of the offsite workforce occupations was calculated using Equation (34). The obtained results and the associated ranks of the offsite construction occupations are shown in Table 6.3.

Table 6.3 Overall impact and rank for the offsite construction workforce.

Offsite workforce occupations	Overall impact on skillset[a]	Rank
Computer-aided manufacturing (CAM) and information modeling professionals	5.41	1
Assembly, fabrication, and production personnel	4.54	2
Planners, expeditors, facilitators, sequence management, and supply chain personnel	4.52	3
Technology and configuration specialists	4.18	4
Instrumentation and controls personnel	3.93	5
Logistics and transportation management personnel	3.76	6
Quality assurance, quality control, and reliability personnel	3.44	7
Engineering personnel (industrial, mechanical, electrical, manufacturing, systems, etc.)	3.35	8
Equipment and machine operations personnel and technicians	3.21	9
Procurement and contract management personnel	3	10
Heavy lifting, rigging, and signal personnel	2.99	11
Start-up, testing, and commissioning personnel	2.97	12
Material handling and warehouse management personnel	2.79	13
Maintenance, programming, and troubleshooting personnel	2.46	14
Detailers	1.74	15
Specification writers	1.74	16
Safety personnel	1.72	17
Truck drivers	1.08	18

[a]Calculated using Equation (34)

6.5.4. The Impact of Offsite Construction on the Onsite Workforce. For the impact of offsite construction on the *skillset* of the onsite workforce occupations, the obtained Cronbach's Alpha is 0.947 with a 95% confidence interval of [0.93, 0.961], and the obtained ICC is 0.962 with a 95% confidence interval of [0.94, 0.98]. As for the impact of offsite construction on the *demand* for the onsite workforce occupations, the obtained Cronbach's Alpha is 0.954 with a 95% confidence interval of [0.94, 0.966], and the obtained ICC is 0.968 with a 95% confidence interval of [0.95, 0.98]. Since all obtained Cronbach's Alpha coefficients are higher than 0.75, it could be concluded that the

established survey is valid and reliable (Christmann and Aelstb, 2006). Also, since all obtained ICC are greater than 0.75, it could be concluded that there is an excellent agreement and consistency between the 100 responses obtained from the survey (Cicchetti and Sparrow, 1981).

After checking the reliability and validity of this part of the survey, the average impact on skillset and the average impact on demand was calculated for each onsite workforce occupation. The obtained results for the onsite construction workforce are shown in Table 6.4.

Table 6.4 Quantified impacts on the onsite construction workforce.

Onsite workforce occupations	Impact on skillset			Impact on demand		
	Average score for impact on skillset	Average impact on skillset (in %)	Nature of impact on skillset	Average score for impact on demand	Average impact on demand (in %)	Nature of change in demand
Boilermakers	0.31	1.55%	Upskill	-0.66	-3.3%	Shrink
Carpenters	0.4	2.0%	Upskill	-0.58	-2.9%	Shrink
Concrete, brick, block, stone, and plastering personnel	0.26	1.3%	Upskill	-0.56	-2.8%	Shrink
Drywall personnel	0.1	0.5%	Upskill	-0.82	-4.1%	Shrink
Electrical personnel	1.32	6.6%	Upskill	-0.01	-0.05%	Shrink
Equipment operators	0.95	4.75%	Upskill	0.25	1.25%	Grow
Floor layers/installers/setters	0.17	0.85%	Upskill	-0.42	-2.1%	Shrink
General Laborers/Helpers	0.41	2.05%	Upskill	-0.67	-3.35%	Shrink
Glaziers	0.16	0.8%	Upskill	-0.52	-2.6%	Shrink
Heavy civil personnel (earthwork, utilities, highway, etc.)	0.56	2.8%	Upskill	0.13	0.65%	Grow

Table 6.4 Quantified impacts on the onsite construction workforce. (Continued).

Offsite modules/components installation and set-up personnel	2.33	13.3%	Upskill	2.26	12.6%	Grow
Instrumentation and control personnel	1.51	7.55%	Upskill	0.88	4.4%	Grow
Insulation personnel	0.46	2.3%	Upskill	-0.41	-2.05%	Shrink
Ironworkers	0.72	3.6%	Upskill	-0.43	-2.15%	Shrink
Lifting, cranes, hoisting, rigging, and signal personnel	1.46	7.3%	Upskill	1.06	5.3%	Grow
Mechanical personnel	1.26	6.3%	Upskill	-0.01	-0.05%	Shrink
Millwrights	0.74	3.7%	Upskill	-0.12	-0.6%	Shrink
Painters	0.19	0.95%	Upskill	-0.82	-4.1%	Shrink
Pipefitters, pipelayers, and steamfitters	0.95	4.75%	Upskill	-0.55	-2.75%	Shrink
Plumbing personnel	0.66	3.3%	Upskill	-0.5	-2.5%	Shrink
Roofers and waterproofers	0.36	1.8%	Upskill	-0.32	-1.6%	Shrink
Scaffold builders	0.39	1.95%	Upskill	-0.57	-2.85%	Shrink
Sheet Metal	0.62	3.1%	Upskill	-0.53	-2.65%	Shrink
Welders	1.29	6.45%	Upskill	-0.51	-2.55%	Shrink

Furthermore, the overall impact of offsite construction on each one of the onsite workforce occupations was calculated using Equation (34). The obtained results and the associated ranks of the onsite construction occupations are shown in Table 6.5.

Table 6.5 Overall impact and rank for the onsite construction workforce.

Onsite workforce occupations	Overall impact on skillset[a] (absolute value)	Rank
Offsite modules/components installation and set-up personnel	6.52	1
Lifting, cranes, hoisting, rigging, and signal personnel	2.67	2
Instrumentation and control personnel	2.65	3
Electrical personnel	1.22	4

Table 6.5 Overall impact and rank for the onsite construction workforce. (Continued).

Equipment operators	0.85	5
Millwrights	0.66	6
Mechanical personnel	0.65	7
Scaffold builders	0.57	8
Boilermakers	0.55	9
Insulation personnel	0.54	10
Carpenters	0.49	11
Heavy civil personnel (earthwork, utilities, highway, etc.)	0.48	12
General Laborers/Helpers	0.38	13
Pipefitters, pipelayers, and steamfitters	0.35	14
Ironworkers	0.19	15
Glaziers	0.17	16
Welders	0.12	17
Concrete, brick, block, stone, and plastering personnel	0.1	18
Painters	0.09	19
Drywall personnel	0.06	20
Floor layers/installers/setters	0.05	21
Sheet Metal	0.05	22
Roofers and waterproofers	0.01	23
Plumbing personnel	0.01	24

[a]Calculated using Equation (34)

As shown in Table 6.5, the top 5 onsite construction occupations that are most impacted by the increased use of offsite construction include: (1) offsite modules/components installation and set-up personnel; (2) lifting, cranes, hoisting, rigging, and signal personnel; (3) instrumentation and control personnel; (4) electrical personnel; and (5) equipment operators.

6.5.5. The Impact of Offsite Construction on Labor Characteristics. For the labor characteristics, the obtained Cronbach's Alpha is 0.881 with a 95% confidence interval of [0.845, 0.912], and the obtained ICC is 0.971 with a 95% confidence interval of [0.96, 0.98]. Since all obtained Cronbach's Alpha coefficients are higher than 0.75, it could

be concluded that the established survey is valid and reliable (Christmann and Aelstb, 2006). Also, since all obtained ICC are greater than 0.75, it could be concluded that there is an excellent agreement and consistency between the 100 responses obtained from the survey (Cicchetti and Sparrow, 1981). After checking the reliability and validity of this part of the survey, the average score and percentage impact was calculated for each labor characteristic. The obtained results are shown in Table 6.6.

Table 6.6 Impact of offsite construction on the labor characteristics.

Aspect	Labor characteristics	Average score impact	Average percentage impact	Nature of impact
Workforce diversity	Opportunities for women	2.1	11.0%	Increase
	Opportunities for minorities (women not included) and diverse ethnic groups	1.92	9.6%	Increase
	Opportunities for international/immigrant workforce	1.58	7.9%	Increase
	Opportunities for workforce with disabilities	1.22	6.1%	Increase
	Opportunities for converting the workforce from non-construction industries to the construction industry	1.31	6.55%	Increase
	Opportunities for veterans or active military members	1.62	8.1%	Increase
Workforce generations	Opportunities for 'silent generation' or 'traditionalists'	-1.29	-6.45%	Decrease
	Opportunities for 'baby boomers'	-0.68	-3.4%	Decrease
	Opportunities for 'generation X'	0.33	1.65%	Increase
	Opportunities for 'generation Y' or 'Millennials'	1.79	8.95%	Increase
	Opportunities for 'generation Z'	2.4	14.0%	Increase
	Opportunities for 'generation alpha'	2.21	12.1%	Increase
Workforce membership	Opportunities for union labor	0.07	0.35%	Increase
	Opportunities for non-union labor	1.77	8.85%	Increase

Table 6.6 Impact of offsite construction on the labor characteristics. (Continued).

Workforce education level	Opportunities for diploma holders	1.82	9.1%	Increase
	Opportunities for associate degree holders	1.67	8.35%	Increase
	Opportunities for bachelor's degree holders	1.68	8.4%	Increase
	Opportunities for master's degree holders	1.38	6.9%	Increase
	Opportunities for professional degree holders	1.23	6.15%	Increase
	Opportunities for doctoral (or Ph.D) degree holders	0.59	2.95%	Increase
Workforce attributes	Workforce productivity	1.84	9.2%	Increase
	Workforce rework (man-hours attributed to rework)	-0.48	-2.4%	Decrease
	Idle (stand-by) workforce (paid man-hours not spent working)	-0.75	-3.75%	Decrease
	Labor disputes	-0.18	-0.9%	Decrease
	Workplace theft and fraud	-0.36	-1.8%	Decrease
	Workforce (or staffing) turnover	-0.08	-0.4%	Decrease
	Workforce mobility, transfer, or relocation from one geographical area to the other	-0.03	-0.15%	Decrease
	Workforce overtime (man-hours attributed to overtime)	0.26	1.3%	Increase
	Quality of the working conditions	1.77	8.85%	Increase
	Workforce health and safety	1.94	9.7%	Increase
	Workforce absenteeism	-0.19	-0.95%	Decrease
	Workforce fatigue	-0.4	-2.0%	Decrease
	Age of retirement	1.24	6.2%	Increase
	Crew or team size	-0.4	-2.0%	Decrease
	Job security	0.48	2.4%	Increase
	Workforce's quality of work	1.28	6.4%	Increase
	Workforce compensation (overall package: pay and benefits)	0.66	3.3%	Increase
	Length of workforce career path progression/development	0.57	2.85%	Increase
	Travel and per diem rate	0.46	2.3%	Increase
	Workforce learning rate	0.89	4.45%	Increase
	Workforce smartness, adaptability, and flexibility	1.1	5.5%	Increase
	Cost of workforce training and development	1.14	5.7%	Increase
	Shifting work hours to international low-cost labor locations	1.63	8.15%	Increase

As shown in Table 6.6, some of the labor characteristics are perceived to increase as a result of the use of offsite construction while other labor characteristics are perceived to decrease.

6.5.6. The Impact of Offsite Construction on the Engineering Workforce. This subsection provides the results and associated analysis in relation to the collected data for the impact of offsite construction on the technical and managerial skills of the engineering workforce.

6.5.6.1. Impact on technical skillset. The measure of internal reliability (internal consistency) was assessed using the Cronbach's Alpha and the inter-rater agreement (external reliability) was assessed using the ICC. For the impact on the technical skillset of the engineering workforce, the obtained Cronbach's Alpha is 0.927 with a 95% confidence interval of [0.905, 0.946], and the obtained ICC is 0.925 with a 95% confidence interval of [0.87, 0.97]. Since the obtained Cronbach's Alpha coefficient is higher than 0.75, it could be concluded that this part of the survey is valid and reliable (Christmann and Aelstb, 2006). Also, since the obtained ICC is greater than 0.75, it could be concluded that there is an excellent agreement and consistency between the 100 surveyed industry practitioners (Cicchetti and Sparrow, 1981) on the impact of offsite construction on the technical skillset of the engineering workforce. The obtained results are shown in Table 6.7.

As shown in Table 6.7, the technical skillset of many engineering occupations is considerably impacted by offsite construction. The top 5 impacted occupations in terms of technical skillset include: (1) BIM and 3D/4D/nD information modeling and rendering; (2) mechanical engineering; (3) electrical engineers; (4) quality assurance and quality control engineers; and (5) planning engineers.

Table 6.7 Results for the impact of offsite construction on the technical skillset for the engineering workforce.

Engineering occupations	Mean	Stan. dev.	Rank
Project principals/directors	3.13	0.84	14
Structural engineers	3.4	0.84	8
Quantity surveying engineers	3.27	0.87	12
Cost estimation engineers	3.39	0.84	9
Planning engineers	3.47	0.82	5
Geotechnical engineers	2.9	1.0	17
Electrical engineers	3.55	0.88	3
Mechanical engineers	3.58	0.79	2
Plumbing engineers	3.22	0.89	13
Civil engineers	3.06	0.96	15
Health and safety engineers	3.04	0.88	16
Quality assurance and quality control engineers	3.49	0.78	4
Specialty engineers	3.43	0.88	6
BIM and 3D/4D/nD information modeling and rendering	3.81	0.85	1
Detailing professionals	3.41	0.98	7
Specification writers	3.38	0.93	10
Land surveying and landscape professionals	2.81	1.08	18
Architects	3.28	1.01	11

6.5.6.2. Impact on managerial skillset. For the impact on the engineering workforce's managerial skill set, the obtained Cronbach's Alpha is 0.951 with a 95% confidence interval of [0.936, 0.964], and the obtained ICC 0.939 with a 95% confidence interval of [0.89, 0.97]. Since the obtained Cronbach's Alpha coefficient is higher than 0.75, it could be concluded that this part of the survey is valid and reliable (Christmann and Aelstb, 2006). Also, since the obtained ICC is greater than 0.75, it could be concluded that there is an excellent agreement and consistency between the 100 surveyed industry practitioners (Cicchetti and Sparrow, 1981) on the impact of offsite construction on the managerial skillset of the engineering workforce. The obtained results are shown in Table 6.8.

Table 6.8 Results for the impact of offsite construction on the managerial skillset for the engineering workforce.

Engineering occupations	Mean	Stan. dev.	Rank
Project principals/directors	3.48	0.87	1
Structural engineers	3.07	0.88	9
Quantity surveying engineers	2.98	0.96	14
Cost estimation engineers	3.23	1.0	7
Planning engineers	3.41	1.01	4
Geotechnical engineers	2.61	0.96	17
Electrical engineers	3.2	0.9	8
Mechanical engineers	3.27	0.92	5
Plumbing engineers	2.87	0.88	15
Civil engineers	2.85	0.96	16
Health and safety engineers	2.99	0.95	11
Quality assurance and quality control engineers	3.42	0.92	3
Specialty engineers (e.g. fire protection, process, operations, industrial, hydraulic, etc.)	3.24	0.98	6
BIM and 3D/4D/nD information modeling and rendering	3.45	1.03	2
Detailing professionals	2.98	0.95	13
Specification writers	2.98	0.94	12
Land surveying and landscape professionals	2.58	1.07	18
Architects	3.04	1.04	10

As shown in Table 6.8, the managerial skillset of many engineering occupations is considerably impacted by offsite construction but to a lower extent as compared to the impact on the technical skillset shown in Table 6.7. The top 5 impacted occupations in terms of managerial skillset include: (1) project principals/directors; (2) BIM and 3D/4D/nD information modeling and rendering; (3) quality assurance and quality control engineers; (4) planning engineers; and (5) mechanical engineers.

6.5.6.3. Prioritization of the impacted occupations. The prioritization of the impacted occupations was performed using *k*-means clustering, and the optimal number of

clusters was determined using the silhouette score. The obtained results are shown in Figure 6.4.

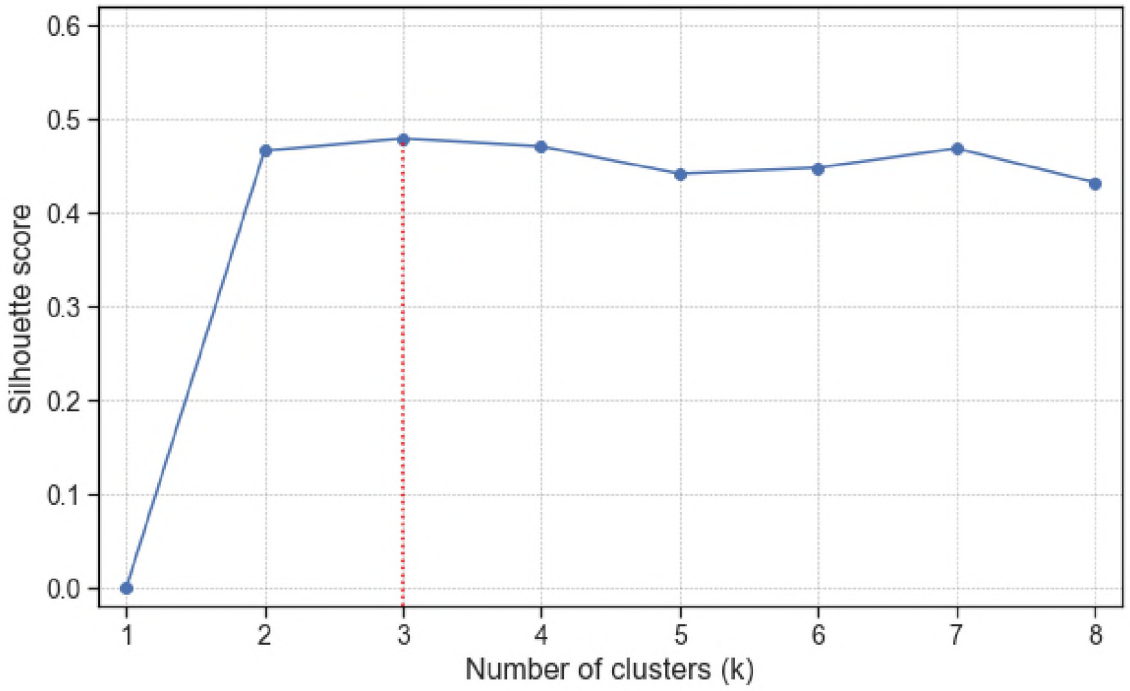

Figure 6.4 Determining the optimal number of clusters.

As shown in Figure 6.4, the highest silhouette score is for $k = 3$, thus 3 clusters are considered for the *k*-means clustering algorithm. Once the optimal number of clusters is determined, the *k*-means clustering algorithm was used to prioritize the impacted occupations where the first cluster includes occupations with a high overall impact, the second cluster includes occupations with a medium overall impact, and the third cluster includes occupations with a low overall impact. The obtained results are shown in Figure 6.5.

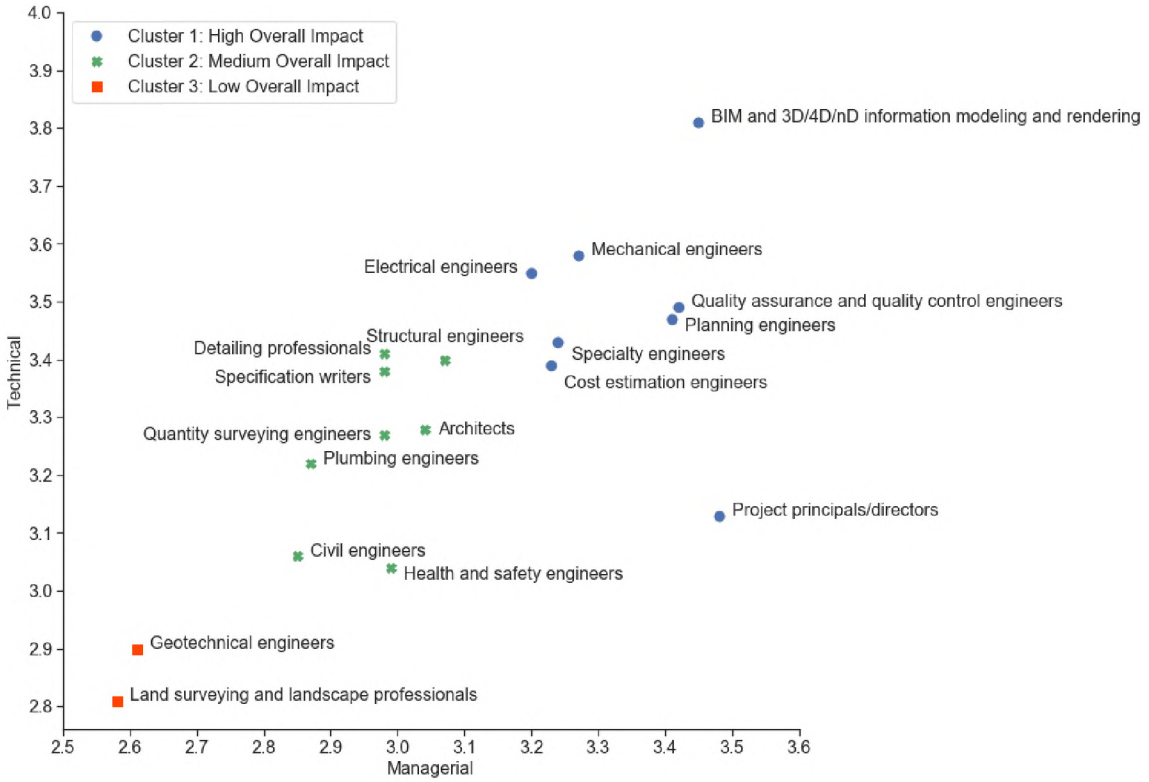

Figure 6.5 Obtained clustering results for the prioritization of the engineering workforce occupations.

As shown in Figure 6.5, construction companies shall invest considerable resources in training and developing the following prioritized engineering occupations in relation to the offsite construction skillset: BIM and 3D/4D/nD information modeling and rendering, mechanical engineers, quality assurance and quality control engineers, planning engineers, electrical engineers, specialty engineers, cost estimation engineers, and project principals/directors.

6.5.7. The Impact of Offsite Construction on the Construction Workforce. This subsection provides the results and associated analysis in relation to the collected data for the impact of offsite construction on the technical and managerial skills of the construction workforce.

6.5.7.1. Impact on technical skillset. For the impact on the construction workforce's technical skills, the obtained Cronbach's Alpha is 0.933 with a 95% confidence interval of [0.912, 0.951], and the obtained ICC is 0.928 with a 95% confidence interval of [0.87, 0.97]. Since the obtained Cronbach's Alpha coefficient is higher than 0.75, it could be concluded that this part of the survey is valid and reliable. Also, since the obtained ICC is greater than 0.75, it could be concluded that there is an excellent agreement and consistency between the 100 surveyed industry practitioners. The obtained results are shown in Table 6.9.

Table 6.9 Results for the impact of offsite construction on the technical skillset for the construction workforce.

Construction occupations	Mean	Stan. dev.	Rank
Project management professionals	3.36	0.8	12
Procurement professionals	3.41	0.78	10
Expeditors	3.08	0.91	15
Logistics professionals	3.46	0.89	8
Construction managers	3.56	0.86	4
Planning and control professionals	3.49	0.86	6
Safety professionals	3.13	0.97	14
Quality assurance and quality control professionals	3.47	0.88	7
Materials management professionals	3.42	0.96	9
Field coordination and interface management personnel	3.56	0.81	3
Superintendents	3.4	0.83	11
Foremen	3.33	0.88	13
Waste management personnel	2.69	0.94	16
Start-up, testing, and commissioning personnel	3.57	0.84	2
Constructability professionals	3.62	0.84	1
Specialized lifts and heavy haul professionals	3.56	0.9	5

As shown in Table 6.9, the technical skillset of many construction occupations is considerably impacted by offsite construction where the top 5 impacted occupations

include: (1) constructability professionals; (2) start-up, testing, and commissioning personnel; (3) field coordination and interface management personnel; (4) construction managers; and (5) specialized lifts and heavy haul professionals.

6.5.7.2. Impact on managerial skillset. For the impact on the construction workforce's managerial skills, the obtained Cronbach's Alpha is 0.942 with a 95% confidence interval of [0.924, 0.958], and the obtained ICC is 0.936 with a 95% confidence interval of [0.88, 0.97]. Since the obtained Cronbach's Alpha coefficient is higher than 0.75, it could be concluded that this part of the survey is valid and reliable. Since the obtained ICC is greater than 0.75, it could be concluded that there is an excellent agreement/consistency between the 100 surveyed industry practitioners. The obtained results are shown in Table 6.10.

Table 6.10 Results for the impact of offsite construction on the managerial skillset for the construction workforce.

Construction occupations	Mean	Stan. dev.	Rank
Project management professionals	3.66	1.02	2
Procurement professionals	3.48	0.93	7
Expeditors	3.07	0.98	15
Logistics professionals	3.44	0.97	8
Construction managers	3.69	1.0	1
Planning and control professionals	3.53	0.95	5
Safety professionals	3.15	1.02	14
Quality assurance and quality control professionals	3.42	0.99	10
Materials management professionals	3.38	0.98	11
Field coordination and interface management personnel	3.59	0.92	3
Superintendents	3.43	1.06	9
Foremen	3.27	1.0	13
Waste management personnel	2.58	1.0	16
Start-up, testing, and commissioning personnel	3.53	0.97	6
Constructability professionals	3.58	0.96	4
Specialized lifts and heavy haul professionals	3.3	1.01	12

As shown in Table 6.10, the managerial skillset of many engineering occupations is considerably impacted by offsite construction. The top 5 impacted occupations in terms of managerial skillset include: (1) construction managers; (2) project management professionals; (3) field coordination and interface management personnel; (4) constructability professionals; and (5) planning and control professionals.

6.5.7.3. Prioritization of the impacted occupations. Similar to the prioritization of the engineering workforce, the *k*-means clustering algorithm was used to prioritize the construction workforce occupations. The obtained results are shown in Figure 6.6.

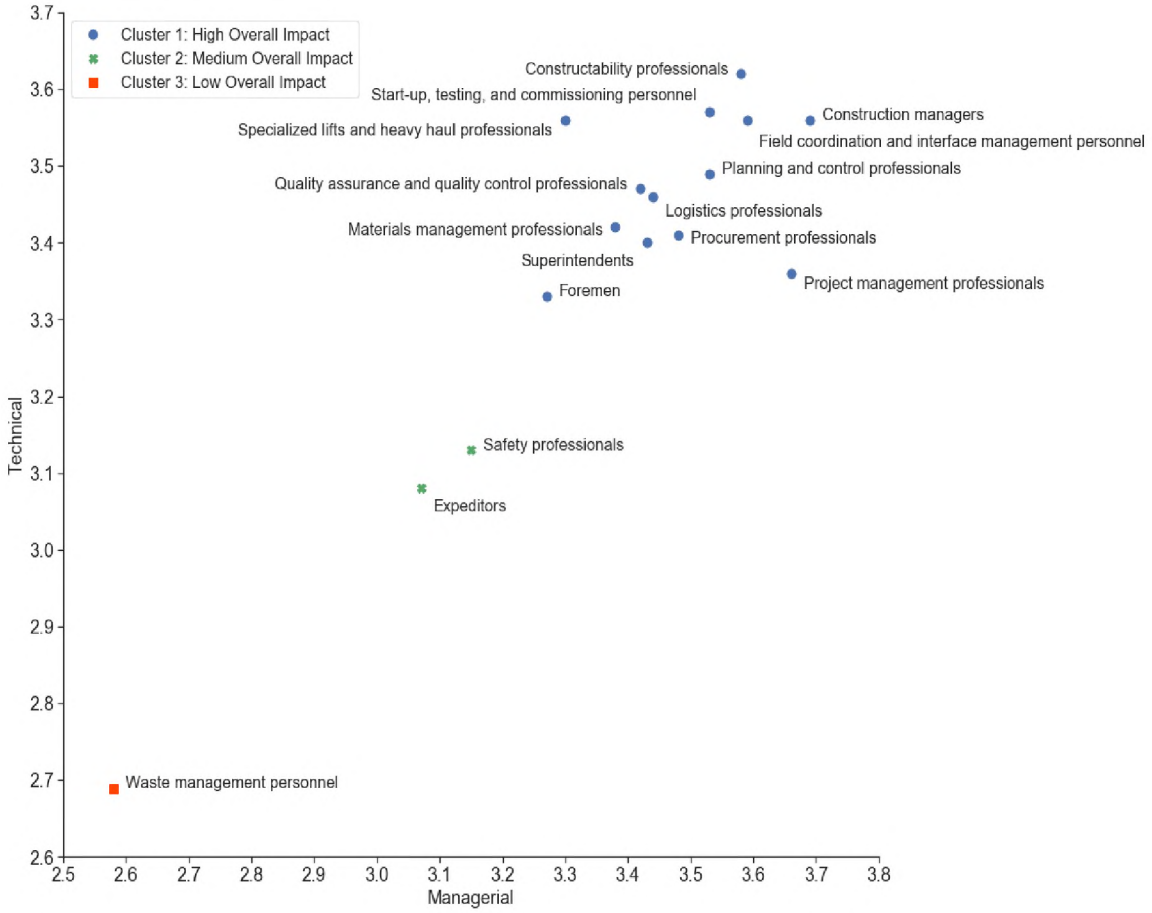

Figure 6.6 Obtained clustering results for the prioritization of the construction workforce occupations.

Figure 6.6 reflects that construction companies shall invest considerable resources in training and developing the following prioritized construction occupations in relation to the offsite construction skillset: construction managers; field coordination and interface management personnel; constructability professionals; start-up, testing, and commissioning personnel; planning and control professionals; logistics professionals; procurement professionals; project management professionals; specialized lifts and heavy haul professionals; quality assurance and quality control professionals; materials management professionals; superintendents; and foremen.

6.5.8. The Impact of Offsite Construction on the Administrative Workforce. This subsection provides the results and associated analysis in relation to the collected data for the impact of offsite construction on the technical and managerial skills of the administrative workforce.

6.5.8.1. Impact on technical skillset. For the impact on the technical skillset of the administrative workforce, the obtained Cronbach's Alpha is 0.946 with a 95% confidence interval of [0.929, 0.96], and the obtained ICC is 0.936 with a 95% confidence interval of [0.88, 0.97].

Since the obtained Cronbach's Alpha coefficient is higher than 0.75, it could be concluded that this part of the survey is valid and reliable (Christmann and Aelstb, 2006). Also, since the obtained ICC is greater than 0.75, it could be concluded that there is an excellent agreement and consistency between the 100 surveyed industry practitioners (Cicchetti and Sparrow, 1981) on the impact of offsite construction on the technical skillset of the administrative workforce. The obtained results are shown in Table 6.11.

Table 6.11 Results for the impact of offsite construction on the technical skillset for the administrative workforce.

Administrative occupations	Mean	Stan. dev.	Rank
Project executive professionals	3.04	0.95	9
Legal and contract professionals	3.02	0.99	10
Environmental, green, and sustainability professionals	3.15	0.99	6
Project finance professionals	2.92	0.92	13
Permitting and regulation professionals	3.14	0.92	7
Risk management professionals	3.25	0.88	4
Sales, operations, technology, and construction (SOTC) professionals	3.01	0.87	11
Owner's representative	3.22	0.82	5
Sub-contractor(s) administration personnel	2.92	0.9	12
Marketing and business development professionals	2.83	0.97	14
Human resources (HR) professionals	2.62	0.95	16
Project controls professionals	3.39	0.91	3
Computer and information technology (IT) professionals	3.53	0.99	1
Insurance professionals	2.77	0.91	15
Advanced work packaging (AWP) professionals	3.52	1.01	2
Document control professionals	3.08	0.99	8

As shown in Table 6.11, the technical skillset of many administrative occupations is considerably impacted by offsite construction. The top 5 impacted occupations in terms of technical skillset include: (1) computer and information technology (IT) professionals; (2) advanced work packaging (AWP) professionals; (3) project controls professionals; (4) risk management professionals; and (5) owner's representative.

6.5.8.2. Impact on managerial skillset. For the impact on the administrative workforce's managerial skills, the obtained Cronbach's Alpha is 0.952 with a 95% confidence interval of [0.936, 0.964], and the obtained ICC is 0.883 with a 95% confidence interval of [0.78, 0.95].

Since the obtained Cronbach's Alpha coefficient is higher than 0.75, it could be concluded that this part of the survey is valid and reliable (Christmann and Aelstb, 2006).

Since the obtained ICC is greater than 0.75, it could be concluded that there is an excellent agreement/consistency between the 100 surveyed industry practitioners (Cicchetti and Sparrow, 1981) on the impact of offsite construction on the managerial skillset of the administrative workforce. The obtained results are shown in Table 6.12.

As shown in Table 6.12, the managerial skillset of many administrative occupations is considerably impacted by offsite construction. The top 5 impacted occupations in terms of managerial skillset include: (1) advanced work packaging (AWP) professionals; (2) owner's representative; (3) project executive professionals; (4) project controls professionals; and (5) risk management professionals.

Table 6.12 Results for the impact of offsite construction on the managerial skillset for the administrative workforce.

Administrative occupations	Mean	Stan. dev.	Rank
Project executive professionals	3.4	0.96	3
Legal and contract professionals	3.2	1.05	7
Environmental, green, and sustainability professionals	3.16	1.02	8
Project finance professionals	3.05	1.01	11
Permitting and regulation professionals	3.12	1.07	9
Risk management professionals	3.27	0.97	5
Sales, operations, technology, and construction (SOTC) professionals	2.99	0.96	13
Owner's representative	3.41	0.96	2
Sub-contractor(s) administration personnel	3.06	0.98	10
Marketing and business development professionals	2.83	1.02	16
Human resources (HR) professionals	2.86	1.09	15
Project controls professionals	3.36	1.01	4
Computer and information technology (IT) professionals	3.24	1.09	6
Insurance professionals	2.88	1.06	14
Advanced work packaging (AWP) professionals	3.43	1.02	1
Document control professionals	3.01	1.07	12

6.5.8.3. Prioritization of the impacted occupations. Similar to the prioritization of the engineering and construction workforce, the *k*-means clustering algorithm was used to prioritize the administrative workforce occupations where the first cluster includes occupations with a high overall impact, the second cluster includes occupations with a medium overall impact, and the third cluster includes occupations with a low overall impact. The obtained results are shown in Figure 6.7.

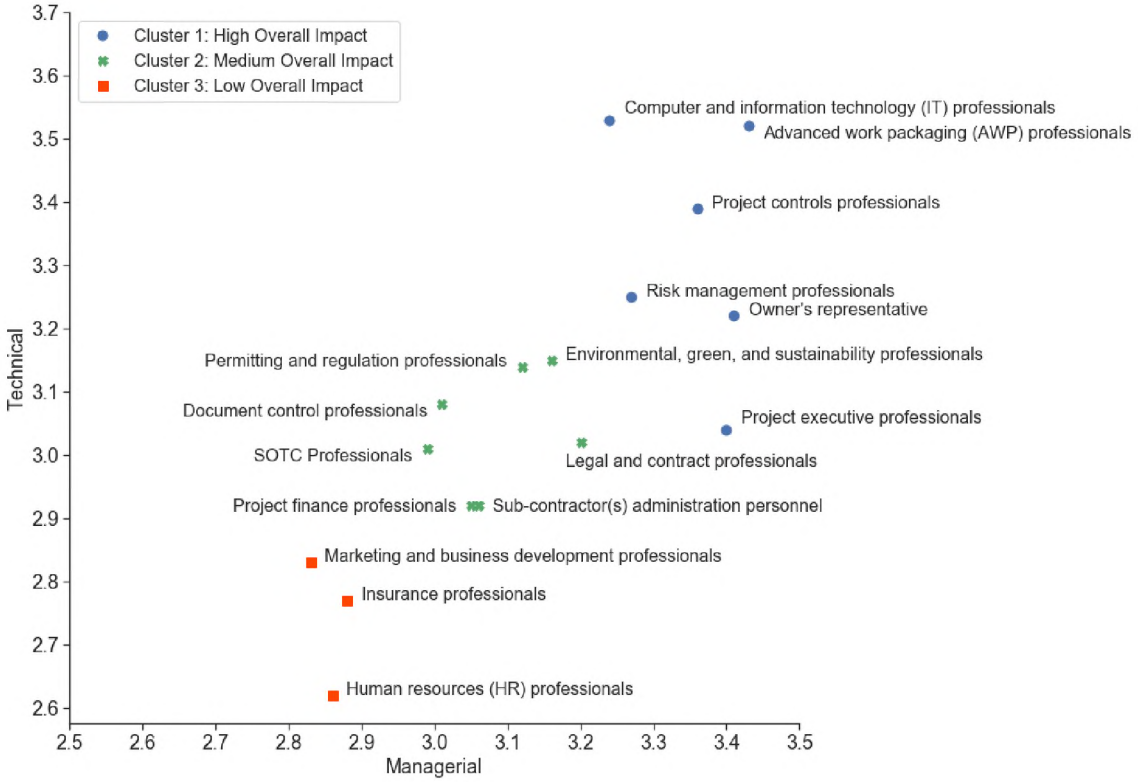

Figure 6.7 Obtained clustering results for the prioritization of the administrative workforce occupations.

Figure 6.7 reflects that construction companies shall invest considerable resources in training and developing the following prioritized occupations in relation to the offsite

construction skillset: construction managers; field coordination and interface management personnel; constructability professionals; start-up, testing, and commissioning personnel; planning and control professionals; logistics professionals; procurement professionals; project management professionals; specialized lifts and heavy haul professionals; quality assurance and quality control professionals; materials management professionals; superintendents; and foremen.

6.6. SUMMARY

This section of the dissertation (1) quantified the impacts of offsite construction on the skill set of onsite and offsite workforce; (2) evaluated the impacts of offsite construction on the demand for the onsite and offsite workforce; (3) assessed the impacts of offsite construction on the labor characteristics; (4) quantified the impacts of offsite construction on the technical skill set of the identified engineering, construction, and administrative workforce occupations; (5) evaluated the impacts of offsite construction on the managerial skill set for each one of the engineering, construction, and administrative workforce categories; and (6) prioritized the impactions of offsite construction on the overall engineering, construction, and administrative workforce skillset.

6.7. CREDIT

The research reported in this section of the dissertation (i.e., Section 6) was a collaboratively work carried through funding provided by the Construction Industry Institute (CII) at Missouri University of Science and Technology, Purdue University, and the University of Arkansas under CII RT-371. To this end, the collaborators (Islam El-

adaway and Rayan Hassane Assaad from Missouri University of Science and Technology, Makarand Hastak from Purdue University, and Kim LaScola Needy from the University of Arkansas) are very thankful for the financial and logistical support provided by CII. Also, the collaborators deeply appreciate and value the efforts and the insightful guidance provided by the CII RT-371 industry team members who offered a wealth of knowledge and real-world experience with exceptional dedication. It is worth noting that the research present in this section of the dissertation does not include the entire research efforts performed by the collaborators under CII RT-371, but rather it reports a part of the full research efforts conducted by CII RT-371.

6.8. RELATED APPENDIX

Appendix E presents the expert-based survey's part that is related to the scope of this section of the dissertation. It is worth mentioning that the entire survey that was developed by the RT-371 team included other questions that are not present in this appendix because they were not part of the focus of this section of the dissertation. These other questions are related to assessing the current state of practice of offsite construction and its future prospects as well as studying the labor productivity factors in offsite construction projects. In addition, the RT-371 team conducted semi-structured interviews with offsite construction industry experts and collected business case studies from leading national and international offsite construction companies. However, the collected information is not present or discussed in this section of the dissertation since it was not part of the scope and focus of this section of the dissertation.

7. CONCLUSION

7.1. RESEARCH SUMMARY

Modeling techniques and computational analytics have showed great potential in solving critical challenges that are facing the aging infrastructure as well as the engineering, construction, and technology industry. In relation to that, this dissertation focused on providing and developing innovative models, computational methods, and approaches to help in solving the following three critical issues faced in the industry: (1) safety hazard potential of dam infrastructures; (2) engineering and construction operations including project performance (schedule and time), construction productivity, and fatalities; (3) and leveraging of offsite construction methods and technologies through proper training and prioritization of workforce development programs to the difference workforce occupations. In relation to that, the research summary of this dissertations is provided in the below paragraphs.

First, because visual formal inspection requires a huge effort, long time, high cost, and even substantial danger, the recent availability of dam data opens unprecedented opportunities to artificial intelligence algorithms and data analytics for better evaluation and prediction of the hazard potential level of dams. In relation to that, this dissertation reflected that artificial intelligence methods have promising capabilities in devising a data-driven framework or decision support tool that is valuable for dam owners and authorities to evaluate and predict the hazard potential level of their dams with good accuracy while minimizing the effort, time, and costs associated with formal inspection of the dams. The provided conclusions also offered insights on the factors affecting the hazard potential of

dams in the US. This empowers the associated agencies and authorities to direct their efforts toward ameliorating these critical factors for current dams, if possible, or ultimately for future dams. As such, this will help in fostering a proper and more effective distribution and allocation of funds for the management and repair of dams. Further, this dissertation provided that because no considerable cost is generally associated with the use of artificial intelligence methods and because considerable time savings are entertained as well, agencies and authorities responsible for the management of dams are recommended to utilize such methods as a first recourse for the prediction of hazard potential of their dams as well as the conditions of their infrastructure systems.

Second, this dissertation presented a holistic model to predict project performance. The model forecasts two prevailing indicators: cost and schedule overruns. The developed model allows project administrators to predict the performance of construction projects based on 25 project risks that are retrieved from the literature and that have shown to be the most important based on a meta-analysis of the literature. The model presents explicit equations that could be used to forecast the project performance in terms of cost and schedule and a set of graphs that could be adopted as an alternative method for estimating project performance to minimize the computational efforts. In addition, using numerical examples of a hypothetical case study, the proposed model reflected its practicality and reliability in predicting project performance. Further, this dissertation aims to trigger project stakeholders to include the considered 25 project risks, that affect the project performance in the construction industry, in their standard project assessment endeavors. Doing so will not only help improve the performance of projects but will enable gathering comprehensive data on the forecasted and actual overruns—an optimal method for fine-

tuning the model to better reflect construction projects. In addition, the developed model allows decision makers to take timely corrective actions and assess the impacts of such actions. Ultimately, doing so will assist in ensuring more effective and efficient delivery and execution of projects.

Third, based on the findings of this dissertation, it was concluded that (1) the new concept of gross construction productivity is statistically justified and should be implemented in the construction industry, because this dissertation reflected statistical significant causalities between the productivity of the construction industry and different dynamic workforce and workplace variables; (2) gross construction productivity is additional valuable information that construction companies should consider to make different insightful and well-educated industry-related decisions; (3) the health of the construction industry needs to be studied based on the productivity of the industry as a whole and not only based on labor productivity alone; and (4) the construction industry should move toward the development of a notion of a gross construction productivity indicator used to measure, evaluate, and predict the performance of the entire industry. Ultimately, this dissertation proposed a new indicator or index for gross construction productivity.

Fourth, this dissertation provided an additional dimension to the traditional analysis of safety accidents by examining possible associations and combinations between different causes of fatalities on construction jobsites. More specifically, this dissertation categorized the fatality causes into five main clusters and identified the main individual causes that lead to fatalities in the construction industry within each cluster as well as the critical combinations of causes that could lead to a fatal accident. In relation to that, the emphasis

was mainly on the combinations of causes that result in fatalities in the construction industry. It was concluded also that safety managers and supervisors shall focus on the identified relationships that can serve as an early warning such that if some of these causes are seen on the construction site, then a fatal accident is very likely to happen. That said, by focusing on the causes as combinations, rather than as simple events, and understanding the strength of their relationships, this dissertation provided a proactive approach that can help safety practitioners decide on countermeasures to address the combinations of causes. This would ultimately result in a more effective and efficient prevention of fatal accidents on construction sites. Furthermore, the results of this dissertation can be applied in the safety management practice in real construction sites through the development of data-driven accident causation systems. These systems are models that attribute accidents to many contributory factors, causes, and sub-causes by establishing that combinations of, and not only individual, factors give rise to accidents (Raouf, 2020).

Fifth, the findings of this dissertations showed that the skillsets for all offsite and onsite workforce occupations need to be upskilled rather than reskilled, the workforce market for the offsite construction workforce will prosper, and the demand for most of the onsite workforce occupations will decrease due to the expected increase in the use of offsite construction in the future. The findings also reflected that companies and industry practitioners shall prioritize their training programs and plans for the following offsite workforce occupations: (1) computer-aided manufacturing (CAM) and information modeling professionals; (2) assembly, fabrication, and production personnel; (3) planners, expeditors, facilitators, sequence management, and supply chain personnel; (4) technology and configuration specialists; and (5) instrumentation and controls personnel; as well as for

the following onsite workforce occupations: (1) offsite modules/components installation and set-up personnel; (2) lifting, cranes, hoisting, rigging, and signal personnel; (3) instrumentation and control personnel; (4) electrical personnel; and (5) equipment operators. Also, research outcomes showed that offsite construction could create opportunities and challenges alike for the workforce. On the opportunities side, offsite construction is expected to yield higher compensation, improved job security, reduced fatigue, fewer rework, greater productivity, higher learning rate, better working conditions, enhanced quality, and improved safety and health, among many others. However, on the challenges side, offsite construction could result in longer career path progression, higher cost of training and development, and displacement of local workers, among others. The findings showed that the skillsets of around 44% of the engineering workforce occupations, of around 81% of the construction workforce occupations, and of around 38% of the administrative workforce occupations are highly impacted by offsite construction. The outcomes also reflected that companies and industry practitioners shall prioritize their training programs and plans for the following (1) engineering workforce occupations: BIM and 3D/4D/nD information modeling and rendering, mechanical engineers, quality assurance and quality control engineers, planning engineers, electrical engineers, specialty engineers, cost estimation engineers, and project principals/directors; (2) construction workforce occupations: construction managers; field coordination and interface management personnel; constructability professionals; start-up, testing, and commissioning personnel; planning and control professionals; logistics professionals; procurement professionals; project management professionals; specialized lifts and heavy haul professionals; quality assurance and quality control professionals; materials

management professionals; superintendents; and foremen; and (3) administrative workforce occupations: computer and information technology (IT) professionals; advanced work packaging (AWP) professionals; project controls professionals; risk management professionals; owner's representative; and project executive professionals.

7.2. RESEARCH CONTRIBUTIONS

The research performed in this dissertation has many contributions and implications. In fact, this research is unique compared to previous related research with respect to focus, purpose, and methods. In relation to that, the contributions of this research could be either considered collectively or based on each individual section of the dissertation by its own. In other words, each section in this dissertation possesses its own recognizable and distinguishable intellectual and methodological merits (meaning, contributing to the infrastructure and construction management body of knowledge) as well as practical merits (meaning, pragmatic application in the industry). Also, all sections of the dissertation collectively share the all-encompassing or overarching contribution of providing innovative management and modeling approaches to help addressing the emerging challenges that are faced by the industry. To this end, the following paragraphs highlight the intellectual and methodological merits or contributions of each section of the dissertation as well as the practical industry merits/contributions.

Section 2 – Supervised Computational Artificial Intelligence Model for the Evaluation and Prediction of the Hazard Potential Level of Dam Infrastructures: The intellectual merits and methodological contributions of this section lie in the fact that this research is among the very few scholarly efforts (if any) that tried to leverage the recent

advancements in computational and machine learning methods for a better management and administration of dam infrastructures in the US, especially as related to their hazard potential level. In other words, most of the methodologies in the previous research efforts focused on one special structural or geotechnical behavioral aspect of dams while disregarding the important potential hazard aspect. As such, this section bridged this knowledge gap by developing a novel methodology, based on artificial intelligence algorithms, that can be adopted quickly and easily to provide accurate forecasts for the hazard potential level of dams in the US. In addition, this is the first research work that relies on a methodology that ensures that there was no data leakage when training and validating the developed model. Moreover, the provided model was evaluated on unseen dams' data, and it reported a relative high prediction accuracy, which reflects its robustness in making generalized predictions to any new dams. The methodological contributions are also reflected by the fact that the followed approach provided a prediction accuracy percentage that is considerably higher than that obtained in previous research works that focused on different aspects of dams. As for the practical merits and contributions of this section, they are reflected by equipping dam owners and authorities with a valuable data-driven framework that could help dam regulatory organizations to evaluate and predict the hazard potential level of their dams with a good accuracy while minimizing the efforts, time, and costs associated with the needed periodic formal inspections of the dams by authorized engineers. More specifically, this section attempted to address the deficiency in resources, funding, and staff by developing a decision support tool that could be used by US dam authorities to help minimize the difficulties faced in relation to conducting dam safety inspections. The proposed decision support tool is easy to implement, does not

require substantial investments, and does not need considerable time to report the dam's hazard level. To this end, the developed decision tool in this section is believed to be unique because there is no similar framework present in the literature. In addition, the developed decision tool could be applied to any dam in the US as it is not specific to any US state.

Section 3 – Mathematical and Risk Model for the Prediction of Project Performance in the Construction Industry: The intellectual merits and methodological contributions of this section are reflected by the novelty of the followed methodology for predicting time and cost performance of construction projects. More specifically, the methodology presented in this section was entirely developed and created for the purpose of this section, in the sense that the followed methodology is not reported in any other previous research efforts or studies, especially as related to the mapping between the different project risks and the project performance. In other words, the added value of the proposed model is that it combines the different project risks using a reliable and unique methodological process that integrates mathematical, statistical, and risk modeling concepts to provide an enhanced ability to assess and evaluate the cost and schedule overruns of construction projects. In relation to that, the model follows an unprecedented approach for the matching between the different inputs and outputs of the developed framework in a way that maximizes information usage and that allows non-linear relationships and dependencies between project risks and project performance. Furthermore, the provided mathematical formulations offer explicit equations that are not present in any other related previous research efforts on project performance. These mathematical formulations or equations could revolutionize the way project performance is assessed in the construction industry as they reduce subjective assessments by the project stakeholders and maximizes objectivity

in their evaluations. Ultimately, the developed framework contributes to the body of knowledge by providing a novel model that improves project performance in terms of prediction, control, management, analysis, and decision making. As for the practical merits and contributions, this section developed a model that allows practitioners to make better and improved predictions of project performance by incorporating all pertinent and available information on a wide spectrum of project risks. Furthermore, the presented model allows construction professionals to assess the impact of their decisions on the performance of the project and helps them take the appropriate corrective and preventive actions to minimize cost and schedule overruns. Moreover, the proposed model enables timely and targeted feedback to be executed by project administrators to determine different project issues early on and to detect warning signs before project failure. This situation helps decision makers make adjustments that ensure the smooth execution of the project by keeping it on time and on budget. Furthermore, the model improves the capabilities of project stakeholders for adopting the needed actions and decisions based on an educated judgment that incorporates information related to different project risks that affect performance.

Section 4 – A Statistical and Time Series Model to Study the Impact of Dynamic Workforce and Workplace Variables on the Productivity of the Construction Industry: The intellectual merits and methodological contributions of this section are reflected by this section's novelty in proposing a new concept of gross construction productivity. More specifically, the intellectual implications of this section include the opportunities that this research could generate in relation to studying the health of the construction industry based on the productivity of the industry as a whole and not only based on labor productivity

alone. Therefore, this section helps in moving towards the development of a theoretical notion of gross construction productivity. This notion was inspired by the use of the gross domestic product to study and evaluate the health of economies and countries, as well as by the use of the word 'gross' by the US Bureau of Labor Statistics for other variables such as gross job gains and gross job losses in the entire construction industry (US Bureau of Labor Statistics, 2020a). Furthermore, the methodology proposed and followed in this research opens exceptional opportunities to develop an indicator or index that models and incorporates the relationship between different dynamic variables and the broad economic output. In addition, and while many previous research efforts reflected the health or the importance of the construction industry by examining its share of or percentage contribution to GDP, this research's uniqueness is reflected by proposing that the health of the industry is be also examined in terms of gross construction productivity because the construction industry is one of the largest industries nationwide and a major contributors to many other industries. Thus, the intellectual merits or methodological contributions of this section are reflected by being the steppingstone or first motivational research work that would inspire and encourage scholars and practitioners to delve into this new notion of gross construction productivity more deliberately. As for the practical merits or contributions, this section opens practical opportunities in relation to using the productivity of the entire construction industry to anticipate market volatility in the construction industry as well as to identify potential improvements in the overall productivity of the industry. This is because construction productivity is an important metric that provides feedback about the industry trends and improvements (Vereen et al., 2016). This also could lead to a better execution of projects since productivity affects the performance of projects

in the construction industry (Soekiman et al., 2011). Moreover, the practical implications of this section are reflected in the importance of the research to the different entities that supply information to US construction companies. This section provides them with an additional and important piece of information (i.e., the gross construction productivity) that could be used to make different decisions. For example, the gross construction productivity could be added as an additional piece of information to the value of construction spending that the Information Handling Services (IHS) closely observes to enlighten construction firms about business risks in the construction industry (HIS, 2020). Another example is the Associated Builders and Contractors (ABC), which can provide its members with the value of gross construction productivity in addition to residential and nonresidential construction spending, which ABC carefully tracks and reports to its members on a monthly basis (ABC, 2020). In fact, construction firms utilize information about the future trends of the overall construction industry indicators in making their strategic business decisions, such as market entry, consolidation (mergers and acquisitions), business expansion and contraction, and staffing (Abediniangerabi et al., 2017). Therefore, the practical contributions of this section are to add gross construction productivity to these industrywide indicators. In summary, gross construction productivity is additional valuable information that construction companies can use to make different insightful and well-educated industry-related decisions.

Section 5 – A Hybrid Unsupervised Computational Model for Determining the Critical Combinations of Safety Fatality Causes: The intellectual merits and methodological contributions of this section lie in its novel integration of spectral clustering and data mining computational methods in a way that has not been attempted

before. More specifically, there is a lack of studies that have used clustering methods and data mining techniques to study the fatal safety accidents in the construction industry. As far as the clustering methods used by previous studies are concerned, these methods mainly included traditional approaches such as principal component analysis and k-means clustering, among others, which have many limitations in terms of their reliance on the individual properties of the data points rather than on the strength of the interconnectivities or associations between the different data points. Therefore, it was necessary to implement more advanced clustering methods to study the associations between different fatal safety causes. In relation to that, this section's reliance on spectral clustering have helped in addressing the limitations of traditional clustering methods. Similarly, as far as the data mining approaches used by previous studies are concerned, these methods did not factor the possibility of having combinations or causations between different factors or causes. More specifically, the traditional way of looking at safety incidents focuses on analyzing the weakest link in the chain of events by identifying the only one main cause for the accident and what went wrong that allowed the incident to occur (Goldberg, 2003). However, the rigid adherence to this way of thinking can lead to some significant errors in improving safety performance (Goldberg, 2003). That said, it was necessary to use other data mining techniques that has the capabilities of studying the accident causations through examining the interconnectivities and associations between different fatality causes. In relation to that, the implemented frequent pattern mining algorithm allowed for valuable combinations, associations, and relationships to be identified and for complex, interesting, and hidden associations to be discovered that otherwise could not be found by traditional techniques. In summary, compared to previous studies and the traditional techniques to

safety management, this research presents a different approach to analyze construction fatality accidents using accident causation principles. As for the practical merits and contributions, this section equips safety practitioners with a data-driven approach that can take into consideration the fact that, while safety accidents could happen due to factors or causes that are individually critical, fatalities on construction sites could also result due to a combination of factors that might not be perceived to be critical on the individual level but rather become critical when combined with other factors. That said, this research equips safety managers and supervisors with a proactive approach that allows them to take the needed preventive actions to avoid fatalities on construction sites by identifying, in hindsight, the critical combinations and associations of fatality causes. Ultimately, the outcomes of this research would enhance the safety performance in the construction industry and prevent construction fatalities by exercising more effective and efficient practices to proactively prevent the occurrence of the identified safety fatality combinations on construction sites.

Section 6 – Studying the Impact of Offsite Construction Technology on the Workforce and Labor Characteristics: The intellectual metrics and methodological contributions of this section lie in the fact that this section addresses a persistent missing piece in the body of knowledge as it is one of the first research endeavors to investigate the future implications, rather than the current impacts, of offsite construction methods and technologies on the workforce in the construction industry. Comprehensive lists of workforce occupations including onsite, offsite, engineering, construction, and administrative workforce have been identified. For each of these occupations, either the impact of offsite construction technology on the demand and skillset or the technical skills

and managerial skills of the workforce occupations were quantified; which has never been attempted before by previous research efforts. In addition, 43 labor-related characteristics were identified, rated, and quantified. Moreover, this section offered prioritization of the impacted occupations based quantitative methods as well as computational techniques such as clustering methods. Ultimately, this section led to discovering that were not addressed before. As for the practical merits and contributions, this section plays a critical role in helping offsite construction industry practitioners in workforce planning and management, in the prioritization of training needs and programs, and in improving the quality of the workforce involved in the offsite construction operations. Thus, this section presents opportunities to expand the knowledge base of the workforce occupations involved in offsite construction projects and to strengthen the skills that each employee needs to improve. Ultimately, this would help construction companies in creating and possessing an overall knowledgeable team of workers and employees that are capable of take over for one another as needed and that can work on teams or work independently without constant help and supervision from others. This research could also help industry practitioners in improving the productivity and performance of their workforce as it provides them with the opportunity to be adequately and properly trained to perform the different tasks and activities as well as to have a stronger understanding of the offsite construction aspects and the responsibilities of their jobs. Consequently, this research is valuable for construction firms in keeping their employees on the cutting edge of industry developments. This could ultimately help organizations in holding a position as a leader and strong competitor within the offsite construction industry.

7.3. FUTURE WORK

The research conducted in this dissertation opens multiple opportunities for scholars and/or practitioners to build on the presented efforts in this dissertation in their future endeavors. In relation to that, future work related to Section 2 could include investigating the performance of different artificial intelligence algorithms or developing other machine learning models to predict the safety potential level of dams at the state level and not only on the national level. This will result in having a set of reliable models or frameworks that are capable of predicting the hazard potential of dams specific to each state. Also, while the research in this dissertation is applied to dams in the US, the proposed technique or approach is scalable and malleable so it could be used on any similar international available data set. Future work and developments related to Section 3 could be to develop a methodology that provides different weights for cost and schedule overruns, respectively. Additionally, a comparative study and an in-depth sensitivity analysis of the considered 25 project risks could be performed to reflect the importance of each risk on project performance. These efforts will assist in prioritizing the different project risks based on their contributions to cost and schedule overruns. Furthermore, future work could incorporate an additional dimension for the developed model that allows the reflection of potential correlations between the different project risks. Doing so will ensure a model that is better able to factor the different project uncertainties in the construction industry. Future work related to Section 4 could be to address one of the limitations of this section, which is the focus on the unidirectional relationship between the different dynamic workforce and workplace variables on the one hand and the productivity of the construction industry on the other hand. That is, this research was limited to studying

how the response variable (construction productivity) was influenced by the predictor variables (workforce and workplace variables), but not vice versa. Since this section focused on the workforce and workplace variables that can affect construction productivity rather than other industry-, government-, or economic-related variables, future research work could include studying the effects of other variables on construction productivity. Moreover, since this section mainly aimed to be a proof-of-concept for a new notion of gross construction productivity, it is desired that such research acts as a step toward the development of a new gross construction productivity indicator by future scholars or practitioners. Future work related to Section 5 could include to devised methods or approached to integrate the accident causation systems with the different technologies—that are currently used on construction sites to monitor the progress of the work and identify possible unsafe events—to provide real-time and automated warning signs of the possibility of an accident in hindsight (i.e., before the occurrence of the incident). For instance, future work could focus on applied the proposed approach in this research in the safety management practice in real construction sites by integrating it with the unmanned aerial vehicles (or drones) and computer vision that can automatically raise a red flag or symbol when the identified combinations of causes occur on the construction site. It is worth mentioning that scholars as well as construction companies could also rely on their own data set of construction incidents or fatal accidents in order to identify other possible critical combinations of factors that are more specific to the companies' construction sector or operations. Finally, future research work related to Section 6 could include providing specific training aspects or skills that need to be acquired by each one of the prioritized workforce occupations that were reported in this dissertation. Other potential future efforts

could include designing specific workforce development programs and training materials for the different workforce occupations considered in this section or those that are involved in the offsite construction projects. This would help construction organizations in having more specific guidelines and practices for properly and adequately training their workforce so that to have a competent, well-rounded, and multi-skilled labor force that could perform the different offsite construction tasks and activities.

APPENDIX A.

DATA AND PYTHON CODE FOR THE DEVELOPED SUPERVISED MODEL TO EVALUATE AND PREDICT THE HAZARD POTENTIAL LEVEL OF DAM INFRASTRUCTURES

The data used for developing the supervised artificial intelligence model is consisting of 91,468 readings that were downloaded from the official website of the National Inventory of Dams that could be accessed at the following link: https://nid.sec.usace.army.mil/ords/f?p=105:1::::::

It is worth mentioning that since the data includes 91,468 readings, it was not possible to include it in this dissertation or this appendix. However, since the data is publicly available on the previous provided link then there should be no problem with that. It is to be noted that the data used in this dissertation was the one collected for the 2018 year and for all the states in the US. The variables that were considered in the data are the following: (1) DISTANCE; (2) DAM_TYPE; (3) CORE; (4) FOUNDATION; (5) PURPOSES; (6) AGE; (7) Modified/Maintenance?; (8) DAM_LENGTH; (9) DAM_HEIGHT; (10) STRUCTURAL_HEIGHT; (11) HYDRAULIC_HEIGHT; (12) NID_HEIGHT; (13) MAX_DISCHARGE; (14) MAX_STORAGE; (15) NORMAL_STORAGE; (16) NID_STORAGE; (17) SURFACE_AREA; (18) DRAINAGE_AREA; (19) INSPECTION_FREQUENCY; (20) SPILLWAY_WIDTH; (21) VOLUME; (22) NUMBER_OF_LOCKS; (23) LENGTH_OF_LOCKS; (24) WIDTH_OF_LOCKS; (25) NUMSEPARATESTRUCTURES; and, (26) HAZARD.

It is to be noted that all the first 25 variables were considered as potential input variables for the supervised artificial intelligence model while the last variable 'HAZARD' is the output variable. It is worth mentioning that while all variables were directly taken from the provided raw data from the National Inventory of Dams; two variables were indirectly calculated from such data. The first variable is the 'AGE' which was calculated as 2018 minus the variable 'YEAR_COMPLETED' that is present in the raw data

published by the National Dam Inventory of Dams. The second variable is the 'Modified/Maintenance?' which is a categorical variable that could take a value of 'N' representing 'No' or a value of 'Y' representing 'Yes'. This variable takes the value of 'N' if the corresponding variable 'YEAR_MODIFIED' (that is present in the raw data published by the National Inventory of Dams) does not include any reading, and it takes a value of 'Y' if the corresponding variable 'YEAR_MODIFIED' does include a reading. This could be achieved via excel using the following formula '=IF(Cell#,"N","Y")' where Cell# refers to the associated readings of the 'YEAR_MODIFIED' variable.

To be able to run the implemented code below, the 26 variables need to be saved in a csv (comma-separated values) file named: 'NID2018_Initial Data_25 Input Variables.csv'. The implemented Python code (the code's comments are shown as underlined) for the developed supervised artificial intelligence model is as follows:

```
#importing packages
import pandas as pd
import numpy as np
from sklearn.compose import ColumnTransformer
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.compose import make_column_transformer
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
```

```
from sklearn.ensemble import RandomForestClassifier

from boruta import BorutaPy

from sklearn.neighbors import KNeighborsClassifier

from sklearn.neural_network import MLPClassifier

from sklearn.model_selection import learning_curve, GridSearchCV

from sklearn.model_selection import StratifiedKFold

from sklearn.model_selection import PredefinedSplit

from imblearn.combine import SMOTETomek

from imblearn.over_sampling import SMOTE

from collections import Counter

from sklearn import metrics

from itertools import chain

from sklearn.metrics import classification_report, confusion_matrix,

balanced_accuracy_score, accuracy_score,average_precision_score, brier_score_loss

%matplotlib inline

'''The following versions were used for the different packages:

Version of Sklearn is: 0.24.1

Version of Pandas is: 1.1.5

Version of Numpy is: 1.20.1

Version of Boruta is: 0.3

Version of Imblearn is: 0.8.0

'''

#reading the data as a dataframe
```

```
original_data_categorical_representation = pd.read_csv("NID2018_Initial Data_25
Input Variables.csv")
#visualizing the head of the dataframe
original_data_categorical_representation.head()
#checking the type of the variables
original_data_categorical_representation.info() #5 categorical input variables:
'DAM_TYPE', 'CORE', 'FOUNDATION', 'PURPOSES', and 'Modified/Maintenance?'
#we need to transform the categorical input variables to numerical variables using the
label encoding method by transforming the categorical features to numerical values
between 0 and the number of attributes minus 1
#first, get the unique values (excluding Null values) for each categorical variable
#second, sort the unique values by alphabetical order for ease of interpretation
#third, we create a dictionary where the keys (starting at 0) are the indices of the unique
values and the dictionary's values are the unique values themselves
#fourth, we flip the keys and the values of the dictionary
#first categorical variable: DAM_TYPE
dic_DamType = dict(map(reversed,
dict(enumerate(sorted(original_data_categorical_representation['DAM_TYPE'].dropna().
unique()), 0)).items()))
#second categorical variable: PURPOSES
dic_Purposes = dict(map(reversed,
dict(enumerate(sorted(original_data_categorical_representation['PURPOSES'].dropna().u
nique()), 0)).items()))
#third categorical variable: FOUNDATION
```

```
dic_Foundation = dict(map(reversed,
dict(enumerate(sorted(original_data_categorical_representation['FOUNDATION'].dropn
a().unique()), 0)).items()))
#fourth categorical variable: CORE
dic_Core = dict(map(reversed,
dict(enumerate(sorted(original_data_categorical_representation['CORE'].dropna().unique
()), 0)).items()))
#fifth categorical variable: Modified/Maintenance?
dic_Maintenance = dict(map(reversed,
dict(enumerate(sorted(original_data_categorical_representation['Modified/Maintenance?'
].dropna().unique()), 0)).items()))
#transforming the categorical data into numerical data based on the created dictionaries
original_data_numerical_representation =
original_data_categorical_representation.copy()
original_data_numerical_representation['DAM_TYPE'].replace(dic_DamType,
inplace=True)
original_data_numerical_representation['CORE'].replace(dic_Core, inplace=True)
original_data_numerical_representation['FOUNDATION'].replace(dic_Foundation,
inplace=True)
original_data_numerical_representation['PURPOSES'].replace(dic_Purposes,
inplace=True)
original_data_numerical_representation['Modified/Maintenance?'].replace(dic_Maintena
nce, inplace=True)
```

```
#saving the data
original_data_numerical_representation.to_csv('NID2018_Initial Data_25 Input Variables_Numerical.csv')
#checking the class distribution of the output variable
original_data_categorical_representation['HAZARD'].value_counts() #U means undetermined
#checking how many null values are present in the output variable
original_data_categorical_representation['HAZARD'].isnull().sum()
#since the output variable is 'HAZARD' which is a very critical aspect of dams, dams with undetermined hazard level (i.e., with 'U') and with null values are dropped from the dataset rather than imputed using data imputation methods so that to avoid assigning a value which is not 100% correct
#dropping null values
original_data_numerical_representation = original_data_numerical_representation[original_data_numerical_representation.HAZARD.notnull()]
#dropping dams with undetermined hazard level
original_data_numerical_representation = original_data_numerical_representation[original_data_numerical_representation.HAZARD != 'U']
#resetting the index
original_data_numerical_representation.reset_index(drop=True, inplace=True)
#changing the output variable such as Low is 0, Significant is 1, and H is 2
```

```
original_data_numerical_representation['HAZARD'] =
original_data_numerical_representation['HAZARD'].map({'L':0,'S':1,'H':2})
#selecting the input variables
X_original=original_data_numerical_representation.drop('HAZARD', axis=1)
#selecting the output variable
y_original=original_data_numerical_representation['HAZARD']
#spliting
X_trainVal,X_test,y_trainVal,y_test=train_test_split(X_original,y_original,test_size=0.2,
random_state=1,stratify=y_original)
#getting the values of the splited data
X_trainVal=X_trainVal.values
y_trainVal=y_trainVal.values
```

```
#The Python code for the KNN algorithm is as follows:
skf=StratifiedKFold(n_splits=5,random_state=1,shuffle=True) #choosing how many
folds we want through the 'n_splits' parameter
summary_accuracies_perFold_perHyperparameterCombination=pd.DataFrame()
#initializing a dataframe that will store the obtained accuracies on the validation set
iteration_number=0 #initializing the iteration counter
list_of_selected_features = [[] for i in range(skf.n_splits)] #initializing a list of list to
store the selected features for each iteration
for train_index, val_index in skf.split(X_trainVal,y_trainVal):
    #increasing the iteration
    iteration_number=iteration_number+1
```

```
    #printing with color (red)
    print('\033[31m' + '\nCross-Validation ITERATION #',iteration_number,'out
of',skf.n_splits,'ITERATIONS\n'+ '\033[0m')
    #defining the training set and the validation set for each iteration
    X_train, X_val = X_trainVal[train_index], X_trainVal[val_index]
    y_train, y_val = y_trainVal[train_index], y_trainVal[val_index]


    #scaling the training set
    print('\nScaling the training folds STARTED for iteration #',iteration_number,'\n')
    scaler = MinMaxScaler()
    X_train_scaled=scaler.fit_transform(X_train)
    print('Scaling the training folds ENDED for iteration #',iteration_number,'\n')
    #imputing the training set
    print('\nImputing the training folds STARTED for iteration #',iteration_number,'\n')
    imputer=IterativeImputer(max_iter=100,random_state=1,
verbose=0,min_value=0,max_value=1)
    X_train_scaled_imputed=imputer.fit_transform(X_train_scaled)
    print('Imputing the training folds ENDED for iteration #',iteration_number,'\n')
    #feature selection
    print('\nFeature selection on the training folds STARTED for iteration
#',iteration_number)
    rf=RandomForestClassifier(n_jobs=-1, class_weight='balanced') #defining random
forest classifier
```

```
    feat_selector = BorutaPy(rf, n_estimators='auto',max_iter=500, verbose=0,
random_state=1) #defining Boruta feature selection method
    feat_selector.fit(X_train_scaled_imputed, y_train) #finding all relevant features
    print ('\nNumber of selected features:',feat_selector.n_features_)
    X_train_scaled_imputed_filtered = feat_selector.transform(X_train_scaled_imputed)
#calling the transform() to filter the data down to selected features
    print('\nVariables to keep based on Boruta for iteration
#',iteration_number,':\n',X_original.columns[feat_selector.support_].values)
    #storing the selected features
    list_of_selected_features[iteration_number-1] =
X_original.columns[feat_selector.support_].values
    print('\nFeature selection on the training folds ENDED for iteration
#',iteration_number,'\n')
    #SMOTE
    print("\nThe number of classes before SMOTE is:", Counter(y_train))
    print ('\nApplying SMOTE STARTED for iteration #',iteration_number,'\n')
    X_train_scaled_imputed_filtered_sm,y_train_sm=SMOTE(random_state=1,n_jobs=-
1).fit_resample(X_train_scaled_imputed_filtered,y_train)
    print ('Applying SMOTE ENDED for iteration #',iteration_number,'\n')
    print("The number of classes after SMOTE is:", Counter(y_train_sm),'\n')
    #transforming the validation set
    print('\nScaling, imputing, and dimension reduction for the validation fold STARTED
for iteration #',iteration_number,'\n')
```

```
    X_val_scaled = scaler.transform(X_val)
    X_val_scaled_imputed = imputer.transform(X_val_scaled)
    X_val_scaled_imputed_filtered=feat_selector.transform(X_val_scaled_imputed)
    print('Scaling, imputing, and dimension reduction for the validation fold ENDED for iteration #',iteration_number,'\n')
    #we want to do gridsearch now for each one of the k-fold iterations, but sklearn's GridSearchCV (with cv=integer) includes cross validation (which is not needed for our case because we already used StratifiedKFold)
    #in fact, what is needed for our case is to do grid search through training the different models (with different parameters) on the training and getting the accuracy (for each model) on the validation set (this approach cannot be done easily with sklearn's GridSearchCV (with cv=integer))
    #therefore we need to define the cv parameter in sklearn's GridSearchCV differently (i.e., by including a generator) which could be done with the help of sklearn's PredefinedSplit()
    x = np.concatenate([X_train_scaled_imputed_filtered_sm, X_val_scaled_imputed_filtered]) #combining the training and validation samples for the input variables
    y = np.concatenate([y_train_sm, y_val]) #combining the training and validation samples for the output variable
    #PredefinedSplit() requires one parameter (test_fold) thus we need to define it properly
    test_fold = np.concatenate([
    #setting the test_fold to -1 for the training sample
```

```
    np.full(shape=X_train_scaled_imputed_filtered_sm.shape[0],fill_value=-1),
    #setting the test_fold to 0 for the validation sample
    np.zeros(X_val_scaled_imputed_filtered.shape[0])
    ]) #test_fold will be in the form of [-1,-1,-1,-1,-1,.....,0,0,0,0] where the number of -1s is size of the training set and the number of 0s is the size of the validation set
    cv = PredefinedSplit(test_fold)
    #defining the gridsearch
    gridsearch = GridSearchCV(KNeighborsClassifier(algorithm='auto'),{"n_neighbors":list(range(1,41)),'weights':['uniform', 'distance'],'p':[1,2]},cv=cv,verbose=2,n_jobs=-1) #since the 'scoring' parameter is not specified for the used GridSearchCV, the estimator's (here, KNN Classifier) default score method is used (here, 'Accuracy').
    #fitting the gridsearch
    print('\nTraining STARTED for iteration #',iteration_number,'\n')
    gridsearch.fit(x,y)
    print('Training ENDED for iteration #',iteration_number,'\n')
    #printing the results of the gridsearch
    print('\nValidation RESULTS for iteration #',iteration_number,'\n')
    print(pd.DataFrame(gridsearch.cv_results_)[['params', 'split0_test_score']]) #printing only the results of interest
    #saving the results into the created dataframe

summary_accuracies_perFold_perHyperparameterCombination=summary_accuracies_pe
```

```
rFold_perHyperparameterCombination.append(pd.DataFrame(gridsearch.cv_results_)[['p
arams', 'split0_test_score']])
#getting the average accuracies
average_accuracies_perHyperparameterCombination=summary_accuracies_perFold_per
HyperparameterCombination.groupby(level=0).mean() #getting the average accuracies of
all folds PER HYPERPARAMETER COMBINATION
#changing the name of the column from 'split0_test_score' to 'average accuracies'
average_accuracies_perHyperparameterCombination.rename(columns={"split0_test_scor
e": "average accuracies"},inplace=True)
#adding the hyperparameters' combination to the associated averages
average_accuracies_perHyperparameterCombination['params']=summary_accuracies_per
Fold_perHyperparameterCombination.iloc[:average_accuracies_perHyperparameterCom
bination.index.values.max()+1,:]['params']
#getting the best combination of hyperparameter and the best accuracy
average_accuracies_perHyperparameterCombination.sort_values(by='average
accuracies',ascending=False,inplace=True) #sorting from highest to lowest accuracy
print('Highest average accuracy (on the validation set)
is:',round(average_accuracies_perHyperparameterCombination['average
accuracies'].values[0]*100,2),'%\n')
best_parameters=average_accuracies_perHyperparameterCombination['params'].values[0]
print('The best hyperparameters are:',best_parameters)
#data preparation for the testing set
#scaling the training/validation set
```

```
print('\nScaling the 80% training set STARTED')
scaler = MinMaxScaler()
X_trainVal_scaled=scaler.fit_transform(X_trainVal)
print('Scaling the 80% training set ENDED\n')

#imputing the training/validation set
print('Imputing the 80% training set STARTED')
imputer=IterativeImputer(max_iter=100,random_state=1,verbose=0,min_value=0,max_v
alue=1)
X_trainVal_scaled_imputed=imputer.fit_transform(X_trainVal_scaled)
print('Imputing the 80% training set ENDED\n')

#feature selection
print('Reducing the dimension of the 80% training set STARTED')
#getting all the features that were selected (i.e., all the unique values in the variable
'list_of_selected_features')
all_selected_features=list(set(chain(*list_of_selected_features))) #this is a list
#getting the indices (based on their initial sequence) of the selected features
indices_selected_features=[X_original.columns.get_loc(c) for c in all_selected_features
if c in X_original]
#reducing the dimension of the training/validation set
X_trainVal_scaled_imputed_filtered=X_trainVal_scaled_imputed[:,indices_selected_feat
ures]
```

```
print('Reducing the dimension of the 80% training set ENDED\n')


#SMOTE
print("\nThe number of classes before SMOTE is:", Counter(y_trainVal))
print ('Applying SMOTE STARTED for the 80% training set STARTED')
X_trainVal_scaled_imputed_filtered_sm,y_trainVal_sm=SMOTE(random_state=1,n_job
s=-1).fit_resample(X_trainVal_scaled_imputed_filtered,y_trainVal)
print ('Applying SMOTE ENDED for the 80% training set STARTED')
print("The number of classes after SMOTE is:", Counter(y_trainVal_sm),'\n')


#transforming the testing set
print('Scaling, imputing, and reducing the dimensions of the 20% testing set STARTED')
X_test_scaled = scaler.transform(X_test)
X_test_scaled_imputed = imputer.transform(X_test_scaled)
X_test_scaled_imputed_filtered=X_test_scaled_imputed[:,indices_selected_features]
print('Scaling, imputing, and reducing the dimensions of the 20% testing set ENDED\n')
#assigning the best parameters for the KNN algorithm
best_model=KNeighborsClassifier(**best_parameters)
#training on the Train/Val set
best_model.fit(X_trainVal_scaled_imputed_filtered_sm,y_trainVal_sm)
#prediction
y_pred=best_model.predict(X_test_scaled_imputed_filtered)
#testing accuracy
```

```
print('Testing Accuracy:',round(accuracy_score(y_test, y_pred)*100,2),'%')
#classification report
print('\n Classification Report:\n',classification_report(y_test,y_pred))
#getting the confusion matrix
cm = confusion_matrix(y_test, y_pred)
#printing the confusion matrix
print('\n Classification Matrix:\n',cm)
#plotting the confusion matrix
plt.figure(figsize=(6,6))
plt.matshow(cm,cmap=plt.cm.Blues,fignum=1)
plt.colorbar()
plt.ylabel('True Hazard Level\n',fontsize=20)
plt.xticks(np.arange(3),['Low','Significant','High'],fontsize=20,rotation=45)
plt.yticks(np.arange(3),['Low','Significant','High'],fontsize=20)
plt.xlabel('\nPredicted Hazard Level',fontsize=20)
for (i, j), z in np.ndenumerate(cm):
    plt.text(j, i, '{:0.0f}'.format(z), ha='center',
va='center',color='red',fontweight='bold',fontsize=20)
plt.show()

#getting the normalized confusion matrix
cmn = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
#printing the normalized confusion matrix
```

```
print('\n Normalized Classification Matrix:\n',cmn)
#plotting the normalized confusion matrix
plt.figure(figsize=(6,6))
plt.matshow(cmn,cmap=plt.cm.Blues,vmin=0, vmax=1,fignum=1)
plt.colorbar()
plt.ylabel('True Hazard Level\n',fontsize=20)
plt.xticks(np.arange(3),['Low','Significant','High'],fontsize=20,rotation=45)
plt.yticks(np.arange(3),['Low','Significant','High'],fontsize=20)
plt.xlabel('\nPredicted Hazard Level',fontsize=20)
for (i, j), z in np.ndenumerate(cmn):
    plt.text(j, i, '{:0.2f}'.format(z), ha='center',
va='center',color='red',fontweight='bold',fontsize=20)
plt.show()
#getting the importance of the selected features
#defining the algorithm to be used
forest = RandomForestClassifier(random_state=1)
#fitting the algorithm on the training/validation data
forest.fit(X_trainVal_scaled_imputed_filtered_sm,y_trainVal_sm)
#getting the importance of each feature
importances = forest.feature_importances_
#getting the standard deviation of each feature
std = np.std([tree.feature_importances_ for tree in forest.estimators_],
             axis=0)
```

```
#getting the indices (based on the reduced size; i.e., from 0 to maximum number of
selected features -1) of each feature
indices = np.argsort(importances)[::-1]


#printing the feature ranking
print("Feature ranking:\n")
for i,j in zip (X_original.columns[indices_selected_features][indices],
importances[indices]):
   print(i,j)


#plotting the impurity-based feature importance of the forest
plt.figure(figsize=(10,8))
plt.title("Feature importances\n",fontsize=16)
plt.barh(range(X_trainVal_scaled_imputed_filtered_sm.shape[1]),
np.flip(importances[indices]), #np.flip is used so that to visualize the plot from the
highest importance to the lowest
      color="r", xerr=np.flip(std[indices]), align="center") #np.flip is used so that to
visualize the plot from the highest importance to the lowest
plt.yticks(range(X_trainVal_scaled_imputed_filtered_sm.shape[1]),
np.flip(X_original.columns[indices_selected_features][indices]),fontsize=16) #np.flip is
used so that to visualize the plot from the highest importance to the lowest
plt.xticks(fontsize=16)
plt.ylim([-1, X_trainVal_scaled_imputed_filtered_sm.shape[1]])
```

```
plt.show()
```

```
#The Python code for the ANN algorithm is as follows:
skf=StratifiedKFold(n_splits=5,random_state=1,shuffle=True) #choosing how many folds we want through the 'n_splits' parameter
summary_accuracies_perFold_perHyperparameterCombination=pd.DataFrame() #initializing a dataframe that will store the obtained accuracies on the validation set
iteration_number=0 #initializing the iteration counter
list_of_selected_features = [[] for i in range(skf.n_splits)] #initializing a list of list to store the selected features for each iteration
for train_index, val_index in skf.split(X_trainVal,y_trainVal):
    #increasing the iteration
    iteration_number=iteration_number+1
    #printing with color (red)
    print('\033[31m' + '\nCross-Validation ITERATION #',iteration_number,'out of',skf.n_splits,'ITERATIONS\n'+ '\033[0m')
    #defining the training set and the validation set for each iteration
    X_train, X_val = X_trainVal[train_index], X_trainVal[val_index]
    y_train, y_val = y_trainVal[train_index], y_trainVal[val_index]

    #scaling the training set
    print('\nScaling the training folds STARTED for iteration #',iteration_number,'\n')
    scaler = MinMaxScaler()
    X_train_scaled=scaler.fit_transform(X_train)
```

```
    print('Scaling the training folds ENDED for iteration #',iteration_number,'\n')
    #imputing the training set
    print('\nImputing the training folds STARTED for iteration #',iteration_number,'\n')
    imputer=IterativeImputer(max_iter=100,random_state=1,
verbose=0,min_value=0,max_value=1)
    X_train_scaled_imputed=imputer.fit_transform(X_train_scaled)
    print('Imputing the training folds ENDED for iteration #',iteration_number,'\n')
    #feature selection
    print('\nFeature selection on the training folds STARTED for iteration
#',iteration_number)
    rf=RandomForestClassifier(n_jobs=-1, class_weight='balanced') #defining random
forest classifier
    feat_selector = BorutaPy(rf, n_estimators='auto',max_iter=500, verbose=0,
random_state=1) #defining Boruta feature selection method
    feat_selector.fit(X_train_scaled_imputed, y_train) #finding all relevant features
    print ('\nNumber of selected features:',feat_selector.n_features_)
    X_train_scaled_imputed_filtered = feat_selector.transform(X_train_scaled_imputed)
#calling the transform() to filter the data down to selected features
    print('\nVariables to keep based on Boruta for iteration
#',iteration_number,':\n',X_original.columns[feat_selector.support_].values)
    #storing the selected features
    list_of_selected_features[iteration_number-1] =
X_original.columns[feat_selector.support_].values
```

```
    print('\nFeature selection on the training folds ENDED for iteration
#',iteration_number,'\n')

    #SMOTE
    print("\nThe number of classes before SMOTE is:", Counter(y_train))
    print ('\nApplying SMOTE STARTED for iteration #',iteration_number,'\n')
    X_train_scaled_imputed_filtered_sm,y_train_sm=SMOTE(random_state=1,n_jobs=-
1).fit_resample(X_train_scaled_imputed_filtered,y_train)
    print ('Applying SMOTE ENDED for iteration #',iteration_number,'\n')
    print("The number of classes after SMOTE is:", Counter(y_train_sm),'\n')

    #transforming the validation set
    print('\nScaling, imputing, and dimension reduction for the validation fold STARTED
for iteration #',iteration_number,'\n')
    X_val_scaled = scaler.transform(X_val)
    X_val_scaled_imputed = imputer.transform(X_val_scaled)
    X_val_scaled_imputed_filtered=feat_selector.transform(X_val_scaled_imputed)
    print('Scaling, imputing, and dimension reduction for the validation fold ENDED for
iteration #',iteration_number,'\n')

    #we want to do gridsearch now for each one of the k-fold iterations, but sklearn's
GridSearchCV (with cv=integer) includes cross validation (which is not needed for our
case because we already used StratifiedKFold)
```

```
    #in fact, what is needed for our case is to do grid search through training the different models (with different parameters) on the training and getting the accuracy (for each model) on the validation set (this approach cannot be done easily with sklearn's GridSearchCV (with cv=integer))


    #therefore we need to define the cv parameter in sklearn's GridSearchCV differently (i.e., by including a generator) which could be done with the help of sklearn's PredefinedSplit()
    x = np.concatenate([X_train_scaled_imputed_filtered_sm, X_val_scaled_imputed_filtered]) #combining the training and validation samples for the input variables
    y = np.concatenate([y_train_sm, y_val]) #combining the training and validation samples for the output variable
    #PredefinedSplit() requires one parameter (test_fold) thus we need to define it properly
    test_fold = np.concatenate([
    #setting the test_fold to -1 for the training sample
    np.full(shape=X_train_scaled_imputed_filtered_sm.shape[0],fill_value=-1),
    #setting the test_fold to 0 for the validation sample
    np.zeros(X_val_scaled_imputed_filtered.shape[0])
    ]) #test_fold will be in the form of [-1,-1,-1,-1,-1,.....,0,0,0,0] where the number of -1s is size of the training set and the number of 0s is the size of the validation set
    cv = PredefinedSplit(test_fold)
    #defining the gridsearch
```

```
    gridsearch = GridSearchCV
(MLPClassifier(random_state=1),{'hidden_layer_sizes': [(1,1), (2,2), (3,3), (4,4), (5,5),
(6,6), (7,7), (8,8), (9,9), (10,10),
                                                    (1,1,1), (2,2,2), (3,3,3), (4,4,4),
(5,5,5), (6,6,6), (7,7,7), (8,8,8), (9,9,9), (10,10,10),
                                                    (1,1,1,1), (2,2,2,2), (3,3,3,3),
(4,4,4,4), (5,5,5,5), (6,6,6,6), (7,7,7,7), (8,8,8,8), (9,9,9,9), (10,10,10,10),
                                                    (1,1,1,1,1), (2,2,2,2,2), (3,3,3,3,3),
(4,4,4,4,4), (5,5,5,5,5), (6,6,6,6,6), (7,7,7,7,7), (8,8,8,8,8), (9,9,9,9,9), (10,10,10,10,10),
                                                    (1,1,1,1,1,1), (2,2,2,2,2,2),
(3,3,3,3,3,3), (4,4,4,4,4,4), (5,5,5,5,5,5), (6,6,6,6,6,6), (7,7,7,7,7,7), (8,8,8,8,8,8),
(9,9,9,9,9,9), (10,10,10,10,10,10)],
                                                    'solver':['adam'],'alpha': [0.0001,
0.05],'batch_size':[32, 64, 128, 256],'learning_rate_init':[0.1,
0.001]},cv=cv,verbose=2,n_jobs=-1) #since the 'scoring' parameter is not specified for
the used GridSearchCV, the estimator's (here, ANN Classifier) default score method is
used (here, 'Accuracy').
    #fitting the gridsearch
    print('\nTraining STARTED for iteration #',iteration_number,'\n')
    gridsearch.fit(x,y)
    print('Training ENDED for iteration #',iteration_number,'\n')
    #printing the results of the gridsearch
    print('\nValidation RESULTS for iteration #',iteration_number,'\n')
```

```
    print(pd.DataFrame(gridsearch.cv_results_)[['params', 'split0_test_score']])
#printing only the results of interest

    #saving the results into the created dataframe

summary_accuracies_perFold_perHyperparameterCombination=summary_accuracies_pe
rFold_perHyperparameterCombination.append(pd.DataFrame(gridsearch.cv_results_)[['p
arams', 'split0_test_score']])

#getting the average accuracies
average_accuracies_perHyperparameterCombination=summary_accuracies_perFold_per
HyperparameterCombination.groupby(level=0).mean() #getting the average accuracies of
all folds PER HYPERPARAMETER COMBINATION
#changing the name of the column from 'split0_test_score' to 'average accuracies'
average_accuracies_perHyperparameterCombination.rename(columns={"split0_test_scor
e": "average accuracies"},inplace=True)
#adding the hyperparameters' combination to the associated averages
average_accuracies_perHyperparameterCombination['params']=summary_accuracies_per
Fold_perHyperparameterCombination.iloc[:average_accuracies_perHyperparameterCom
bination.index.values.max()+1,:]['params']
#getting the best combination of hyperparameter and the best accuracy
average_accuracies_perHyperparameterCombination.sort_values(by='average
accuracies',ascending=False,inplace=True) #sorting from highest to lowest accuracy
```

```
print('Highest average accuracy (on the validation set)
is:',round(average_accuracies_perHyperparameterCombination['average
accuracies'].values[0]*100,2),'%\n')
best_parameters=average_accuracies_perHyperparameterCombination['params'].values[0
]
print('The best hyperparameters are:',best_parameters)
#data preparation for the testing set

#scaling the training/validation set
print('\nScaling the 80% training set STARTED')
scaler = MinMaxScaler()
X_trainVal_scaled=scaler.fit_transform(X_trainVal)
print('Scaling the 80% training set ENDED\n')

#imputing the training/validation set
print('Imputing the 80% training set STARTED')
imputer=IterativeImputer(max_iter=100,random_state=1,verbose=0,min_value=0,max_v
alue=1)
X_trainVal_scaled_imputed=imputer.fit_transform(X_trainVal_scaled)
print('Imputing the 80% training set ENDED\n')

#feature selection
print('Reducing the dimension of the 80% training set STARTED')
```

```
#getting all the features that were selected (i.e., all the unique values in the variable
'list_of_selected_features')
all_selected_features=list(set(chain(*list_of_selected_features))) #this is a list
#getting the indices (based on their initial sequence) of the selected features
indices_selected_features=[X_original.columns.get_loc(c) for c in all_selected_features
if c in X_original]
#reducing the dimension of the training/validation set
X_trainVal_scaled_imputed_filtered=X_trainVal_scaled_imputed[:,indices_selected_feat
ures]
print('Reducing the dimension of the 80% training set ENDED\n')

#SMOTE
print("\nThe number of classes before SMOTE is:", Counter(y_trainVal))
print ('Applying SMOTE STARTED for the 80% training set STARTED')
X_trainVal_scaled_imputed_filtered_sm,y_trainVal_sm=SMOTE(random_state=1,n_job
s=-1).fit_resample(X_trainVal_scaled_imputed_filtered,y_trainVal)
print ('Applying SMOTE ENDED for the 80% training set STARTED')
print("The number of classes after SMOTE is:", Counter(y_trainVal_sm),'\n')

#transforming the testing set
print('Scaling, imputing, and reducing the dimensions of the 20% testing set STARTED')
X_test_scaled = scaler.transform(X_test)
X_test_scaled_imputed = imputer.transform(X_test_scaled)
```

```
X_test_scaled_imputed_filtered=X_test_scaled_imputed[:,indices_selected_features]
print('Scaling, imputing, and reducing the dimensions of the 20% testing set ENDED\n')
#assigning the best parameters for the ANN algorithm
best_model =MLPClassifier(**best_parameters)
#training on the Train/Val set
best_model.fit(X_trainVal_scaled_imputed_filtered_sm,y_trainVal_sm)
#prediction
y_pred=best_model.predict(X_test_scaled_imputed_filtered)
#testing accuracy
print('Testing Accuracy:',round(accuracy_score(y_test, y_pred)*100,2),'%')
#classification report
print('\n Classification Report:\n',classification_report(y_test,y_pred))
#getting the confusion matrix
cm = confusion_matrix(y_test, y_pred)
#printing the confusion matrix
print('\n Classification Matrix:\n',cm)
#plotting the confusion matrix
plt.figure(figsize=(6,6))
plt.matshow(cm,cmap=plt.cm.Blues,fignum=1)
plt.colorbar()
plt.ylabel('True Hazard Level\n',fontsize=20)
plt.xticks(np.arange(3),['Low','Significant','High'],fontsize=20,rotation=45)
plt.yticks(np.arange(3),['Low','Significant','High'],fontsize=20)
```

```
plt.xlabel('\nPredicted Hazard Level',fontsize=20)
for (i, j), z in np.ndenumerate(cm):
    plt.text(j, i, '{:0.0f}'.format(z), ha='center',
va='center',color='red',fontweight='bold',fontsize=20)
plt.show()

#getting the normalized confusion matrix
cmn = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
#printing the normalized confusion matrix
print('\n Normalized Classification Matrix:\n',cmn)
#plotting the normalized confusion matrix
plt.figure(figsize=(6,6))
plt.matshow(cmn,cmap=plt.cm.Blues,vmin=0, vmax=1,fignum=1)
plt.colorbar()
plt.ylabel('True Hazard Level\n',fontsize=20)
plt.xticks(np.arange(3),['Low','Significant','High'],fontsize=20,rotation=45)
plt.yticks(np.arange(3),['Low','Significant','High'],fontsize=20)
plt.xlabel('\nPredicted Hazard Level',fontsize=20)
for (i, j), z in np.ndenumerate(cmn):
    plt.text(j, i, '{:0.2f}'.format(z), ha='center',
va='center',color='red',fontweight='bold',fontsize=20)
plt.show()
```

APPENDIX B.

EQUATIONS OF THE FITTED DISTRIBUTIONS FOR EACH PROJECT RISK

The following equations, Equation B1 to B25, show the fitted cumulative distribution functions for the 25 project risks present in Section 3.

$$F(c) = \begin{cases} 0 & if\ 0 \le c < 0.11425 \\ 1.453c - 0.166 & if\ 0.11425 \le c \le 0.80242 \\ 1 & if\ 0.80242 < c \le 1 \end{cases} \quad \text{(B1)}$$

$$F(c) = \begin{cases} 0 & if\ 0 \le c < 0.068548 \\ 1.07360c - 0.07360 & if\ 0.068548 \le c \le 1 \end{cases} \quad \text{(B2)}$$

$$F(c) = \frac{1.011}{1 + \left(\frac{1}{\frac{c + 0.087114}{0.40462}}\right)^{4.5594}} \quad for\ 0 \le c \le 1 \quad \text{(B3)}$$

$$F(c) = \begin{cases} 0 & if\ 0 \le c < 0.0059025 \\ 1 - \left[1 - \left(\frac{c - 0.0059025}{0.9620375}\right)^{1.7421}\right]^{3.7056} & if\ 0.0059025 \le c \le 0.96794 \\ 1 & if\ 0.96794 < c \le 1 \end{cases} \quad \text{(B4)}$$

$$F(c) = \begin{cases} -0.009 + 22.88 \int_0^{1.157c+0.04} t^{1.166}(1-t)^{2.834} dt & if\ 0 \le c \le 0.82989 \\ 1 & if\ 0.82989 < c \le 1 \end{cases} \quad \text{(B5)}$$

$$F(c) = -0.022 + 1.034e^{-\left(\frac{c+0.43238}{0.57212}\right)^{-4.8001}} \quad \text{(B6)}$$

$$F(co) = \begin{cases} 0 & if\ 0 \le c < 0.1008 \\ 13.18c^2 - 2.66c + 0.134 & if\ 0.1008 \le c \le 0.18519 \\ -1.37c^2 + 2.73c - 0.36 & if\ 0.18519 < c \le 0.99934 \\ 1 & if\ 0.99934 < c \le 1 \end{cases} \quad \text{(B7)}$$

$$F(c) = 1.012\left[1 + \left(\frac{c + 0.035198}{0.58009}\right)^{-6.2338}\right]^{-0.33014} \quad for\ 0 \le c \le 1 \quad \text{(B8)}$$

$$F(c) = \begin{cases} 0 & if\ 0 \le c \le 0.042335 \\ \left[1 + \left(\frac{c - 0.042335}{0.3575}\right)^{-6.9303}\right]^{-0.21902} & if\ 0.042335 < c \le 1 \end{cases} \quad \text{(B9)}$$

$$F(c) = 22.141 \int_0^{0.8855c+0.01962} t^{1.0687}(1-t)^{3.0317} dt \ \ for\ 0 \le c \le 1 \quad \text{(B10)}$$

$$F(c) = \int_0^c \frac{(4.463)10^{-5} e^{\frac{-3.5147}{k}}}{\left(\frac{k}{3.5147}\right)^{9.1282}} dk \ \ for\ 0 \le c \le 1 \quad \text{(B11)}$$

$$F(c) = \begin{cases} 0 & if\ 0 \le c < 0.11089 \\ 1.125c - 0.1247 & if\ 0.11089 \le c \le 1 \end{cases} \quad \text{(B12)}$$

$$F(c) = 1.01\left[1 + \left(\frac{c + 0.0034896}{0.61057}\right)^{-6.1958}\right]^{-0.19603} \quad for\ 0 \le c \le 1 \quad \text{(B13)}$$

$$F(c) = \frac{1.010}{1 + \left(\frac{1}{\frac{c + 0.079286}{0.35404}}\right)^{4.1572}} \quad for\ 0 \le c \le 1 \quad \text{(B14)}$$

$$F(c) = 29.27 \int_0^{0.9615c+0.0084} t^{1.750}(1-t)^{2.25} dt \ for\ 0 \le c \le 1 \quad \text{(B15)}$$

$$F(c) = \begin{cases} 0 & if\ 0 \le c \le 0.03676 \\ 1.011\left[1 + \left(\frac{c - 0.03676}{0.41611}\right)^{-4.1837}\right]^{-0.39925} & if\ 0.03676 < c \le 1 \end{cases} \quad \text{(B16)}$$

$$F(c) = \begin{cases} 0 & if\ 0 \le c \le 0.081452 \\ 1.019\left[1 + \left(\frac{c - 0.081452}{0.70933}\right)^{-7.7021}\right]^{-0.14617} & if\ 0.081452 < c \le 1 \end{cases} \quad \text{(B17)}$$

$$F(c) = -0.017 + \frac{1.022}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{t^2}{2}} dt \quad for\ 0 \le c \le 1$$
$$where\ Z = \frac{\left(\frac{c + 0.060074}{0.22331}\right)^{0.5} - \left(\frac{c + 0.060074}{0.22331}\right)^{-0.5}}{0.66179} \tag{B18}$$

$$F(c) = \frac{1.021}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{t^2}{2}} dt\ for\ 0 \le c \le 1$$
$$where\ Z = \frac{\left(\frac{c + 0.18434}{0.52761}\right)^{0.5} - \left(\frac{c + 0.18434}{0.52761}\right)^{-0.5}}{0.40818} \tag{B19}$$

$$F(c) = \frac{1.035}{e^{e^{\frac{0.35947 - c}{0.18994}}}}\ for\ 0 \le c \le 1 \tag{B20}$$

$$F(c) = 1.001 - 1.001e^{\frac{-c^2}{0.14898}} \qquad for\ 0 \le c \le 1 \tag{B21}$$

$$F(c) = 1.006\left[1 + \left(\frac{c + 0.0010638}{0.42538}\right)^{-4.3989}\right]^{-0.25447} \qquad for\ 0 \le c \le 1 \tag{B22}$$

$$F(c) = \begin{cases} A & if\ 0 \le c \le 0.3336 \\ B & if\ 0.3336 < c \le 1 \end{cases}$$
$$A = 0.27c + 4.80c^2 + 113.28c^3 - 3091.13c^4 + 23578.41c^5 - 70802.94c^6 + 74063.73c^7 \tag{B23}$$
$$and\ B = 33.42c - 110.77c^2 + 178.23c^3 - 135.99c^4 + 39.60c^5 - 3.49$$

$$F(c) = \begin{cases} 0.50c - 2.89c^2 + 74.15c^3 - 216.57c^4 + 170.18c^5 & if\ 0 \le c \le 0.3889 \\ -9.66c + 23.85c^2 - 23.39c^3 + 8.19c^4 + 2.01 & if\ 0.3889 < c \le 1 \end{cases} \tag{B24}$$

$$F(c) = 0.54c + 10.70c^2 - 36.63c^3 + 43.84c^4 - 17.45c^5 \quad for\ 0 \le c \le 1 \tag{B25}$$

APPENDIX C.

DATA AND PYTHON CODE FOR THE STATISTICAL AND TIME SERIES VECTOR AUTOREGRESSION PRODUCTIVITY MODEL

Table C.1 Data for the first set of 6 variables.

Date	Gross Productivity ($/man-hour)	Total Compensation ($ per hour worked)	Average Hourly Earnings ($)	Unemployment Rate (%)	Fatalities (in numbers)	GDP ($)
1/1/2006	1056.894	28.65	21.85	9	1255.278	13603.93
2/1/2006	1048.206	28.69293	21.85	8.6	1262.382	13664.78
3/1/2006	1044.798	28.86102	21.85	8.5	1261.296	13708.8
4/1/2006	1013.085	29.11	22.05	6.9	1260.335	13749.81
5/1/2006	1015.113	29.33439	22.03	6.6	1267.633	13786.08
6/1/2006	1006.86	29.46426	22.21	5.6	1271.137	13824.65
7/1/2006	992.9074	29.41	22.29	6.1	1279.799	13867.47
8/1/2006	981.825	29.1274	22.25	5.9	1279.875	13920.44
9/1/2006	976.1741	28.74907	22.41	5.6	1283.001	13979.21
10/1/2006	964.5282	28.46	22.53	4.5	1292.093	14037.23
11/1/2006	968.22	28.37105	22.5	6	1289.998	14095.05
12/1/2006	961.4026	28.44973	22.51	6.9	1297.282	14150.15
1/1/2007	972.0795	28.62	22.59	8.9	1303.483	14208.57
2/1/2007	995.4996	28.80875	22.66	10.5	1300.185	14269.8
3/1/2007	979.0355	28.96985	22.72	9	1287.019	14325.05
4/1/2007	987.5038	29.12	22.83	8.6	1305.638	14382.36
5/1/2007	990.7804	29.22595	22.98	6.9	1290.565	14431.87
6/1/2007	995.3039	29.31047	23.02	5.9	1290.666	14481.7
7/1/2007	994.6377	29.39	23.1	5.9	1293.052	14535
8/1/2007	1007.08	29.48828	23.13	5.3	1279.661	14597.64
9/1/2007	1003.117	29.60409	23.21	5.8	1268.704	14653.01
10/1/2007	1002.74	29.73	23.18	6.1	1260.117	14681.5
11/1/2007	986.2817	29.86944	23.32	6.2	1253.416	14674.7
12/1/2007	974.1951	30.00134	23.43	9.4	1240.834	14654.76
1/1/2008	973.1726	30.12	23.42	11	1221.199	14651.04
2/1/2008	966.5919	30.2102	23.51	11.4	1202.137	14686.53
3/1/2008	975.2069	30.27405	23.58	12	1187.024	14742.83
4/1/2008	986.3708	30.33	23.63	11.1	1169.582	14805.61
5/1/2008	998.5931	30.38443	23.82	8.6	1150.647	14849.16
6/1/2008	993.7385	30.46688	23.89	8.2	1140.33	14862.98
7/1/2008	1003.955	30.6	23.98	8	1109.353	14835.19
8/1/2008	994	30.80567	24.21	8.2	1093.037	14759.17
9/1/2008	1005.803	31.02947	24.23	9.9	1057.465	14656.83
10/1/2008	1010.795	31.2	24.25	10.8	1054.126	14559.54
11/1/2008	1018.855	31.28213	24.41	12.7	1032.659	14482.74

Table C.1 Data for the first set of 6 variables. (Continued).

12/1/2008	1002.295	31.29672	24.56	15.3	1018.19	14430.89
1/1/2009	989.6352	31.29	24.56	18.2	994.254	14394.55
2/1/2009	1001.259	31.29803	24.6	21.4	985.8044	14370.5
3/1/2009	1023.779	31.31903	24.76	21.1	967.3639	14357.81
4/1/2009	1012.606	31.35	24.8	18.7	955.7074	14352.85
5/1/2009	1003.545	31.37915	24.77	19.2	940.5363	14357.55
6/1/2009	1014.074	31.39436	24.85	17.4	924.2447	14378.14
7/1/2009	1020.849	31.38	24.85	18.2	910.3684	14420.31
8/1/2009	1021.127	31.32835	24.89	16.5	904.3094	14489.22
9/1/2009	1016.82	31.26635	24.88	17.1	905.184	14566.45
10/1/2009	1041.731	31.23	25.01	18.7	896.5954	14628.02
11/1/2009	1000.234	31.24098	25.03	19.4	889.7758	14663.9
12/1/2009	1002.073	31.28237	24.99	22.7	882.2238	14686.83
1/1/2010	977.6822	31.33	25.12	24.7	868.9971	14721.35
2/1/2010	1009.765	31.36331	25.25	27.1	864.9468	14783.08
3/1/2010	990.6377	31.39154	25.17	24.9	855.475	14852.31
4/1/2010	976.5654	31.44	25.1	21.8	840.3096	14926.1
5/1/2010	983.7437	31.51238	25.15	20.1	841.4248	14981.87
6/1/2010	983.0746	31.57615	25.12	20.1	837.8137	15030.27
7/1/2010	955.1823	31.58	25.17	17.3	819.9853	15079.92
8/1/2010	950.4578	31.49669	25.17	17	822.1393	15141.42
9/1/2010	934.6278	31.38317	25.19	17.2	819.2775	15201.05
10/1/2010	948.4633	31.32	25.31	17.3	805.9452	15240.84
11/1/2010	952.0005	31.35827	25.31	18.8	813.3458	15254.34
12/1/2010	946.2288	31.46404	25.35	20.7	804.8053	15260.06
1/1/2011	927.2099	31.59	25.46	22.5	799.1726	15285.83
2/1/2011	916.547	31.68715	25.44	21.8	802.8194	15349.92
3/1/2011	922.8337	31.74286	25.34	20	795.0442	15424.01
4/1/2011	917.2929	31.77	25.38	17.8	794.2954	15496.19
5/1/2011	909.2263	31.77557	25.35	16.3	780.3985	15536.75
6/1/2011	932.583	31.82988	25.35	15.6	782.4489	15562
7/1/2011	924.0165	32.01	25.36	13.6	774.9769	15591.85
8/1/2011	950.9169	32.3702	25.46	13.5	770.1903	15645.35
9/1/2011	942.9655	32.78399	25.49	13.3	769.8317	15717.62
10/1/2011	940.8893	33.08	25.49	13.7	769.0893	15796.46
11/1/2011	944.518	33.16779	25.44	13.1	779.3141	15878.55
12/1/2011	947.4409	33.12414	25.49	16	779.6418	15953.03
1/1/2012	952.7736	33.08	25.5	17.7	788.9539	16019.76
2/1/2012	947.4198	33.13834	25.55	17.1	795.8784	16072.7

Table C.1 Data for the first set of 6 variables. (Continued).

3/1/2012	948.6982	33.25505	25.66	17.2	796.4462	16113.18
4/1/2012	964.0777	33.38	25.65	14.5	806.8653	16152.26
5/1/2012	981.9633	33.452	25.69	14.2	807.0543	16189.86
6/1/2012	986.8385	33.48967	25.71	12.8	812.9467	16226.69
7/1/2012	988.2228	33.52	25.75	12.3	822.1695	16257.15
8/1/2012	987.2418	33.5735	25.77	11.3	829.4309	16282.73
9/1/2012	981.1104	33.66954	25.86	11.9	836.4394	16312.85
10/1/2012	988.7591	33.82	25.84	11.4	848.3528	16358.86
11/1/2012	983.2946	34.03805	25.95	12.2	837.3026	16430.63
12/1/2012	968.0937	34.25875	25.95	13.5	848.2236	16506.91
1/1/2013	954.2513	34.43	25.96	16.1	855.9949	16569.59
2/1/2013	956.5025	34.492	26.03	15.7	849.8537	16600.39
3/1/2013	948.4274	34.46479	25.98	14.7	858.006	16614.52
4/1/2013	969.6409	34.37	26.03	13.2	855.509	16637.93
5/1/2013	968.7046	34.24836	26.04	10.8	851.477	16686.12
6/1/2013	981.2357	34.16154	26.09	9.8	848.2184	16760.35
7/1/2013	1007.58	34.19	26.17	9.1	849.2733	16848.75
8/1/2013	1010.471	34.37867	26.19	9.1	854.9707	16947.33
9/1/2013	1014.721	34.63348	26.17	8.5	847.2871	17033.29
10/1/2013	1030.049	34.82	26.18	9	849.3723	17083.14
11/1/2013	1028.275	34.86339	26.25	8.6	855.6744	17089.3
12/1/2013	1056.552	34.82572	26.38	11.4	855.7703	17082.31
1/1/2014	1057.266	34.81	26.41	12.3	865.2	17104.56
2/1/2014	1061.364	34.89259	26.76	12.8	865.2018	17187.58
3/1/2014	1038.963	35.00762	26.48	11.3	868.0346	17298.96
4/1/2014	1054.835	35.1	26.56	9.4	875.5426	17432.91
5/1/2014	1052.84	35.1084	26.61	8.6	879.8082	17548.53
6/1/2014	1046.381	35.12471	26.67	8.2	882.8547	17647.26
7/1/2014	1040.541	35.27	26.66	7.5	895.7764	17721.66
8/1/2014	1039.491	35.62572	26.73	7.7	907.8311	17777.13
9/1/2014	1038.056	36.04185	26.81	7	912.9484	17817.5
10/1/2014	1063.548	36.31	26.86	6.4	923.9882	17849.91
11/1/2014	1057.089	36.31433	26.92	7.5	928.2045	17884.4
12/1/2014	1064.644	36.16775	26.87	8.3	935.0578	17925.82
1/1/2015	1076.194	36.04	27.07	9.8	938.0063	17984.18
2/1/2015	1064.354	36.07458	27.12	10.6	942.6191	18061.96
3/1/2015	1096.528	36.21601	27.24	9.5	957.0795	18139.51
4/1/2015	1116.61	36.42	27.28	7.5	950.2514	18219.41
5/1/2015	1137.997	36.60245	27.32	6.7	968.5878	18279.36

Table C.1 Data for the first set of 6 variables. (Continued).

6/1/2015	1145.991	36.75752	27.34	6.3	961.0014	18321.7
7/1/2015	1139.681	36.87	27.36	5.5	967.8246	18344.71
8/1/2015	1139.13	36.9446	27.44	6.1	967.9606	18352.33
9/1/2015	1147.384	36.98457	27.35	5.5	978.2144	18351.33
10/1/2015	1112.092	37	27.52	6.2	970.044	18350.83
11/1/2015	1111.631	37.00106	27.67	6.2	988.7475	18359.12
12/1/2015	1096.805	36.993	27.62	7.5	988.7305	18381.34
1/1/2016	1123.488	36.98	27.66	8.5	1001.263	18424.28
2/1/2016	1118.095	36.97888	27.76	8.7	996.1386	18489.7
3/1/2016	1147.921	37.04518	27.88	8.7	999.9664	18561.79
4/1/2016	1133.942	37.26	27.98	6	1004.742	18637.25
5/1/2016	1142.106	37.63933	28.08	5.2	1009.632	18699
6/1/2016	1159.916	38.06763	28.13	4.6	1016.089	18754.51
7/1/2016	1152.297	38.34	28.2	4.5	1015.891	18806.74
8/1/2016	1161.606	38.35346	28.2	5.1	1024.155	18864.76
9/1/2016	1157.527	38.22555	28.23	5.2	1022.944	18927.43
10/1/2016	1153.921	38.15	28.4	5.7	1024.918	18991.88
11/1/2016	1182.149	38.25906	28.35	5.7	1041.355	19061.14
12/1/2016	1191.433	38.48486	28.42	7.4	1038.818	19127.25
1/1/2017	1175.746	38.73	28.53	9.4	1038.778	19190.43
2/1/2017	1181.436	38.90284	28.52	8.8	1033.655	19246.24
3/1/2017	1180.228	39.02445	28.61	8.4	1029.173	19295.52
4/1/2017	1163.955	39.18	28.6	6.3	1034.148	19356.65
5/1/2017	1166.649	39.38015	28.72	5.3	1034.281	19428.55
6/1/2017	1159.785	39.54161	28.9	4.5	1017.811	19516
7/1/2017	1158.208	39.53	28.95	4.9	1016.124	19611.7
8/1/2017	1154.133	39.27294	29	4.7	1016.551	19718.91
9/1/2017	1157.314	38.92374	29.18	4.7	1013.485	19825.69
10/1/2017	1142.033	38.7	29.11	4.5	1017.023	19918.91
11/1/2017	1161.543	38.73103	29.22	5	1010.925	19999.31
12/1/2017	1165.214	38.91283	29.29	5.9	1013.564	20073.51
1/1/2018	1172.106	39.09	29.39	7.3	1011.107	20163.16
2/1/2018	1179.871	39.13549	29.55	7.8	1002.615	20276.99
3/1/2018	1159.125	39.12464	29.49	7.4	1011.096	20390.27
4/1/2018	1158.714	39.17	29.66	6.5	1017.742	20510.18
5/1/2018	1161.529	39.34327	29.7	4.4	1015.684	20607.15
6/1/2018	1154.116	39.55479	29.78	4.7	1017.226	20687.57
7/1/2018	1147.978	39.65	29.94	3.4	1024.692	20749.75
8/1/2018	1140.891	39.54054	30.02	3.4	1026.287	20802.14

Table C.1 Data for the first set of 6 variables. (Continued).

9/1/2018	1150.164	39.33194	30.18	4.1	1027.771	20849.49
10/1/2018	1112.676	39.19	30.24	3.6	1028.472	20897.8
11/1/2018	1113.488	39.22088	30.3	3.9	1037.414	20956.42
12/1/2018	1086.992	39.36979	30.44	5.1	1038.172	21022.19
1/1/2019	1078.322	39.55	30.32	6.4	1038.14	21098.83
2/1/2019	1118.833	39.68359	30.45	6.2	1047.976	21182.6
3/1/2019	1107.502	39.78438	30.5	5.2	1034.191	21259.64
4/1/2019	1116.124	39.92	30.63	4.7	1045.154	21340.27
5/1/2019	1109.362	40.10366	30.7	3.2	1053.416	21409.5
6/1/2019	1090.597	40.30446	30.74	4	1048.522	21476.27
7/1/2019	1100.223	40.45	30.75	3.8	1047.81	21542.54
8/1/2019	1103.718	40.50569	30.87	3.6	1044.606	21615.99
9/1/2019	1100.784	40.49086	30.87	3.2	1046.396	21684.67
10/1/2019	1114.17	40.45	30.98	4	1052.47	21734.27
11/1/2019	1127.915	40.42238	31.09	4.4	1042.919	21755.32
12/1/2019	1127.24	40.45	31.14	5	1032.507	21734.27

Table C.2 Data for the second set of 6 variables.

Date	Hires (in thousands)	Occupational Injuries and Illnesses (in numbers)	Gross Job Gains (in thousands)	Gross Job Losses (in thousands)	Job Openings (in thousands)	Turnover (in thousands)
1/1/2006	453	413.0758	942	764	156	556
2/1/2006	387	418.1069	903.3244	787.7211	146	511
3/1/2006	502	420.4098	872.1328	812.1707	185	504
4/1/2006	646	419.3637	843	839	222	629
5/1/2006	629	419.6343	821.4564	861.1734	182	468
6/1/2006	540	421.6188	807.3961	876.3771	185	388
7/1/2006	541	421.1558	803	880	199	379
8/1/2006	418	419.1869	808.6112	870.5323	202	370
9/1/2006	404	415.9535	822.436	853.9528	174	270
10/1/2006	426	415.6521	841	839	157	361
11/1/2006	351	416.0448	862.0584	831.0408	105	325
12/1/2006	269	413.6727	878.0884	830.1155	103	412
1/1/2007	389	408.7923	883	834	242	366

Table C.2 Data for the second set of 6 variables. (Continued).

2/1/2007	262	412.0494	871.3053	840.9861	231	428
3/1/2007	507	404.9561	850.7359	849.6815	191	375
4/1/2007	566	404.6202	824	862	221	369
5/1/2007	523	405.0592	800.649	875.9471	191	352
6/1/2007	527	397.4354	783.6897	888.6778	189	394
7/1/2007	483	397.0124	778	895	208	303
8/1/2007	421	393.1	784.8066	892.3941	178	267
9/1/2007	389	392.043	796.831	883.9558	144	284
10/1/2007	410	386.9473	804	875	161	282
11/1/2007	281	384.2113	799.7831	869.4969	107	215
12/1/2007	227	381.4332	787.9593	870.8504	106	253
1/1/2008	337	374.8058	774	882	145	279
2/1/2008	315	376.3804	763.1892	903.3227	121	274
3/1/2008	400	371.3954	755.2577	923.4601	108	269
4/1/2008	548	365.2331	747	933	122	311
5/1/2008	456	356.9832	737.5235	923.2811	189	370
6/1/2008	456	354.4786	725.9083	906.2426	145	326
7/1/2008	459	346.1938	713	900	132	349
8/1/2008	398	346.5615	697.5727	916.4747	88	389
9/1/2008	317	337.87	677.8028	947.2114	115	304
10/1/2008	396	335.5281	652	977	71	288
11/1/2008	268	328.4604	618.4219	996.6109	52	267
12/1/2008	204	322.2353	588.8944	1004.372	46	288
1/1/2009	325	316.2489	574	1003	38	208
2/1/2009	284	312.1471	582.9003	993.5528	73	298
3/1/2009	332	304.3975	602.1405	976.7982	48	281
4/1/2009	448	300.0802	621	947	29	329
5/1/2009	409	293.1755	625.9415	906.6648	51	321
6/1/2009	333	281.382	622.008	860.8727	67	315
7/1/2009	383	281.7317	618	822	50	406
8/1/2009	260	274.7731	619.9161	794.837	65	406
9/1/2009	304	263.8404	623.7923	778.6319	66	455
10/1/2009	316	265.8393	623	769	54	442
11/1/2009	239	259.1593	614.026	761.043	49	326
12/1/2009	198	249.784	605.9235	752.2174	55	344
1/1/2010	271	244.9652	610	739	54	269
2/1/2010	236	239.1259	633.8469	719.8897	65	283
3/1/2010	419	235.6429	662.1586	700.2406	86	237
4/1/2010	487	228.3155	685	680	101	268

Table C.2 Data for the second set of 6 variables. (Continued).

5/1/2010	388	225.5568	687.1867	665.6885	88	314
6/1/2010	335	219.6246	675.5726	658.8643	83	351
7/1/2010	400	216.9353	662	662	113	314
8/1/2010	334	210.9531	654.3041	675.2736	59	380
9/1/2010	301	212.4792	650.6233	690.1664	78	433
10/1/2010	357	203.7425	646	696	67	368
11/1/2010	254	205.1206	637.0498	686.2236	66	359
12/1/2010	213	200.2139	628.7477	667.9345	32	331
1/1/2011	250	202.1962	627	650	61	300
2/1/2011	290	198.9177	636.7864	640.6076	45	324
3/1/2011	379	196.4113	652.9018	636.9737	67	305
4/1/2011	504	195.5508	674	633	91	279
5/1/2011	497	194.6532	692.6708	625.8822	124	272
6/1/2011	476	193.3722	703.7197	619.8896	77	258
7/1/2011	402	196.1443	700	622	94	313
8/1/2011	336	198.969	678.4884	636.028	111	355
9/1/2011	361	193.1778	651.8544	653.0892	87	335
10/1/2011	331	196.8098	637	661	82	371
11/1/2011	241	192.4275	643.6901	652.5741	63	401
12/1/2011	179	197.8606	661.0367	636.4633	56	368
1/1/2012	298	193.9435	674	624	85	386
2/1/2012	291	193.2695	670.17	624.6709	61	321
3/1/2012	315	193.7027	657.4193	632.0374	94	251
4/1/2012	380	193.7529	645	638	134	343
5/1/2012	412	190.5859	642.3851	635.2832	105	308
6/1/2012	427	189.1838	647.3578	626.4274	70	289
7/1/2012	439	189.8211	655	617	95	242
8/1/2012	333	190.8171	661.9601	610.4692	134	306
9/1/2012	348	184.3637	667.7797	606.9102	92	256
10/1/2012	317	189.7742	673	605	107	326
11/1/2012	291	185.4342	678.2311	603.7572	73	335
12/1/2012	176	186.7479	681.7665	604.2494	93	376
1/1/2013	304	193.2036	682	608	116	480
2/1/2013	321	188.2274	677.8429	615.2264	121	468
3/1/2013	353	190.9694	672.695	620.5033	121	471
4/1/2013	402	192.2712	669	619	133	427
5/1/2013	433	191.9278	669.7135	607.2105	137	355
6/1/2013	407	192.6297	672.0879	593.6054	146	336
7/1/2013	363	196.7599	672	591	111	376

Table C.2 Data for the second set of 6 variables. (Continued).

8/1/2013	298	195.4925	666.8209	606.6841	129	316
9/1/2013	303	202.146	659.8507	628.7669	121	300
10/1/2013	368	203.0408	656	640	141	295
11/1/2013	233	199.5572	658.5464	629.9195	117	329
12/1/2013	148	200.2909	666.1657	609.1527	86	384
1/1/2014	259	203.4374	677	592	152	383
2/1/2014	240	203.9252	688.1879	590.5788	122	392
3/1/2014	268	203.0733	695.5381	597.2697	130	471
4/1/2014	406	204.3851	697	603	137	537
5/1/2014	412	201.6348	689.757	599.397	137	504
6/1/2014	341	203.708	678.0871	590.7391	171	379
7/1/2014	394	205.2303	669	585	158	368
8/1/2014	334	204.8846	666.4297	587.5523	148	368
9/1/2014	282	199.9705	667.4846	596.6746	105	353
10/1/2014	322	198.9889	667	608	156	311
11/1/2014	241	199.8284	661.936	618.6021	104	339
12/1/2014	217	200.2463	657.6221	625.0268	100	342
1/1/2015	322	199.5268	661	625	150	378
2/1/2015	278	200.2984	676.1644	617.0048	146	400
3/1/2015	318	200.7605	691.6072	607.0414	177	389
4/1/2015	470	201.8227	698	599	179	459
5/1/2015	407	202.8271	685.7037	598.3255	184	423
6/1/2015	416	205.7316	665.8096	601.4882	160	439
7/1/2015	361	203.2301	656	603	156	402
8/1/2015	322	205.1545	667.0546	599.2529	168	388
9/1/2015	321	205.8774	688.663	593.5276	106	339
10/1/2015	332	206.4191	704	591	145	349
11/1/2015	260	205.7988	702.3633	595.3497	89	357
12/1/2015	180	204.1943	689.1034	605.4638	115	292
1/1/2016	261	205.6088	673	620	161	271
2/1/2016	295	203.2047	662.4842	635.9626	188	331
3/1/2016	351	206.1694	658.5923	647.3717	219	297
4/1/2016	494	203.6163	660	650	196	400
5/1/2016	424	200.4804	665.3731	640.1673	188	371
6/1/2016	340	204.9147	670.5166	625.5759	186	444
7/1/2016	408	206.583	670	618	244	492
8/1/2016	337	203.6869	661.1731	624.0577	205	445
9/1/2016	304	198.6108	652.0139	634.8316	240	499
10/1/2016	351	206.2531	653	637	221	409

Table C.2 Data for the second set of 6 variables. (Continued).

11/1/2016	255	199.5149	670.6801	622.394	176	370
12/1/2016	247	205.1216	695.1312	602.1088	138	345
1/1/2017	363	206.2146	715	590	150	320
2/1/2017	306	202.731	718.9298	597.8685	179	365
3/1/2017	379	203.4095	709.0146	616.6448	186	325
4/1/2017	545	201.8673	686	640	229	309
5/1/2017	505	200.2184	656.3367	656.4176	184	352
6/1/2017	449	201.7813	630.7592	664.0084	220	410
7/1/2017	433	197.3331	625	661	247	363
8/1/2017	387	199.4471	647.6463	647.1514	226	498
9/1/2017	380	198.6394	685.7031	627.4991	178	497
10/1/2017	399	199.6062	719	609	203	428
11/1/2017	268	201.6376	735.3923	595.8214	212	525
12/1/2017	179	198.1586	736.6447	590.8303	180	347
1/1/2018	326	197.1742	730	595	245	350
2/1/2018	331	194.9667	721.024	608.7444	198	327
3/1/2018	359	197.5757	710.6732	625.4122	234	391
4/1/2018	510	198.4253	695	643	258	289
5/1/2018	514	196.1746	675.5204	654.4988	279	317
6/1/2018	447	198.9234	657.7336	659.3805	323	304
7/1/2018	450	196.3109	651	657	314	323
8/1/2018	381	194.3102	660.2249	647.5532	315	328
9/1/2018	353	201.5932	678.5426	634.1396	299	389
10/1/2018	362	199.6757	695	621	278	427
11/1/2018	290	199.8214	702.792	610.8965	279	483
12/1/2018	221	201.4901	702.6295	607.587	299	472
1/1/2019	394	201.4053	698	614	313	525
2/1/2019	308	201.5535	692.0263	631.5232	287	441
3/1/2019	358	197.4927	686.9573	650.9797	364	475
4/1/2019	593	203.2864	683	668	434	348
5/1/2019	499	203.4239	681.6933	673.7878	376	319
6/1/2019	491	199.4174	682.1143	671.9231	331	339
7/1/2019	432	198.3338	683	668	360	414
8/1/2019	411	200.3989	683.3898	666.2734	384	319
9/1/2019	431	199.6764	683.286	666.7332	327	476
10/1/2019	493	200.0679	683	668	326	312
11/1/2019	325	197.905	682.8067	668.8561	217	392
12/1/2019	278	197.6653	683	668	239	487

Note: all the values of the variables were taken to be at the first of the month so that to ensure consistency between all variables. In addition, since the 'Fatalities (in numbers)' variable and the 'Occupational Injuries and Illnesses (in numbers)' variable are initially reported per year from the source publishing this data, these variables were transformed to monthly data using interpolation and then a noise distribution was added to them to account for some variability.

To use this data, please make sure to aggregate it so that the first column is the 'Date' and the other variables are stacked afterwards in columns in the following order: Gross Productivity ($/man-hour), Total Compensation ($ per hour worked), Average Hourly Earnings ($), Unemployment Rate (%), Fatalities (in numbers), GDP ($), Hires (in thousands), Occupational Injuries and Illnesses (in numbers), Gross Job Gains (in thousands), Gross Job Losses (in thousands), Job Openings (in thousands), and Turnover (in thousands). To be able to run the Python code below, this aggregated data needs to be saved in a csv (comma-separated values) file named 'Combined_Data.csv'.

It is worth mentioning that the implemented Python code was inspired and based on Prabhakaran (2019), but it was tailored to accommodate for the needs and scope of the research in this dissertation. The implemented Python code (the code's comments are shown as underlined) for the statistical and times series model is as follows:

```
#importing all needed packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels
```

```
from statsmodels.tsa.stattools import adfuller

from statsmodels.tsa.vector_ar.vecm import coint_johansen

from statsmodels.tsa.api import VAR

from statsmodels.tsa.stattools import grangercausalitytests

from statsmodels.stats.stattools import durbin_watson

'''

Note: the following versions were used for the different packages:

Version of Pandas is: 1.1.5

Version of Numpy is: 1.20.1

Version of Pandas is: 0.12.2

'''

#reading the data

Data = pd.read_csv('Combined_Data.csv',index_col=0)

#transforming the index to date format (i.e., datetime)

Data.index = pd.to_datetime(Data.index)

#plotting the data

fig, axes = plt.subplots(nrows=4, ncols=3, dpi=300, figsize=(12,8))

for i, ax in enumerate(axes.flatten()):

    ax.plot(Data[Data.columns[i]], color='red', linewidth=1) #plotting each column

    ax.set_title(Data.columns[i]) #setting the title

    ax.tick_params(labelsize=10) #setting the size of labels

plt.tight_layout()
```

```
#since we have a total of 168 readings (which is considered to be a low sample size), the training size was taken to be around 90% and the testing size was taken to be around 10%
#this means that around 0.1*168=16.8 months will be used for testing. Therefore, 17 months were used for testing, which means starting on 2018-08-01 and ending on 2019-12-01
#hence, we need to get the index of the row that starts with 2018-08-01, this will be used to split the data between training and testing
for i in range(Data.shape[0]):
    if (Data.iloc[i][0]==Data.loc['2018-08-01'][0]) & (Data.iloc[i][1]==Data.loc['2018-08-01'][1]):
        training_size=i
#getting the testing size
testing_size=Data.shape[0]-training_size
#dividing the data between training and testing sets
df_train, df_test = Data[0:-testing_size], Data[-testing_size:]
#defining a function to check the stationarity of the data using the ADF statistica test
def adfuller_test(series, signif=0.05, name='', verbose=False):
    """This function performs ADF test to check the stationarity of given series, and it prints a report"""
    r = adfuller(series)
    output = {'test_statistic':round(r[0], 4), 'pvalue':round(r[1], 4), 'n_lags':round(r[2], 4), 'n_obs':r[3]}
```

```
    p_value = output['pvalue']
    def adjust(val, length= 6): return str(val).ljust(length)
    # Print Summary
    print(f'    Augmented Dickey-Fuller Test on "{name}"', "\n   ", '-'*47)
    print(f' Null Hypothesis: Data has unit root. Non-Stationary.')
    print(f' Significance Level    = {signif}')
    print(f' Test Statistic        = {output["test_statistic"]}')
    print(f' No. Lags Chosen       = {output["n_lags"]}')
    for key,val in r[4].items():
        print(f' Critical value {adjust(key)} = {round(val, 3)}')
    if p_value <= signif:
        print(f" => P-Value = {p_value}. Rejecting Null Hypothesis.")
        print(f" => Series is Stationary.")
    else:
        print(f" => P-Value = {p_value}. Weak evidence to reject the Null Hypothesis.")
        print(f" => Series is Non-Stationary.")
#applying the ADF test by calling the previous function
for name, column in df_train.iteritems():
    adfuller_test(column, name=column.name)
    print('\n')
#the ADF test shows that some of the variables are not stationary, therefore we need to
difference all of them and check again for stationarity
df_train_differenced_order_1 = df_train.diff().dropna() #first difference
```

```
#applying the ADF test on the differenced (order =1) data
for name, column in df_train_differenced_order_1.iteritems():
    adfuller_test(column, name=column.name)
    print('\n')
#some variables are still not stationary, we have to difference once again
df_train_differenced_order_2 = df_train_differenced_order_1.diff().dropna() #second
difference
#applying the ADF test on the differenced (order =2) data
for name, column in df_train_differenced_order_2.iteritems():
    adfuller_test(column, name=column.name)
    print('\n')
#applying the granger causality test with Gross Productivity ($/man-hour) being the
response variable and all other variables being the predictors
P_values=pd.DataFrame() #initializing a dataframe to store the different p_values
#choosing the maximum desired lag
maxlag=24 # I chose 24 because it spans for 24 months or 2 years
for c in df_train_differenced_order_2.columns: #looping over all columns
    #retrieving the p-value and rounding it to 4 decimal places
    P_values[c]=[round(grangercausalitytests(df_train_differenced_order_2[['Gross
Productivity ($/man-
hour)',c]],maxlag=maxlag,verbose=False)[i+1][0]['ssr_chi2test'][1],4)for i in range
(maxlag)]
#changing the indices' names to indicate the lag value
```

```
for i in range (P_values.shape[0]):
    P_values=P_values.rename(index={i:'Lag'+str(i+1)})
#changing the columns names to indicate which variable is the response and which
variables are the predictors
for j in range (1,P_values.shape[1]):
   P_values=P_values.rename(columns={P_values.columns[j]:P_values.columns[j]+'_x'})
P_values=P_values.rename(columns={'Gross Productivity ($/man-hour)':'Gross
Productivity ($/man-hour)_y'})
#getting the minimum p value for each variable so that we include all relevant predictor
variable that could have either short-term or long-term impact on the response variable
P_values_min=pd.DataFrame(P_values.min()).T.rename(index={0:'Min P_All Lags'})
#if the p-value obtained from the test is greater than the significance level of 0.05, then
there is no causality
Irrelevant_predictors_indices= [] #initializing a list that will store the indices of the
irrelevant predictors
for i in range(1,P_values_min.T.shape[0]): #starting from 1 because we do not want to
drop the output variable which present at index 0
    if P_values_min.T.iloc[i,0] > 0.05:
        Irrelevant_predictors_indices.append(i)
#dropping the irrelevant predictors from the differenced training data and from the testing
data
df_train_differenced_order_2.drop(Data.columns[Irrelevant_predictors_indices],axis=1,
inplace=True)
```

```
df_test.drop(Data.columns[Irrelevant_predictors_indices],axis=1, inplace=True)
#having a function that would perform Johanson's cointegration test calculations
def cointegration_test(df, Number_of_lagged_differences,alpha=0.05):
    """This function performs Johanson's Cointegration Test and reports a summary"""
    out = coint_johansen(df,-1,Number_of_lagged_differences)
    d = {'0.90':0, '0.95':1, '0.99':2}
    traces = out.lr1
    cvts = out.cvt[:, d[str(1-alpha)]]
    def adjust(val, length= 6): return str(val).ljust(length)
    # Summary
    print('Rank   ::  Test Stat > C(95%)    =>   Cointegration?  \n', '--'*20)
    for rank, trace, cvt in zip(range(0,len(df.columns)), traces, cvts):
        if rank==0:
            print('r =',rank, ' :: ', adjust(round(trace,2), 9), ">", adjust(cvt, 8), ' =>  ' , trace >
cvt)
        else:
            print('r <=',rank, ':: ', adjust(round(trace,2), 9), ">", adjust(cvt, 8), ' =>  ' , trace >
cvt)
#according to Giles (2011), we should set up a VAR model in the levels of the data,
regardless of the orders of integration of the various time-series. Therefore, we must not
difference the data for the VAR model for the cointegration purposes
model_to_determine_number_of_lagged_differences_for_cointegration=
VAR(df_train.drop(Data.columns[Irrelevant_predictors_indices],axis=1, inplace=False))
```

```
#getting the optimal order for the VAR model corresponding to the lowest AIC (Akaike information criterion) according to Faghih and Kashani (2018) and Giles (2011)
Number_of_lagged_differences=model_to_determine_number_of_lagged_differences_for_cointegration.select_order().selected_orders['aic']
#we need to confirm that for the selected lag length, the residuals of the VAR model are not correlated. If they are not, then we can proceed. If they are, then we may have to modify the lag length. In other words, if the residuals are correlated then we might increase the lagged differences until any autocorrelation issues are resolved according to Giles (2011)
#for that purpose, we will use Durbin Watson statistic
#fitting the VAR model for cointegration purposes
model_fitted_to_determine_number_of_lagged_differences_for_cointegration = model_to_determine_number_of_lagged_differences_for_cointegration.fit(model_to_determine_number_of_lagged_differences_for_cointegration.select_order().selected_orders['aic'])
#getting the summary
model_fitted_to_determine_number_of_lagged_differences_for_cointegration.summary()
def adjust(val, length= 6): return str(val).ljust(length)
for col, val in zip(Data.columns, durbin_watson(model_fitted_to_determine_number_of_lagged_differences_for_cointegration.resid)):
    print(adjust(col), ':', round(val, 2))
```

```
#we are particularly interested with the statistic of the response variable which is
'Gross Productivity ($/man-hour)'
#since the obtained statistic is 2.03 which is close to the value of 2, then we can conclude
that there is no significant serial correlation (which is a good thing). Then we can proceed
with the chosen lag and with the Number_of_lagged_differences as well
#calling the Johanson's cointegration test function
#note that we always need to do the cointegration test on the initial data and not the
differenced one
cointegration_test(df_train.drop(Data.columns[Irrelevant_predictors_indices],axis=1,
inplace=False),Number_of_lagged_differences=Number_of_lagged_differences)
#to select the best order of the FINAL productivity VAR model, we shall iteratively fit
increasing orders and choose the order that gives a model with least BIC.
model = VAR(df_train_differenced_order_2)
x = model.select_order()
x.summary()
#getting the optimal order for the VAR model corresponding to the lowest BIC.
best_order=x.selected_orders['bic']
#fitting the VAR model
model_fitted = model.fit(best_order)
#getting the summary
model_fitted.summary()
#note that we do not need all the fitted model, we only need the one for the Gross
Productivity ($/man-hour)
```

```
#we need to check for serial correlation of residuals using Durbin Watson statistic to ensure that the model is sufficiently able to explain the variances and patterns in the time series.
out = durbin_watson(model_fitted.resid)
def adjust(val, length= 6): return str(val).ljust(length)
for col, val in zip(Data.columns, out):
    print(adjust(col), ':', round(val, 2))
#we are particularly interested with the statistic of the response variable which is 'Gross Productivity ($/man-hour)'
#since the obtained statistic is 2.2 which is close to the value of 2, then we can conclude that there is no significant serial correlation (which is a good thing)
#we need to use the fitted VAR model to forecast on the unseen testing set
lag_order = model_fitted.k_ar #getting the lag order of the fitter VAR model (note: this returns a value of 2)
#defining the input data for our forecast which should be the last 'lag_order' months of the differenced training set
forecast_input = df_train_differenced_order_2.values[-lag_order:]
#forecasting
fc = model_fitted.forecast(y=forecast_input, steps=testing_size)
#transforming it to a dataframe
df_forecast = pd.DataFrame(fc, index=df_test.index, columns=df_test.columns + '_2d')
#defining a function to invert the transformation
def invert_transformation(df_train, df_forecast, second_diff=False):
```

```
    """Revert back the differencing to get the forecast to original scale.
    To be noted is that df_train shall be the original training data (with no differencing) but dimensionally reduced (in terms of columns) based on the selected predictors from the Granger Causality test.
    On the other hand, the df_forecast shall be the predicted differenced data"""
    df_fc = df_forecast.copy()
    columns = df_train.columns
    for col in columns:
        # Roll back 2nd Diff
        if second_diff:
            df_fc[str(col)+'_1d'] = (df_train[col].iloc[-1]-df_train[col].iloc[-2]) + df_fc[str(col)+'_2d'].cumsum()
        # Roll back 1st Diff
        df_fc[str(col)+'_forecast'] = df_train[col].iloc[-1] + df_fc[str(col)+'_1d'].cumsum()
    return df_fc
#calling the previous function to invert the forecasts to the original scale (with no differencing)
df_results = invert_transformation(df_train.drop(Data.columns[Irrelevant_predictors_indices],axis=1, inplace=False), df_forecast, second_diff=True)
#we want to plot the predicted vs actual values
```

```
df_results=df_results.rename(columns={'Gross Productivity ($/man-
hour)_forecast':'Predicted'}) #changing the name of the response variable for the
predicted results
df_test=df_test.rename(columns={'Gross Productivity ($/man-hour)':'Actual'}) #changing
the name of the response variable based on actual values in the testing set
ax=pd.concat([df_results['Predicted'],df_test['Actual']],axis=1).plot(figsize=(12,6),fontsiz
e=12) #concatinating the predicted and actual values into a dataframe and then plotting
ticklabels = [item.strftime("%b\n%Y") for item in
pd.concat([df_results['Predicted'],df_test['Actual']],axis=1).index] #specifying the format
of the tick labels as needed
plt.xticks(pd.concat([df_results['Predicted'],df_test['Actual']],axis=1).index,ticklabels,font
size=12)
plt.ylabel('Gross Productivity ($/man-hour)',fontsize=14) #setting the y label
plt.xlabel('') #we do not want an x label
ax.set_ylim((0,2000)) #setting limits of the y-axis
plt.grid(which='both',color='silver', linestyle='--', linewidth=1) #showing and formatting
the grids
plt.legend(fontsize=14) #formatting the font size of the legend
plt.show()
#defining a function so calculate the different evaluation metrics for the predicted vs
actual values
def forecast_accuracy(forecast, actual):
    mape = np.mean(np.abs(forecast - actual)/np.abs(actual))  # MAPE
```

```
    me = np.mean(forecast - actual)           # ME
    mae = np.mean(np.abs(forecast - actual))    # MAE
    mpe = np.mean((forecast - actual)/actual)   # MPE
    rmse = np.mean((forecast - actual)**2)**.5  # RMSE
    return({'mape':mape, 'me':me, 'mae': mae,
            'mpe': mpe, 'rmse':rmse})
print('Evaluation metrics are as follows:')
#calling the previous function
accuracy_prod = forecast_accuracy(df_results['Predicted'].values,
df_test['Actual'].values)
def adjust(val, length= 6): return str(val).ljust(length)
for k, v in accuracy_prod.items():
    print(adjust(k), ': ', round(v,4))
'''
Note: the results shall print as follows
Evaluation metrics are as follows:
mape   :  0.04
me     :  -40.2963
mae    :  44.5549
mpe    :  -0.0361
rmse   :  56.1412
'''
```

APPENDIX D.

DATA, PYTHON CODE, AND R CODE FOR THE UNSUPERVISED SAFETY MODEL

Table D.1 Data for the first set of 10 fatality causes.

Damaged, defective, or malfunctioning equipment	Lack of necessary equipment	Poor equipment handling and not following proper operation procedures or manufacturer's specs	Use of nonsuitable equipment	Equipment/spoil on excavation edge	Vibration in excavation	Equipment tipping	Poor labeling	Improper use of PPE	No PPE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

Table D.1 Data for the first set of 10 fatality causes. (Continued).

FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE

Table D.1 Data for the first set of 10 fatality causes. (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.1 Data for the first set of 10 fatality causes. (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE

Table D.1 Data for the first set of 10 fatality causes. (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table D.2 Data for the second set of 10 fatality causes (i.e., causes 11 to 20).

No fall arrest system, guardrails, or safety nets	No cable insulation	Not wearing seatbelt	Safety element failure	No/damaged Cave in protection	No protection from traffic	No decking	Lack of employee knowledge	Lack of coordination of site activities	Poor tool handling
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.2 Data for the second set of 10 fatality causes (i.e., causes 11 to 20). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.2 Data for the second set of 10 fatality causes (i.e., causes 11 to 20). (Continued).

TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.2 Data for the second set of 10 fatality causes (i.e., causes 11 to 20). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.2 Data for the second set of 10 fatality causes (i.e., causes 11 to 20). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

Table D.3 Data for the third set of 10 fatality causes (i.e., causes 21 to 30).

Inappropriate tools used	Poor material handling	Poor storage	High exposure to chemical	Wet material	Employer gross negligence	Employer allowed employee to work in an unsafe environment	Lack of knowledge by employer about site conditions	Lack of clear employer instructions	Utility mislocating
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE

Table D.3 Data for the third set of 10 fatality causes (i.e., causes 21 to 30). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.3 Data for the third set of 10 fatality causes (i.e., causes 21 to 30). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table D.3 Data for the third set of 10 fatality causes (i.e., causes 21 to 30). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.3 Data for the third set of 10 fatality causes (i.e., causes 21 to 30). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.4 Data for the fourth set of 10 fatality causes (i.e., causes 31 to 40).

Lack of preventive action	No first aid personnel	Lack of supervision and absence of competent or necessary personnel	Vehicle observer error	Operation carried out by noncompetent individual	Employee misconduct	Willfully exposing self to hazardous situation	Misjudgement of hazardous situation	Lack of specific on the job training	Lack of general health and safety training
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table D.4 Data for the fourth set of 10 fatality causes (i.e., causes 31 to 40). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE
FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table D.4 Data for the fourth set of 10 fatality causes (i.e., causes 31 to 40). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.4 Data for the fourth set of 10 fatality causes (i.e., causes 31 to 40). (Continued).

FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.4 Data for the fourth set of 10 fatality causes (i.e., causes 31 to 40). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table D.5 Data for the fifth set of 10 fatality causes (i.e., causes 41 to 50).

No jobsite inspection	Lack of inspection for equipment and tools	Poor assembling of equipment/ scaffold/de cking/form work	Error in design	Failure of structural element	Collapse of structure	Over excavating	No safe access to site/scaffol d/trench	Poor housekeepi ng	No safe exit to site
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.5 Data for the fifth set of 10 fatality causes (i.e., causes 41 to 50). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table D.5 Data for the fifth set of 10 fatality causes (i.e., causes 41 to 50). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.5 Data for the fifth set of 10 fatality causes (i.e., causes 41 to 50). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.5 Data for the fifth set of 10 fatality causes (i.e., causes 41 to 50). (Continued).

FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.6 Data for the sixth set of 10 fatality causes (i.e., causes 51 to 60).

Working surface condition not suited to task	Site obstruction	No safe walkways	No site survey	Inappropriate lighting	Not following proper work procedures	No hazard identification/communication program	No testing procedure for equipment	No effective emergency plan	Lack of safe working procedures
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.6 Data for the sixth set of 10 fatality causes (i.e., causes 51 to 60). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.6 Data for the sixth set of 10 fatality causes (i.e., causes 51 to 60). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table D.6 Data for the sixth set of 10 fatality causes (i.e., causes 51 to 60). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE

Table D.6 Data for the sixth set of 10 fatality causes (i.e., causes 51 to 60). (Continued).

FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

The data used for the developed model is a Boolean matrix (i.e., comprised of TRUE and FALSE values only) where the columns are the 60 fatality causes and the rows are the 100 case files. A 'TRUE' value means that the corresponding fatality factor/cause was reported in the associated case file. A 'FALSE' value means that the corresponding fatality is not reported in the associated case file. This data was shown previously. To be able to use it, the columns will need to be aggregated so that a Boolean matrix of size 100 by 60 is formed. To be able to run the Python code below, this aggregated data needs to be saved in a csv (comma-separated values) file named 'Reference Matrix_Boolean.csv'.

The implemented Python code (the codes' comments are shown as underlined) for the clustering algorithm is as follows:

```
#importing all needed packages
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from numpy.linalg import matrix_power
from numpy.linalg import multi_dot
from scipy.linalg import fractional_matrix_power
from sklearn.metrics import silhouette_score
import openpyxl
import xlrd
```

```
import xlwt
import csv
%matplotlib inline
'''
Note: the following versions were used for the different packages:
Version of pandas is: 1.1.5
Version of numpy is: 1.20.1
Version of seaborn is: 0.11.1
Version of matplotlib is: 3.3.4
Version of sklearn is: 0.24.1
Version of openpyxl is: 3.0.7
Version of xlrd is: 2.0.1
Version of csv is: 1.0
'''
#reading the reference matrix in a Boolean format (i.e., TRUE and FALSE). Rows are case files and columns are fatality causes
Reference_Matrix_Boolean = pd.read_csv("Reference Matrix_Boolean.csv")
#transforming the Boolean reference matrix to a binary reference matrix
Reference_Matrix_Binary=np.multiply(Reference_Matrix_Boolean, 1)
#taking the transpose of the reference matrix so that fatality causes become on the rows and the case files on the columns
Reference_Matrix_Binary=np.transpose(Reference_Matrix_Binary)
#calculating the adjacency matrix
```

```
Adjacency_Matrix = np.matmul(Reference_Matrix_Binary.values,
np.transpose(Reference_Matrix_Binary.values))
#transforming it to a dataframe
Adjacency_Matrix=pd.DataFrame(Adjacency_Matrix)
#replacing the values on the diagonals with zeros
for i in range(Adjacency_Matrix.shape[0]):
   for j in range(Adjacency_Matrix.shape[1]):
      if i==j:
         Adjacency_Matrix.iloc[i,j]=0
#choosing the optimal number of clusters based on the silhouette score
sil = [] #initializing a list that will store the silhouette scores
kmin = 3 #choosing the minimum number of clusters
kmax = 20 #choosing the maximum number of clusters
#calculating the silhouette score for each value of k
for k in range(kmin, kmax+1):
  kmeans = KMeans(n_clusters = k,random_state=42).fit(Adjacency_Matrix)
  labels = kmeans.labels_
  sil.append(silhouette_score(Adjacency_Matrix, labels, metric = 'euclidean'))
#transforming the sil list to a dictionary where the keys are the k values and the values are
the sil scores
sil_dic = {x : sil[x- kmin] for x in range(kmin, kmax+1)}
#getting the optimal number of clusters corresponding to the maximum sil score
k_optimal=max(sil_dic, key=sil_dic.get)
```

```
print('The optimal number of clusters is:',k_optimal)
#plotting the sil scores
plt.figure(figsize=(8,6))
plt.plot(np.arange(kmin, kmax+1, 1),sil,marker='o')
plt.axvline(x= k_optimal,c='red',ymax=0.95,linestyle=':')
plt.xticks(np.arange(kmin, kmax+1, 1.0))
plt.ylabel('Silhouette Score',fontweight='bold',fontsize=14)
plt.xlabel('Number of Clusters (k)',fontweight='bold',fontsize=14)
plt.xlim((kmin-0.5, kmax+0.5))
plt.ylim((0, max(sil)*1.07))
plt.grid(linestyle=':', linewidth='0.6', color='gray')
plt.show()
#calculating the degree matrix
D = np.diag(np.sum(np.array(Adjacency_Matrix), axis=1))
#calculating the unormalized Laplacian matrix
L_unnormalized = D - Adjacency_Matrix
#calculating the normalized symmetrical Laplacian matrix
L_normalized = multi_dot([fractional_matrix_power(D, -0.5), L_unnormalized,
fractional_matrix_power(D, -0.5)])
#calculating the eigenvalues and eigenvectors
e, v = np.linalg.eig(L_normalized)
#dimension reduction by choosing the largest k eigenvectors
Largest_Eigenvalues=e[np.argsort(e)[-k_optimal:]]
```

```
#getting the indices of the largest
Largest_Eigenvalues_indices=(e[:, None] == Largest_Eigenvalues).argmax(axis=0)


#getting the fatality cause numbers corresponding to the largest k eigenvectors
Largest_Eigenvalues_fatality_causes = [x + 1 for x in Largest_Eigenvalues_indices]
#calculating the U matrix
U=v[:,Largest_Eigenvalues_indices]
#calculating the T matrix
#first, we need to get the squared values of the matrix U
u_squared = np.power(U, 2)
#second, we need to sum the obtained values over the rows
sum_u_squared_rows=np.sum(u_squared, axis = 1)
#third, we need to get the square root of the sums
sum_root=np.power(sum_u_squared_rows, 0.5)
#finally, we need to get the T matrix
T=(U / sum_root[:,None])
#the final step of the spectral clustering is k-means clustering
km = KMeans(init='k-means++', random_state=283, n_clusters=k_optimal) #defining the
k-means parameters
km.fit(T) #fitting the k-means clustering algorithm
Clusters=km.labels_ #getting the labels of the clusters (i.e., 0, 1, 2, 3, and 4 for the 60
fatality causes)
print(Clusters)
```

```
#the printed output is: [1 0 4 2 0 0 3 1 3 2 1 4 1 0 1 4 2 0 1 1 4 3 1 1 1 3 0 4 0 0 1 4 1 3 1 4
0 3 2 2 4 4 3 0 1 0 0 2 3 1 2 2 3 1 2 3 4 2 0 3]. Note that you might have different results
than these since k-means is sensitive to the random initiation and also depending on the
used versions of the packages and the associated random_state number
#saving the fatality causes' numbers in each cluster
#cluster 1
Cluster1_indices=np.where(Clusters == 0)
Cluster1_fatality_causes = [x + 1 for x in Cluster1_indices]
#cluster 2
Cluster2_indices=np.where(Clusters == 2) #note: this 2 could have been any number of
the following: 0, 1, 2, 3, 4. But all numbers need to covered throughout all clusters
Cluster2_fatality_causes = [x + 1 for x in Cluster2_indices]
#cluster 3
Cluster3_indices=np.where(Clusters == 1)
Cluster3_fatality_causes = [x + 1 for x in Cluster3_indices]
#cluster 4
Cluster4_indices=np.where(Clusters == 4)
Cluster4_fatality_causes = [x + 1 for x in Cluster4_indices]
#cluster 5
Cluster5_indices=np.where(Clusters == 3)
Cluster5_fatality_causes = [x + 1 for x in Cluster5_indices]
#printing the clusters
print('Cluster #1 contains the following fatality causes:',*Cluster1_fatality_causes)
```

```
print('Cluster #2 contains the following fatality causes:',*Cluster2_fatality_causes)
print('Cluster #3 contains the following fatality causes:',*Cluster3_fatality_causes)
print('Cluster #4 contains the following fatality causes:',*Cluster4_fatality_causes)
print('Cluster #5 contains the following fatality causes:',*Cluster5_fatality_causes)


'''
This prints as follows:
Cluster #1 contains the following fatality causes: [ 2  5  6 14 18 27 29 30 37 44 46 47 59]
Cluster #2 contains the following fatality causes: [ 4 10 17 39 40 48 51 52 55 58]
Cluster #3 contains the following fatality causes: [ 1  8 11 13 15 19 20 23 24 25 31 33 35
45 50 54]
Cluster #4 contains the following fatality causes: [ 3 12 16 21 28 32 36 41 42 57]
Cluster #5 contains the following fatality causes: [ 7  9 22 26 34 38 43 49 53 56 60]
'''
#saving the different clusters into excel files
Reference_Matrix_Boolean.iloc[:,list(Cluster1_indices[0])].to_excel('Cluster
1.xlsx',index=False)
Reference_Matrix_Boolean.iloc[:,list(Cluster2_indices[0])].to_excel('Cluster
2.xlsx',index=False)
Reference_Matrix_Boolean.iloc[:,list(Cluster3_indices[0])].to_excel('Cluster
3.xlsx',index=False)
Reference_Matrix_Boolean.iloc[:,list(Cluster4_indices[0])].to_excel('Cluster
4.xlsx',index=False)
```

```
Reference_Matrix_Boolean.iloc[:,list(Cluster5_indices[0])].to_excel('Cluster
5.xlsx',index=False)
```

The implemented R code (comments are shown as underlined) for the association rules within Cluster 1 is as follows:

```
#R version 4.0.4 was used
#all versions of R could be found at: https://cran.r-project.org/bin/windows/base/old/
#installing needed packages
install.packages("arules") #This R package provides the infrastructure for representing, manipulating and analyzing transaction data and patterns using frequent itemsets and association rules. Also, it provides a wide range of interest measures and mining algorithms including a interfaces and the code of Borgelt's efficient C implementations of the association mining algorithms Apriori and Eclat.
install.packages("arulesViz") #This R package extends package arules with various visualization techniques for association rules and itemsets. The package also includes several interactive visualizations for rule exploration.
install.packages("readxl") #This package makes it easy to get data out of Excel and into R.
#loading the needed packages
library(arules)
library("arulesViz")
library("readxl")
# setting the number of displayed significant digits to two to make the output easier to read
options(digits = 3)
# setting the seed for the random number generator for predictability
```

```
set.seed(1234)

# free memory

rm(list = ls())

# read xlsx data file into the workspace

AR_Data <- read_excel("Cluster 1.xlsx")

# showing some basic statistics of the data

summary(AR_Data)

# convert the data into the transactions class

basket <- as(AR_Data, "transactions")

# use the Apriori algorithm to mine association rules

rules.all <- apriori(basket)

# quality measurements include support, confidence, lift, and count, round to three decimal
places

quality(rules.all) <- round(quality(rules.all),digits = 3)

# use the Apriori algorithm to mine a subset of association rules

rules <- apriori(basket,parameter = list(minlen=2, supp=0.01,conf=0.75,target = "rules"))

# total number of rules

rules

# inspect the rules

inspect(rules)

summary(rules)

# quantitative measures of the rules' properties

quality(rules)
```

```
# sort rules by lift
rules.sorted <- sort(rules,by="lift")
# quantitative measures of the rules' properties
quality(rules.sorted)
quality(rules) <- round(quality(rules),digits = 3)
# changing the axes of the scatter plot
plot(rules, measure = c("support", "lift"), shading = "lift",
    jitter=0,xlim=c(0.0, 0.02),ylim=c(0, 50))
# re-ordered parallel coordinates plot
plot(rules, method = "paracoord",measure= "confidence",shading = "lift",control =
list(reorder = TRUE))
#Note: you might need to run the above command multiple times to get the exact same
figure reported in this dissertation. However, there is no need for that, since any generated
plot will be equivalent to the one reported in this dissertation.
```

The implemented R code (comments are shown as underlined) for the association rules within Cluster 2 is as follows:

```
#loading the needed packages
library(arules)
library("arulesViz")
library("readxl")
# setting the number of displayed significant digits to two to make the output easier to read
options(digits = 3)
# setting the seed for the random number generator for predictability
```

```
set.seed(1234)

# free memory

rm(list = ls())

# read xlsx data file into the workspace

AR_Data <- read_excel("Cluster 2.xlsx")

# showing some basic statistics of the data

summary(AR_Data)

# convert the data into the transactions class

basket <- as(AR_Data, "transactions")

# use the Apriori algorithm to mine association rules

rules.all <- apriori(basket)

# quality measurements include support, confidence, lift, and count, round to three decimal
places

quality(rules.all) <- round(quality(rules.all),digits = 3)

# use the Apriori algorithm to mine a subset of association rules

rules <- apriori(basket,parameter = list(minlen=2, supp=0.01,conf=0.75,target = "rules"))

# total number of rules

rules

# inspect the rules

inspect(rules)

summary(rules)

# quantitative measures of the rules' properties

quality(rules)
```

```
# sort rules by lift
rules.sorted <- sort(rules,by="lift")
# quantitative measures of the rules' properties
quality(rules.sorted)
quality(rules) <- round(quality(rules),digits = 3)
# changing the axes of the scatter plot
plot(rules, measure = c("support", "lift"), shading = "confidence",
    jitter=0,xlim=c(0.0, 0.05),ylim=c(0, 20))
# re-ordered parallel coordinates plot
plot(rules, method = "paracoord",measure= "lift",shading = "confidence",control =
list(reorder = TRUE))
#Note: you might need to run the above command multiple times to get the exact same
figure reported in this dissertation. However, there is no need for that, since any generated
plot will be equivalent to the one reported in this dissertation.
```

The implemented R code (comments are shown as underlined) for the association rules within Cluster 3 is as follows:

```
#loading the needed packages
library(arules)
library("arulesViz")
library("readxl")
# setting the number of displayed significant digits to two to make the output easier to read
options(digits = 3)
# setting the seed for the random number generator for predictability
```

```
set.seed(1234)

# free memory

rm(list = ls())

# read xlsx data file into the workspace

AR_Data <- read_excel("Cluster 3.xlsx")

# showing some basic statistics of the data

summary(AR_Data)

# convert the data into the transactions class

basket <- as(AR_Data, "transactions")

# use the Apriori algorithm to mine association rules

rules.all <- apriori(basket)

# quality measurements include support, confidence, lift, and count, round to three decimal
places

quality(rules.all) <- round(quality(rules.all),digits = 3)

# use the Apriori algorithm to mine a subset of association rules

rules <- apriori(basket,parameter = list(minlen=2, supp=0.01,conf=0.75,target = "rules"))

# total number of rules

rules

# inspect the rules

inspect(rules)

summary(rules)

# quantitative measures of the rules' properties

quality(rules)
```

```
# sort rules by lift
rules.sorted <- sort(rules,by="lift")
# quantitative measures of the rules' properties
quality(rules.sorted)
quality(rules) <- round(quality(rules),digits = 3)
# changing the axes of the scatter plot
plot(rules, measure = c("support", "lift"), shading = "lift",
    jitter=0,xlim=c(0.0, 0.03),ylim=c(0, 100))
# parallel coordinates plot
plot(rules, method = "paracoord",measure= "confidence",shading = "lift")
```

The implemented R code (comments are shown as underlined) for the association rules within Cluster 3 is as follows:

```
#loading the needed packages
library(arules)
library("arulesViz")
library("readxl")
# setting the number of displayed significant digits to two to make the output easier to read
options(digits = 3)
# setting the seed for the random number generator for predictability
set.seed(1234)
# free memory
rm(list = ls())
# read xlsx data file into the workspace
```

```
AR_Data <- read_excel("Cluster 4.xlsx")

# showing some basic statistics of the data

summary(AR_Data)

# convert the data into the transactions class

basket <- as(AR_Data, "transactions")

# use the Apriori algorithm to mine association rules

rules.all <- apriori(basket)

# quality measurements, round to three decimal places

quality(rules.all) <- round(quality(rules.all),digits = 3)

# use the Apriori algorithm to mine a subset of association rules

rules <- apriori(basket,parameter = list(minlen=2, supp=0.01,conf=0.75,target = "rules"))

# total number of rules

rules

# inspect the rules

inspect(rules)

summary(rules)

# quantitative measures of the rules' properties

quality(rules)

# sort rules by lift

rules.sorted <- sort(rules,by="lift")

# quantitative measures of the rules' properties

quality(rules.sorted)

quality(rules) <- round(quality(rules),digits = 3)
```

```
# changing the axes of the scatter plot
plot(rules, measure = c("support", "lift"), shading = "lift",
    jitter=0,xlim=c(0.0, 0.02),ylim=c(0, 20))
# parallel coordinates plot
plot(rules, method = "paracoord",measure= "support",shading = "lift")
```

The implemented R code (comments are shown as underlined) for the association rules within Cluster 5 is as follows:

```
#loading the needed packages
library(arules)
library("arulesViz")
library("readxl")
# setting the number of displayed significant digits to two to make the output easier to read
options(digits = 3)
# setting the seed for the random number generator for predictability
set.seed(1234)
# free memory
rm(list = ls())
# read xlsx data file into the workspace
AR_Data <- read_excel("Cluster 5.xlsx")
# showing some basic statistics of the data
summary(AR_Data)
# convert the data into the transactions class
basket <- as(AR_Data, "transactions")
```

```
# use the Apriori algorithm to mine association rules

rules.all <- apriori(basket)

# quality measurements, round to three decimal places

quality(rules.all) <- round(quality(rules.all),digits = 3)

# use the Apriori algorithm to mine a subset of association rules

rules <- apriori(basket,parameter = list(minlen=2, supp=0.01,conf=0.75,target = "rules"))

# total number of rules

rules

# inspect the rules

inspect(rules)

summary(rules)

# quantitative measures of the rules' properties

quality(rules)

# sort rules by lift

rules.sorted <- sort(rules,by="lift")

# quantitative measures of the rules' properties

quality(rules.sorted)

quality(rules) <- round(quality(rules),digits = 3)

# changing the axes of the scatter plot

plot(rules, measure = c("support", "lift"), shading = "lift",

    jitter=0,xlim=c(0.0, 0.02),ylim=c(0, 50))

# parallel coordinates plot

plot(rules, method = "paracoord",measure= "support",shading = "lift")
```

APPENDIX E.

QUESTIONS ASKED IN THE SURVEY

- *Project title:* The Impact of Offsite Construction on the Workforce.

- *Brief description:* The Construction Industry Institute (CII) has commissioned Research Team (RT) 371 to study the impact of offsite construction on the 2030 workforce to develop best practices for preparing the current and future construction workforce based on the results of this survey. Offsite construction could be defined as: the offsite fabrication and/or preassembly of project components or systems that have traditionally been stick-built or executed onsite. It includes prefabrication, preassembly, and modularization.

- *Survey Risks:* Although there is risk of loss of confidentiality in case any confidential information is divulged, all data will be saved in a redacted form including the personal information. The investigators will only collect personal information that is absolutely essential to the research activity. The investigators will never release the identities of individual subjects without the express consent of the subject. The investigators will maintain the confidentiality of identifiable information throughout the entire project duration. The research team will only report on the aggregated output of all respondents. As such, the risks of the survey are considered to be "minimal risks" since the probability and magnitude of harm or discomfort anticipated in the proposed research are not greater, in and of themselves, than those ordinarily encountered in daily life or during the performance of routine physical or psychological examination or tests.

- *Benefits:* Any construction firm as well as CII's member companies should substantially benefit from the outcomes of this survey/study as it will: help them position themselves in a more competitive and strategic way; assist in addressing the shortage in skilled labor and in properly training the current and future workforce; and enhance their bottom line, their business operations, and their project performance.

Contact information for the researcher, the advisor, and the IRB chair:

- *Student researcher*: Rayan Hassane Assaad (rayan.assaad@mst.edu).

- *Advisors:* Islam H. El-adaway (eladaway@mst.edu), Makarand (Mark) Hastak (hastak@purdue.edu), and Kim LaScola Needy (kneedy@uark.edu).

- *IRB Chair:* Kathryn Northcut (northcut@mst.edu).

Participation is voluntary:

- You are kindly requested to complete this survey as you have an expert knowledge and experience in the construction industry.

- Please note that your participation in this survey is completely voluntary and you may withdraw at any time if you wish.

The procedures for the study:

- The survey is divided into multiple parts that vary in length.

- To avoid break-off or exhaustion during the survey, it is possible to fill it on several instances in case you do not have time to complete it in one shot. To do so, just exit the survey at any time and your progress will be automatically saved.

- Whenever you would like to continue filling the survey, click on the link that was sent to your email to open the survey. It will open where you last stopped (just remember to use the same computer).

- If you prefer a hard copy version of this survey or you need someone to walk you through it on the phone, please contact: rayan.assaad@mst.edu

- The overall survey will take around 30 minutes to complete.

The method of ensuring subjects' confidentiality:

- All your personal data and individual answers will be used (collected, handled, stored) in compliance with CII's confidentiality rules.

- For the collected personal data (participant name and company name), the data is coded as early in the activity as possible and securely stored so that only the investigators and authorized staff may access it.

- The research team will only report on the aggregated output of all respondents.

- No confidential identifiable information will be disclosed. The investigators will obtain the express permission of the subject to do otherwise.

- In case the investigators wish to use data for a purpose other than the one for which it was originally collected and the data are still identifiable, the investigators will obtain consent from the subjects for the new use of the data.

- Subjects must be aware that they have the rights to be protected against injury or illegal invasions of their privacy and to preservation of their personal dignity.

Expected Growth of Offsite Construction:

- According to recent statistics, the market for offsite/modular construction was estimated to be equal to $112.42 billion (revenues) in 2018 (Markets and Markets 2019).

- A recent analysis conducted by Frost & Sullivan predicted that the global market of offsite/modular construction will grow to reach $215 billion (revenues) in 2025 as a result of the uptick in construction activities and significant cost, labor, and time savings associated with offsite construction (Limaye 2019).

- Since this research focuses on 2030, please keep that information in your mindset while answering the questions present in this survey. In other words, while completing the survey, please keep in mind where the industry will likely be positioned in 2030 based on

the previous statistics (which might require some imagination by thinking outside the box) rather than what will be the state of offsite construction based on the current practices, trends, and advancements.

Part 1: Respondent Data

- Please enter your first and last name_____

- Please enter your e-mail address_____

- Please enter your job title_____

- Please enter your city and state_____

- Please specify the type of your company (check all that apply)

- ☐ Owner/Developer
- ☐ Architect/Engineer
- ☐ General Contractor/Construction Manager
- ☐ Specialty Trade Contractor
- ☐ Supplier/Manufacturer
- ☐ Fabricators (fabrication facility/shop)
- ☐ Other, please specify _____

- Please enter your years of experience in the industry_____

- Please enter your years of experience with offsite construction_____

- Please rank the below industry sectors from the one that you are 'most experienced with' (by entering '1' next to the associated industry sector) to the ones(s) that you have the 'least experience with' (by entering 2, 3, 4, or 5 next to the associated industry sector(s)). No need to rank all sectors, just rank the sectors that you have experience with.

___Residential ___Building (non-residential) and Commercial ___Industrial ___Infrastructure ___Renovation/Revamp

Part 2: The impact of offsite construction on the workforce's occupational skills

The three tables below will capture the expected 2030 impact on the up-skilling in the technical skills (column on the left) and the managerial skills (column on the right) for three categories: (1) Design and engineering workforce occupations; (2) project and construction management workforce occupations; and, (3) project administration workforce occupations. Provide your answer based on where the industry will likely be positioned in 2030 (which might require some imagination by thinking outside the box) rather than what will be the state of offsite construction based on the current practices, trends, and advancements.

Description of the scale

1 = Negligible (will be impacted by <5%)
2 = Minor (will be impacted by 5-10%)
3 = Moderate (will be impacted by 10-20%)
4 = Significant (will be impacted by 20-50%)
5 = Extreme (will be impacted by >50%)

Technical skills: are the teachable and measurable abilities/knowledge needed to perform specific tasks. They are often gained by qualifications (examples include, but not limited to scientific, mathematical, analytical, and information technology skills, etc.)

Managerial skills: are the attributes or abilities that individuals possess to fulfill some specific management activities or tasks. This knowledge/ability is usually learned and/or practiced through experience. Examples include, but not limited to: Conceptual skills, human or interpersonal skills, delegation, problem-solving, motivating, communication, etc.

Upskilling: process of learning new skills within the same job profile due to increased shift towards offsite construction; the employee or worker improves its current skill set.

Table E.1 Design and engineering workforce occupations.

Category 1: Design and engineering workforce occupations	For the following occupations, rate the expected impact on the upskilling in the technical skills as a result of the foreseeable increase in offsite construction in 2030					For the following occupations, rate the expected impact on the upskilling in the managerial skills as a result of the foreseeable increase in offsite construction in 2030				
	1	2	3	4	5	1	2	3	4	5
Project principals/directors	O	O	O	O	O	O	O	O	O	O
Structural engineers	O	O	O	O	O	O	O	O	O	O
Cost estimation engineers	O	O	O	O	O	O	O	O	O	O
Planning engineers	O	O	O	O	O	O	O	O	O	O
Quantity surveying engineers	O	O	O	O	O	O	O	O	O	O
Geotechnical engineers	O	O	O	O	O	O	O	O	O	O
Electrical engineers	O	O	O	O	O	O	O	O	O	O
Mechanical engineers	O	O	O	O	O	O	O	O	O	O
Plumbing engineers	O	O	O	O	O	O	O	O	O	O
Civil engineers	O	O	O	O	O	O	O	O	O	O
Health and safety engineers	O	O	O	O	O	O	O	O	O	O
Quality assurance and quality control engineers	O	O	O	O	O	O	O	O	O	O
Specialty engineers (fire protection, process, operations, industrial, hydraulic, etc.)	O	O	O	O	O	O	O	O	O	O
BIM and 3D/4D/nD information modeling and rendering	O	O	O	O	O	O	O	O	O	O
Detailing professionals	O	O	O	O	O	O	O	O	O	O
Specifications writers	O	O	O	O	O	O	O	O	O	O
Land surveying and landscape professionals	O	O	O	O	O	O	O	O	O	O
Architects	O	O	O	O	O	O	O	O	O	O
Other, please specify	O	O	O	O	O	O	O	O	O	O

Table E.2 Project and construction management workforce occupations.

Category 2: Project and construction management workforce occupations	For the following occupations, rate the expected impact on the upskilling in the technical skills as a result of the foreseeable increase in offsite construction in 2030					For the following occupations, rate the expected impact on the upskilling in the managerial skills as a result of the foreseeable increase in offsite construction in 2030				
	1	2	3	4	5	1	2	3	4	5
Project management professionals	O	O	O	O	O	O	O	O	O	O
Procurement professionals	O	O	O	O	O	O	O	O	O	O
Expeditors	O	O	O	O	O	O	O	O	O	O
Logistics professionals	O	O	O	O	O	O	O	O	O	O
Construction managers	O	O	O	O	O	O	O	O	O	O
Planning and control professionals	O	O	O	O	O	O	O	O	O	O
Safety professionals	O	O	O	O	O	O	O	O	O	O
Quality assurance and quality control professionals	O	O	O	O	O	O	O	O	O	O
Materials management professionals	O	O	O	O	O	O	O	O	O	O
Field coordination and interface management personnel	O	O	O	O	O	O	O	O	O	O
Superintendents	O	O	O	O	O	O	O	O	O	O
Foremen	O	O	O	O	O	O	O	O	O	O
Waste management personnel	O	O	O	O	O	O	O	O	O	O
Start-up, testing, and commissioning personnel	O	O	O	O	O	O	O	O	O	O
Constructability professionals	O	O	O	O	O	O	O	O	O	O
Specialized lifts and heavy haul professionals	O	O	O	O	O	O	O	O	O	O
Other, please specify	O	O	O	O	O	O	O	O	O	O

Table E.3 Project administration workforce occupations.

Category 3: Project administration workforce occupations	For the following occupations, rate the expected impact on the upskilling in the technical skills as a result of the foreseeable increase in offsite construction in 2030					For the following occupations, rate the expected impact on the upskilling in the managerial skills as a result of the foreseeable increase in offsite construction in 2030				
	1	2	3	4	5	1	2	3	4	5
Project executive professionals	O	O	O	O	O	O	O	O	O	O
Legal and contract professionals	O	O	O	O	O	O	O	O	O	O
Environmental, green, and sustainability professionals	O	O	O	O	O	O	O	O	O	O
Project finance professionals	O	O	O	O	O	O	O	O	O	O
Permitting and regulation professionals	O	O	O	O	O	O	O	O	O	O
Risk management professionals	O	O	O	O	O	O	O	O	O	O
Sales, operations, technology, and construction (SOTC) professionals	O	O	O	O	O	O	O	O	O	O
Owner's representative	O	O	O	O	O	O	O	O	O	O
Sub-contractor(s) administration personnel	O	O	O	O	O	O	O	O	O	O
Marketing and business development professionals	O	O	O	O	O	O	O	O	O	O
Human resources (HR) professionals	O	O	O	O	O	O	O	O	O	O
Project controls professionals	O	O	O	O	O	O	O	O	O	O
Computer and information technology (IT) professionals	O	O	O	O	O	O	O	O	O	O
Insurance professionals	O	O	O	O	O	O	O	O	O	O
Advanced work packaging (AWP) professionals	O	O	O	O	O	O	O	O	O	O
Document control professionals	O	O	O	O	O	O	O	O	O	O
Other, please specify	O	O	O	O	O	O	O	O	O	O

Part 3: The impact of offsite construction on the offsite and onsite workforce

The two tables below will capture the expected 2030 offsite construction's impact on the growth/shrinkage of the occupations (column on the left) and on the skillset (column on the right) for two categories: (1) Offsite workforce occupations; and, (2) onsite craft workforce occupations. Provide your answer based on where the industry will likely be positioned in 2030 (which might require some imagination by thinking outside the box) rather than what will be the state of offsite construction based on the current practices, trends, and advancements. Figure E.1 and E.2 provide a description of the used scale.

Description of the scale

± 1 = Negligible (will change by <5%)

± 2 = Minor (will change by 5-10%)

± 3 = Moderate (will change by 10-20%)

± 4 = Significant (will change by 20-50%)

± 5 = Extreme (will change by >50%)

Shrink: the demand or need for the occupation will decrease. Grow: the demand or need for the occupation will increase

Reskilling: is the process of learning new skills to do a different job due to the increased shift toward offsite construction; the employee or worker might earn a completely new degree or certification.

Upskilling: is the process of learning new skills within the same job profile due to the increased shift towards offsite construction; the employee or worker improves its current skill set.

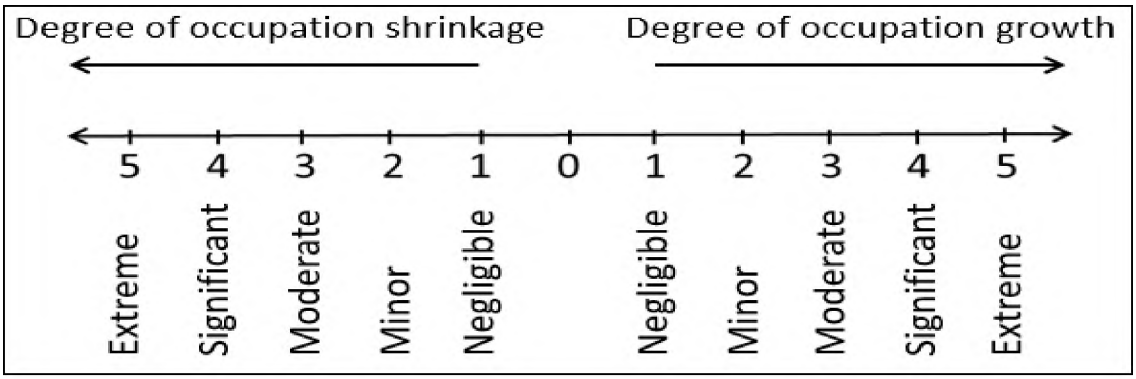

Figure E.1 Description of the growth scale

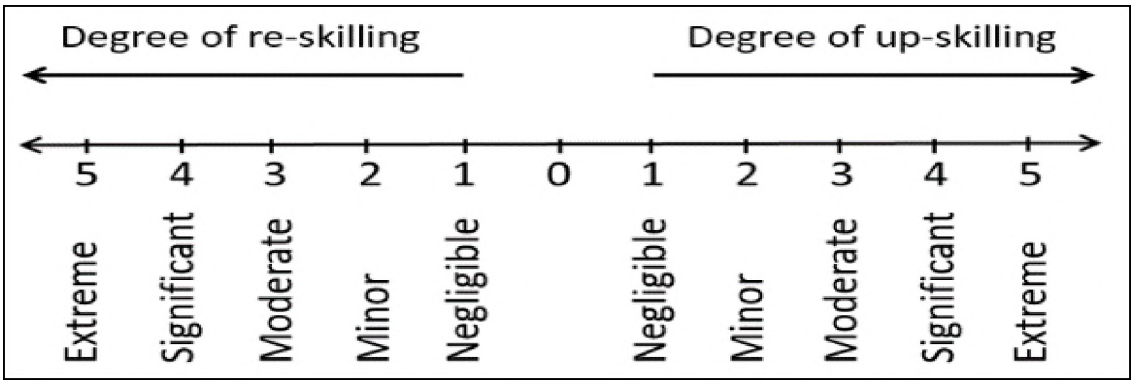

Figure E.2 Description of the skillset scale

Table E.4 Offsite workforce occupations.

Offsite workforce occupations	Will each of the following occupations shrink, grow, or stay the same in 2030 as a result of offsite construction?											How will offsite construction affect the skills of the following occupations in 2030?										
	-5	-4	-3	-2	-1	0	1	2	3	4	5	-5	-4	-3	-2	-1	0	1	2	3	4	5
Assembly, fabrication, and production personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Equipment and machine operations personnel and technicians	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Material handling and warehouse management personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Planners, expeditors, facilitators, sequence management, and supply chain personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Start-up, testing, and commissioning personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Logistics and transportation management personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Engineering personnel (industrial, mechanical, electrical, manufacturing, systems, etc.)	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Quality assurance, quality control, and reliability personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

Table E.4 Offsite workforce occupations. (Continued).

Maintenance, programming, and troubleshooting personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Safety personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Procurement and contract management personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Instrumentation and controls personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Heavy lifting, rigging, and signal personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Technology and configuration analysts	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Detailers	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Specification writers	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Computer-aided manufacturing (CAM) and information modeling professionals	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Truck drivers	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Other, please specify	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

Table E.5 Onsite workforce occupations.

Onsite craft workforce occupations	Will each of the following trades shrink, grow, or stay the same in 2030 as a result of offsite construction?											How will offsite construction affect the skills of the following occupations in 2030?										
	-5	-4	-3	-2	-1	0	1	2	3	4	5	-5	-4	-3	-2	-1	0	1	2	3	4	5
Boilermakers	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Carpenters	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Concrete, brick, block, stone, and plastering personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Drywall personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Electrical personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Equipment operators	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Floor Layers/Installers/Setters	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
General Laborers/Helpers	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Glaziers	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Heavy civil personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Offsite modules/components installation and set-up personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Instrumentation and control personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Insulation personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Ironworkers	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Lifting, cranes, hoisting, rigging, and signal personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Mechanical personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Millwrights	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Painters	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

Table E.5 Onsite workforce occupations. (Continued).

Pipefitters, pipelayers, and steamfitters	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Plumbing personnel	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Roofers and waterproofers	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Scaffold builders	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Sheet metal	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Welders	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Other, please specify	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

Part 4: The impact of offsite construction on the workforce's demographics and attributes

Please rate how you expect the following workforce's demographics and attributes will change (degree of increase or decrease) due to the expected increase of offsite construction in 2030 (as compared to 2020).

Description of the scale

± 1 = Negligible (will change by <5%)

± 2 = Minor (will change by 5-10%)

± 3 = Moderate (will change by 10-20%)

± 4 = Significant (will change by 20-50%)

± 5 = Extreme (will change >50%)

Table E.6 Workforce's demographics and attributes.

Workforce's demographics and attributes	-5 (decrease)	-4	-3	-2	-1	0	1	2	3	4	5 (increase)
Opportunities for women	O	O	O	O	O	O	O	O	O	O	O
Opportunities for minorities (women not included) and diverse ethnic groups	O	O	O	O	O	O	O	O	O	O	O
Opportunities for international/immigrant workforce	O	O	O	O	O	O	O	O	O	O	O
Opportunities for workforce with disabilities	O	O	O	O	O	O	O	O	O	O	O
Opportunities for converting the workforce from non-construction industries to the construction industry	O	O	O	O	O	O	O	O	O	O	O
Opportunities for veterans	O	O	O	O	O	O	O	O	O	O	O
Opportunities for 'silent generation' or 'traditionalists' (they will be more than 85-year-old in 2030)	O	O	O	O	O	O	O	O	O	O	O
Opportunities for 'baby boomers' (they be between 67 and 85 in 2030)	O	O	O	O	O	O	O	O	O	O	O
Opportunities for 'generation X' (they will be between 55 and 66 in 2030)	O	O	O	O	O	O	O	O	O	O	O
Opportunities for 'generation Y' or 'Millennials' (they will be between 34 and 54 in 2030)	O	O	O	O	O	O	O	O	O	O	O
Opportunities for 'generation Z' (they will be between 22 and 33 in 2030)	O	O	O	O	O	O	O	O	O	O	O
Opportunities for 'generation alpha' (they will be younger than 22 in 2030)	O	O	O	O	O	O	O	O	O	O	O
Opportunities for diploma holders (from high schools, trade colleges, and professional schools or any equivalent diploma such as GED: general educational development)	O	O	O	O	O	O	O	O	O	O	O
Opportunities for associate degree holders	O	O	O	O	O	O	O	O	O	O	O
Opportunities for bachelor’s degree holders (recent graduate from undergraduate degrees)	O	O	O	O	O	O	O	O	O	O	O
Opportunities for master's degree holders	O	O	O	O	O	O	O	O	O	O	O
Opportunities for professional degree holders	O	O	O	O	O	O	O	O	O	O	O

Table E.6 Workforce's demographics and attributes. (Continued).

Opportunities for doctoral (or Ph.D) degree holders	O	O	O	O	O	O	O	O	O	O	O
Opportunities for union labor	O	O	O	O	O	O	O	O	O	O	O
Opportunities for non-union labor	O	O	O	O	O	O	O	O	O	O	O
Workforce productivity	O	O	O	O	O	O	O	O	O	O	O
Workforce rework (man-hours attributed to rework)	O	O	O	O	O	O	O	O	O	O	O
Idle/stand-by workforce (paid man-hours not spent working)	O	O	O	O	O	O	O	O	O	O	O
Labor disputes	O	O	O	O	O	O	O	O	O	O	O
Workplace theft and fraud	O	O	O	O	O	O	O	O	O	O	O
Workforce (or staffing) turnover	O	O	O	O	O	O	O	O	O	O	O
Workforce mobility, transfer, or relocation from one geographical area to the other	O	O	O	O	O	O	O	O	O	O	O
Workforce overtime (man-hours attributed to overtime)	O	O	O	O	O	O	O	O	O	O	O
Quality of the working conditions	O	O	O	O	O	O	O	O	O	O	O
Workforce health and safety	O	O	O	O	O	O	O	O	O	O	O
Workforce absenteeism	O	O	O	O	O	O	O	O	O	O	O
Workforce fatigue	O	O	O	O	O	O	O	O	O	O	O
Age of retirement	O	O	O	O	O	O	O	O	O	O	O
Crew or team size	O	O	O	O	O	O	O	O	O	O	O
Job security	O	O	O	O	O	O	O	O	O	O	O
Workforce’s quality of work	O	O	O	O	O	O	O	O	O	O	O
Workforce compensation (overall package: pay and benefits)	O	O	O	O	O	O	O	O	O	O	O
Length of workforce career path progression/development	O	O	O	O	O	O	O	O	O	O	O
Workforce learning rate	O	O	O	O	O	O	O	O	O	O	O
Travel and per diem rate	O	O	O	O	O	O	O	O	O	O	O
Workforce smartness, adaptability, and flexibility	O	O	O	O	O	O	O	O	O	O	O
Cost of workforce training and development	O	O	O	O	O	O	O	O	O	O	O
Shifting work hours to international low-cost labor locations	O	O	O	O	O	O	O	O	O	O	O
Other, please specify	O	O	O	O	O	O	O	O	O	O	O

REFERENCES

Abbas, S., and D. Mosallamy. 2016. "Determinants of FDI flows to developing countries: An empirical study on the MENA region." J. Finance Econ. 4 (1): 30–38.

ABC (Associated Builders and Contractors). 2020. "Construction economic update." Accessed May 7, 2020. https://www.abc.org/NewsMedia/ConstructionEconomics/ConstructionEconomicUpdate/tabid/270/categoryid/46/.

Abdu-Aguye, M. G., and W. Gomaa. 2018. "Novel approaches to activity recognition based on vector auto regression and wavelet transforms." In Proc., 2018 17th IEEE Int. Conf. on Machine Learning and Applications, 951–954. New York: IEEE.

Abdul Nabi, M. and El-adaway, I.H., 2020. "Modular construction: Determining decision-making factors and future research needs." Journal of Management in Engineering, 36 (6): 04020085. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000859.

Abdul Nabi, M. and El-adaway, I.H., 2021. "Understanding the Key Risks Affecting Cost and Schedule Performance of Modular Construction Projects." Journal of Management in Engineering, 37 (4): 04021023.

Abdul Nabi, M., El-adaway, I.H., and Dagli, C., 2020a. "A system dynamics model for construction safety behavior." Procedia Computer Science, 168, pp.249-256. https://doi.org/10.1016/j.procs.2020.02.254.

Abdul Nabi, M., El-adaway, I.H., Fayek, S., Howell, C., and Gambatese, J., 2020b. "Contractual guidelines for construction safety–related issues under design–build standard forms of contract." J. Constr. Eng. Manage. 146 (7): 04020074. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001855.

Abediniangerabi, B., S. M. Shahandashti, N. Ahmadi, and B. Ashuri. 2017. "Empirical investigation of temporal association between architecture billings index and construction spending using time-series methods." J. Constr. Eng. Manage. 143 (10): 04017080. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001391.

Abotaleb, I. S., and I. H. El-Adaway. 2018. "Managing construction projects through dynamic modeling: Reviewing the existing body of knowledge and deriving future research directions." J. Manage. Eng. 34 (6): 04018033. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000633.

AbouRizk, S. M., and Halpin, D. W. 1990. "Probabilistic simulation studies for repetitive construction processes." Journal of construction engineering and management, 116(4), 575-594.

AbouRizk, S., Halpin, D., Mohamed, Y., and Hermann, U. 2011. "Research in modeling and simulation for improving construction engineering operations." Journal of Construction Engineering and Management, 137(10), 843-852.

Abrey, M., and J. J. Smallwood. 2014. "The effects of unsatisfactory working conditions on productivity in the construction industry." Procedia Eng. 85 (2014): 3–9. https://doi.org/10.1016/j.proeng.2014.10.522.

Achimugu, P., Selamat, A. and Ibrahim, R., 2014. "A clustering-based technique for large scale prioritization during requirements elicitation." In Recent Advances on Soft Computing and Data Mining (pp.623-632). Springer, Cham.

Adafin, J., J. O. Rotimi, and S. Wilkinson. 2016. "Risk impact assessments in project budget development: Architects' perspectives." Archit. Eng. Des. Manage. 12 (3): 189–204. https://doi.org/10.1080/17452007.2016.1152228.

Adeleye, B., N. 2018. "Time Series Analysis (Lecture 4 Part 1): Johansen Cointegration Test in EViews." Accessed March 18, 2021. http://cruncheconometrix.blogspot.com/2018/03/time-series-analysis-lecture-4-part-1.html

Agarwal, R., S. Chandrasekaran, and M. Sridhar. 2016. "Imagining construction's digital future." Accessed July 14, 2019. https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/imagining-constructions-digital-future.

Ahn, S., Crouch, L., Kim, T.W. and Rameezdeen, R., 2020. "Comparison of Worker Safety Risks between Onsite and Offsite Construction Methods: A Site Management Perspective." Journal of Construction Engineering and Management, 146(9), p.05020010.

Akbarian, B., and A. Erfanian. 2019. "A framework for seizure detection using effective connectivity, graph theory and deep modular neural networks." Preprint, submitted September 6, 2019. http://arxiv.org/abs/1909.03091.

Akintoye, A. 2000. "Analysis of factors influencing project cost estimating practice." Construction Management & Economics, 18 (1): 77–89. https://doi.org/10.1080/014461900370979.

Alashwal, A. M., Fareed, N. F., and Al-Obaidi, K. M. 2017. "Determining success criteria and success factors for international construction projects for Malaysian contractors." Construction Economics and Building, 17(2), 62-80.

Al-Bayati, A. J., and D. D. York. 2018. "Fatal injuries among Hispanic workers in the US construction industry: Findings from FACE investigation reports." J. Saf. Res. 67 (Dec): 117–123. https://doi.org/10.1016/j.jsr.2018.09.007.

Allison, L., and J. Kaminsky. 2017. "Safety communication networks: Females in small work crews." J. Constr. Eng. Manage. 143 (8): 04017050. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001344.

Allmon, E., C. T. Haas, J. D. Borcherding, and P. M. Goodrum. 2000. "US construction labor productivity trends, 1970–1998." J. Constr. Eng. Manage. 126 (2): 97–104. https://doi.org/10.1061/(ASCE)0733-9364(2000)126:2(97).

Al-Masaeid, H. R., and K. C. Sinha. 1994. "Analysis of accident reduction potentials of pavement markings." J. Transp. Eng. 120 (5): 723–736.

Alsamadani, R., M. R. Hallowell, A. Javernick-Will, and J. Cabello. 2013. "Relationships among language proficiency, communication patterns, and safety performance in small work crews in the United States." J. Constr. Eng. Manage. 139 (9): 1125–1134. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000724.

Anchalia, P.P., Koundinya, A.K. and Srinath, N.K., 2013. "MapReduce design of K-means clustering algorithm." In 2013 International Conference on Information Science and Applications (ICISA) (pp. 1-5). IEEE.

Andersson, J., and M. Burberg. 2015. "Testing for normality of censored data." Accessed October 10, 2018. https://www.diva-portal.org/smash/get/diva2:816450/FULLTEXT01.pdf.

Anggraeni, P. W. 2016. "Transmission mechanism of monetary policy through asset price in Indonesia in the period 2002–2011." J. Indonesian Appl. Econ. 6 (1): 123–141. https://doi.org/10.21776/ub.jiae.2016.006.01.7.

APM (Association for Project Management). 2008. "Interfacing risk and earned value management." Accessed September 22, 2018. https://www.apm.org.uk/media/7584/irevm-sample-chapter.pdf.

Arashpour, M., Bai, Y., Aranda-mena, G., Bab-Hadiashar, A., Hosseini, R. and Kalutara, P., 2017. "Optimizing decisions in advanced manufacturing of prefabricated products: Theorizing supply chain configurations in off-site construction." Automation in Construction, 84, pp.146-153.

Arashpour, M., Wakefield, R., Abbasi, B., Lee, E. W. M., and Minas, J. 2016. "Off-site construction optimization: Sequencing multiple job classes with time constraints." Automation in construction, 71, 262-270.

Arashpour, M., Wakefield, R., Lee, E. W. M., Chan, R., and Hosseini, M. R. 2016. "Analysis of interacting uncertainties in on-site and off-site activities: Implications for hybrid construction." International Journal of Project Management, 34(7), 1393-1402.

Arocho, I., Rasdorf, W., and Hummer, J. 2014. "Methodology to forecast the emissions from construction equipment for a transportation construction project." In Construction Research Congress 2014: Construction in a Global Network (pp. 554-563).

Asaithambi, S. 2017. "Why, how and when to scale your features." Accessed September 14, 2019. https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e.

ASCE (American Society of Civil Engineers). 2016. "Failure to Act: Closing the Infrastructure Investment Gap for America's Economic Future." Accessed September 9, 2020. https://www.infrastructurereportcard.org/wp-content/uploads/2016/10/ASCE-Failure-to-Act-2016-FINAL.pdf

ASCE (American Society of Civil Engineers). 2009. "2009 report card for America's infrastructure: Dams." Accessed September 9, 2019. https://www.infrastructurereportcard.org/2009/fact-sheet/dams.html.

ASCE (American Society of Civil Engineers). 2017. "2017 infrastructure report card: Dams." Accessed September 9, 2019. https://www.infrastructurereportcard.org/wp-content/uploads/2017/01/Dams-Final.pdf.

Ashuri, B., Shahandashti, S.M. and Lu, J., 2012. "Empirical tests for identifying leading indicators of ENR construction cost index." Construction Management and Economics, 30(11), pp.917-927.

Assaad, R. and Abdul-Malak, M.A., 2020a. "Legal perspective on treatment of delay liquidated damages and penalty clauses by different jurisdictions: Comparative analysis." Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 12(2), p.04520013.

Assaad, R. and Abdul-Malak, M.A., 2020b. "Timing of liquidated damages recovery and related liability issues." Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 12(2), p.04520015.

Assaad, R. and El-Adaway, I., 2020a. "A Comprehensive Management Framework for Preventing Operational Bankruptcy of Construction Firms." In Construction Research Congress 2020: Project Management and Controls, Materials, and Contracts (pp. 49-57). Reston, VA: American Society of Civil Engineers.

Assaad, R. and El-Adaway, I., 2020b. "Forecasting and Modeling Bridge Deterioration Using Data Mining Analytics." In Construction Research Congress 2020: Computer Applications (pp. 125-134). Reston, VA: American Society of Civil Engineers.

Assaad, R. and El-adaway, I.H., 2020c. "Bridge infrastructure asset management system: Comparative computational machine learning approach for evaluating and predicting deck deterioration conditions." Journal of Infrastructure Systems, 26(3), p.04020032.

Assaad, R. and El-Adaway, I.H., 2020d. "Enhancing the knowledge of construction business failure: A social network analysis approach." Journal of Construction Engineering and Management, 146(6), p.04020052.

Assaad, R. and El-adaway, I.H., 2020e. "Evaluation and prediction of the hazard potential level of dam infrastructures using computational artificial intelligence algorithms." Journal of Management in Engineering, 36(5), p.04020051.

Assaad, R. and El-adaway, I.H., 2021a. "Determining Critical Combinations of Safety Fatality Causes Using Spectral Clustering and Computational Data Mining Algorithms." Journal of Construction Engineering and Management, 147(5), p.04021035.

Assaad, R. and El-adaway, I.H., 2021b. "Guidelines for Responding to COVID-19 Pandemic: Best Practices, Impacts, and Future Research Directions." Journal of Management in Engineering, 37(3), p.06021001.

Assaad, R. and El-adaway, I.H., 2021c. "Impact of Dynamic Workforce and Workplace Variables on the Productivity of the Construction Industry: New Gross Construction Productivity Indicator." Journal of Management in Engineering, 37(1), p.04020092.

Assaad, R., Ahmed, M.O., El-adaway, I.H., Elsayegh, A. and Siddhardh Nadendla, V.S., 2021a. "Comparing the Impact of Learning in Bidding Decision-Making Processes Using Algorithmic Game Theory." Journal of Management in Engineering, 37(1), p.04020099.

Assaad, R., Dagli, C. and El-adaway, I.H., 2020a. "A system-of-systems model to simulate the complex emergent behavior of vehicle traffic on an urban transportation infrastructure network." Procedia Computer Science, 168, pp.139-146.

Assaad, R., El-Adaway, I. and Abotaleb, I., 2020b. "Holistic Risk Management Approach for Predicting Cost and Schedule Overruns at Project Completion." In Construction Research Congress 2020: Project Management and Controls, Materials, and Contracts (pp. 583-592). Reston, VA: American Society of Civil Engineers.

Assaad, R., El-Adaway, I.H. and Abotaleb, I.S., 2020c. "Predicting project performance in the construction industry." Journal of Construction Engineering and Management, 146(5), p.04020030.

Assaad, R., El-adaway, I.H., Baxmeyer, K., Harman, M., Job, L. and Lashley, H., 2021b. "Allocation of Risks and Responsibilities in Green and Sustainable Buildings." Journal of Architectural Engineering, 27(2), p.04021002.

Assaad, R., El-adaway, I.H., El Hakea, A.H., Parker, M.J., Henderson, T.I., Salvo, C.R. and Ahmed, M.O., 2020d. "Contractual Perspective for BIM Utilization in US Construction Projects." Journal of Construction Engineering and Management, 146(12), p.04020128.

Assaad, R., El-adaway, I.H., Hastak, M. and Needy, K.L., 2020e. "Commercial and Legal Considerations of Offsite Construction Projects and their Hybrid Transactions." Journal of Construction Engineering and Management, 146(12), p.05020019.

Assaad, R., Elsayegh, A., Ali, G., Abdul Nabi, M. and El-Adaway, I.H., 2020f. "Back-to-back relationship under standard subcontract agreements: Comparative study." Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 12(3), p.04520020.

Assaad, R., Elsayegh, A., Ali, G., El-Adaway, I. and Nabi, M.A., 2020g. "Understanding the Sub-Contractual Relationship for Proper Management of Construction Projects." In Construction Research Congress 2020: Project Management and Controls, Materials, and Contracts (pp. 1119-1128). Reston, VA: American Society of Civil Engineers.

Assaf, S., M. A. Hassanain, and A. Abdallah. 2017. "Assessment of deficiencies in design documents for large construction projects." J. Perform. Constr. Facil. 31 (5): 04017086. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001081.

Association of State Dam Safety Officials. 2019a. "Dam failures and incidents." Accessed September 9, 2019. https://www.damsafety.org/dam-failures.

Association of State Dam Safety Officials. 2019b. "Emergency action planning." Accessed September 15, 2019. https://damsafety.org/dam-owners/emergency-action-planning.

Association of State Dam Safety Officials. 2019c. "Frequently asked questions." Accessed September 9, 2019. https://damsafety.org/media/faq.

Attalla, M., T. Hegazy, and R. Haas. 2003. "Reconstruction of the building infrastructure: Two performance prediction models." J. Infrastruct. Syst. 9 (1): 26–34. https://doi.org/10.1061/(ASCE)1076-0342(2003)9:1(26).

Austin, R. B., P. Pishdad-Bozorgi, and J. M. de la Garza. 2016. "Identifying and prioritizing best practices to achieve flash track projects." Journal of Construction Engineering and Management, 142 (2): 04015077.

Avci, O., and O. Abdeljaber. 2016. "Self-organizing maps for structural damage detection: A novel unsupervised vibration-based algorithm." J. Perform. Constr. Facil. 30 (3): 04015043. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000801.

Ayhan, B. U., and O. B. Tokdemir. 2020. "Accident analysis for construction safety using latent class clustering and artificial neural networks." J. Constr. Eng. Manage. 146 (3): 04019114. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001762.

Babar, S., M. J. Thaheem, and B. Ayub. 2017. "Estimated cost at completion: Integrating risk into earned value management." J. Constr. Eng. Manage. 143 (3): 04016104. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001245.

Bai, Y., Huan, J., and Kim, S. 2012. "Measuring bridge construction efficiency using the wireless real-time video monitoring system." Journal of Management in Engineering, 28(2), 120-126.

Balahadia, F. F., B. G. Dadiz, R. R. Ramirez, M. Luvett, P. L. Jay-ar, and A. C. Lagman. 2019. "Application of data mining approach for profiling fire incidents reports of bureau of fire and protection." In Proc., 2019 Int. Conf. on Computational Intelligence and Knowledge Economy (ICCIKE), 713–717. New York: IEEE.

Baldwin, A., C.-S. Poon, L.-Y. Shen, S. Austin, and I. Wong. 2009. "Designing out waste in high-rise residential buildings: Analysis of precasting methods and traditional construction." Renewable Energy 34 (9): 2067–2073.

Barbosa, F., J. Woetzel, J. Mischke, M. J. Ribeirinho, M. Sridhar, M. Parsons, M. Bertram, and S. Brown. 2017. "Reinventing construction: A route to higher productivity." New York: McKinsey & Company. Accessed September 10, 2020. https://www.mckinsey.com/business-functions/operations/our-insights/reinventing-construction-through-a-productivity-revolution

Bartlett, J. E., J. W. Kotrlik, and C. C. Higgins. 2001. "Preview organizational research: Determining appropriate sample size in survey research." Inf. Tech. Learn. Perform. J. 19 (1): 43–50.

Belafi, Z. D., T. Hong, and A. Reith. 2018. "A critical review on questionnaire surveys in the field of energy-related occupant behavior." Energy Effic. 11 (8): 2157–2177. https://doi.org/10.1007/s12053-018-9711-z.

Bell, T. 2017. "Brink of failure? High hazard potential dams." Accessed September 9, 2019. https://www.tbp.org/pubs/Features/Sp17Bell.pdf.

Betz, J. 2018. "27+ impressive project management statistics in 2019." Accessed July 24, 2019. https://learn.g2.com/project-management-statistics.

Bhattacharjee, S., S. Ghosh, and D. Young-Corbett. 2011. "Safety improvement approaches in construction industry: A review and future directions." In Proc., 47th ASC Annual Int. Conf. Los Angeles: American Society of Cinematographers.

Bhimani, J., Leeser, M. and Mi, N., 2015. "Accelerating K-Means clustering with parallel implementations and GPU computing." In 2015 IEEE High Performance Extreme Computing Conference (HPEC) (pp. 1-6). IEEE.

BLS (Bureau of Labor Statistics). 2018. Census of fatal occupational injuries charts, 1992-2017 (final data). Washington, DC: BLS.

Bonham, D. R., P. M. Goodrum, R. Littlejohn, and M. A. Albattah. 2017. "Application of data mining techniques to quantify the relative influence of design and installation characteristics on labor productivity." J. Constr. Eng. Manage. 143 (8): 04017052. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001347.

Boyadzhieva-Georgieva, S. 2014. "A method for choosing a project's planned value curve by integrating earned value and risk management." Econ. Altern. (2): 89–97.

Bröchner, J., and Olofsson, T. 2012. "Construction productivity measures for innovation projects." Journal of construction engineering and management, 138(5), 670-677.

Brown, T., R. Dalton, T. McCleskey, W. Pawlikowski, and N. Ryker. 1988. "Preparing to conduct a dam safety inspection." Accessed November 15, 2019. https://damfailures.org/wp-content/uploads/2015/06/Preparing-to-Conduct-a-Dam-Safety-Inspection.pdf.

Brownlee, J. 2018. "11 classical time series forecasting methods in Python (cheat sheet)." Accessed May 3, 2020. https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/.

Brunskill, E., T. Kollar, and N. Roy. 2007. "Topological mapping using spectral clustering and classification." In Proc., 2007 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 3491–3496. New York: IEEE.

Burkart, M. J. 2002. "Wouldn't it Be Nice if"…!. Practice Periodical on Structural Design and Construction, 7(2), 61-67.

Burson, A. D. 2017. "Determining the feasibility of using abandoned big box stores as modular construction factories." Master thesis, Departement of Construction Science and Management, Clemson University.

Camm, J. D., J. J. Cochran, M. J. Fry, J. W. Ohlmann, and D. R. Anderson. 2019. Business analytics. Boston: Cengage Learning.

Cao, H., and Y. M. Goh. 2019. "Analyzing construction safety through time series methods." Front. Eng. Manage. 6 (2): 262–274. https://doi.org/10.1007/s42524-019-0015-6.

Carmeli, A., and J. Weisberg. 2006. "Exploring turnover intentions among three professional groups of employees." Human Resource Development International, 9(2), pp.191-206.

Carr, S. 2017. "Perceptual errors and the limits of visual performance." In Vol. 4 of The newsletter of the society to improve diagnosis in medicine. New York: IEEE.

Casanovas, M. D. M., J. Armengou, and G. Ramos. 2014. "Occupational risk index for assessment of risk in construction work by activity." J. Constr. Eng. Manage. 140 (1): 04013035. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000785.

Castro-Nova, I., G. M. Gad, A. Touran, B. Cetin, and D. D. Gransberg. 2018. "Evaluating the influence of differing geotechnical risk perceptions on design-build highway projects." ASCE-ASME J. Risk Uncertainty Eng. Syst. Part A: Civ. Eng. 4 (4): 04018038. https://doi.org/10.1061/AJRUA6.0000993.

Ceauşu, F., 2015. "Educational Role of Mental Maps." Review of Artistic Education, (09+ 10), pp.264-272.

Corporate Finance Institute. 2020. "Cointegration." Accessed May 5, 2020. https://corporatefinanceinstitute.com/resources/knowledge/other/cointegration/.

Chan, A. P., J. F. Yeung, C. C. Yu, S. Q. Wang, and Y. Ke. 2011. "Empirical study of risk assessment and allocation of public-private partnership projects in China." J. Manage. Eng. 27 (3): 136–148. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000049.

Chandhok, C., Chaturvedi, S. and Khurshid, A.A., 2012. "An approach to image segmentation using K-means clustering algorithm." International Journal of Information Technology (IJIT), 1(1), pp.11-17.

Chang, C. J., and S. W. Yu. 2018. "Three-variance approach for updating earned value management." J. Constr. Eng. Manage. 144 (6): 04018045. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001491.

Chang, K., and Chi, S. 2019. "Bridge Clustering for Systematic Recognition of Damage Patterns on Bridge Elements." Journal of Computing in Civil Engineering, 33(5), 04019028.

Changali, S., A. Mohammad, and M. van Nieuwland. 2015. "The construction productivity imperative: How to build megaprojects better." New York: McKinsey & Company. Accessed March 24, 2021. https://www.mckinsey.com/business-functions/operations/our-insights/the-construction-productivity-imperative

Chao, Z., and H. J. Kim. 2019. "Removal of computed tomography ring artifacts via radial basis function artificial neural networks." Phys. Med. Biol. 64 (23): 235015. https://doi.org/10.1088/1361-6560/ab5035.

Charfreitag, J. 2020. "Computational Analytics." Accessed September 10, 2020. https://ca.cs.uni-bonn.de/doku.php?id=en:computationalanalytics

Chaturvedi, S., J. J. Thakkar, and R. Shankar. 2018. "Labor productivity in the construction industry: An evaluation framework for causal relationships." Benchmarking 25 (1): 334–356. https://doi.org/10.1108/BIJ-11-2016-0171.

Chemchem, A., F. Alin, and M. Krajecki. 2019. "Combining SMOTE sampling and machine learning for forecasting wheat yields in France." In Proc., 2019 IEEE 2nd Int. Conf. on Artificial Intelligence and Knowledge Engineering (AIKE), 9–14. New York: IEEE.

Chen, C., Q. Wu, G. Zhang, X. C. Liu, and P. D. Prevedouros. 2018. "Extracting arterial access density impacts on safety performance based on clustering and computational analysis." J. Transp. Eng. Part A: Systems 144 (4): 04018008. https://doi.org/10.1061/JTEPBS.0000127.

Chen, H. L. 2014. "Improving forecasting accuracy of project earned value metrics: Linear modeling approach." J. Manage. Eng. 30 (2): 135–145. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000187.

Chen, J., Z. Kira, and Y. K. Cho. 2019. "Deep learning approach to point cloud scene understanding for automated scan to 3D reconstruction." J. Comput. Civ. Eng. 33 (4): 04019027. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000842.

Chen, L., and K. Manley. 2014. "Validation of an instrument to measure governance and performance on collaborative infrastructure projects." J. Constr. Eng. Manage. 140 (5): 04014006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000834.

Chen, L., S. Huang, C. Yang, and Q. Chen. 2020. "Analyzing factors that influence expressway traffic crashes based on association rules: Using the Shaoyang–Xinhuang section of the Shanghai–Kunming expressway as an example." J. Transp. Eng. Part A: Systems 146 (9): 05020007. https://doi.org/10.1061/JTEPBS.0000425.

Chen, X., and D. Cai. 2011. "Large scale spectral clustering with landmark-based representation." In Proc., 25th AAAI Conf. on Artificial Intelligence. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Cheng, C. W., C. C. Lin, and S. S. Leu. 2010. "Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry." Saf. Sci. 48 (4): 436–444. https://doi.org/10.1016/j.ssci.2009.12.005.

Cheong Yong, Y., and N. Emma Mustaffa. 2012. "Analysis of factors critical to construction project success in Malaysia." Engineering, construction and architectural management, 19 (5): 543–556.

Chiang, Y. H., F. K. W. Wong, and S. Liang. 2017. "Fatal construction accidents in Hong Kong." J. Constr. Eng. Manage. 144 (3): 04017121. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001433.

Choi, J. O., X. B. Chen, and T. W. Kim. 2019a. "Opportunities and challenges of modular methods in dense urban environment." International Journal of Construction Management, 19 (2): 93–105.

Choi, J.O., O'Connor, J.T. and Kim, T.W., 2016. "Recipes for cost and schedule successes in industrial modular projects: Qualitative comparative analysis." Journal of Construction Engineering and Management, 142(10), p.04016055.

Choi, J.O., Shrestha, B.K., Kwak, Y.H. and Shane, J.S., 2020. "Innovative technologies and management approaches for facility design standardization and modularization of capital projects." Journal of Management in Engineering, 36(5), p.04020042.

Choi, S. D., L. Guo, J. Kim, and S. Xiong. 2019b. "Comparison of fatal occupational injuries in construction industry in the United States, South Korea, and China." International Journal of Industrial Ergonomics, 71: 64–74. https://doi.org/10.1016/j.ergon.2019.02.011.

Choudhry, R. M., and H. Zahoor. 2016. "Strengths and weaknesses of safety practices to improve safety performance in construction projects in Pakistan." J. Prof. Issues Eng. Educ. Pract. 142 (4): 04016011. https://doi.org/10.1061/(ASCE)EI.1943-5541.0000292.

Choudhry, R. M., J. W. Hinze, M. Arshad, and H. F. Gabriel. 2012. "Subcontracting practices in the construction industry of Pakistan." J. Constr. Eng. Manage. 138 (12): 1353–1359. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000562.

Chowdhury, A., M. Magdon-Ismail, and B. Yener. 2019. "Quantifying contribution and propagation of error from computational steps, algorithms and hyperparameter choices in image classification pipelines." Preprint, submitted February 21, 2019. https://arxiv.org/abs/1903.00405.

Choy, E., and J. Y. Ruwanpura. 2005. "Situation based modeling for construction productivity." In Proc., Construction Research Congress 2005: Broadening Perspectives, 1–10. Reston, VA: ASCE.

Christmann, A., and S. V. Aelst. 2006. "Robust estimation of Cronbach's alpha." Journal of Multivariate Analysis, 97 (7): 1660–1674.

Cicchetti, D.V. and Sparrow, S.A., 1981. "Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior." American journal of mental deficiency, 86(2), 127–137.

CII (Construction Industry Institute). 2009. "The Alberta report: COAA major projects benchmarking summary." Rep. to Alberta Finance and Enterprise and Construction Owner's Association of Alberta (COAA). Edmonton, AB, Canada: CII.

CII (Construction Industry Institute). 2013. "Integrated project risk assessment (IPRA), Version 2.0." Accessed February 21, 2021. https://www.construction-institute.org/resources/knowledgebase/best-practices/project-risk-assessment/topics/rt-181/pubs/ir181-2.

CII (Construction Industry Institute) 2017. "Decision Framework Guide and Tool: Prefabrication, Preassembly, Modularization, and Offsite Fabrication." University of Texas, Austin, TX.

Clemen, R. T., and T. Reilly. 2013. Making hard decisions with DecisionTools. Boston: Cengage Learning.

CLMA (Construction Labor Market Analyzer). 2013."20/20 foresight report." Accessed February 14, 2020. http://www.myclma.com/clma-tools-services/skilled-labor-market-analytics/.

Cochran, W. G. 1977. "Sampling techniques." Third edition. New York: Wiley

Congress Budget Office. 2015. "Estimated Impact of the American Recovery and Reinvestment Act on Employment and Economic Output in 2014." Accessed September 9, 2020. https://www.cbo.gov/sites/default/files/114th-congress-2015-2016/reports/49958-ARRA.pdf

Connelly, L.M., 2008. "Pilot studies." Medsurg Nursing, 17(6), p.411.

Converse, J. M., and S. Presser. 1986. "Survey questions: Handcrafting the standardized questionnaire." Thousand Oaks, CA: SAGE.

Cook, J. 2019. "9 Innovations in Construction Technology Poised to Disrupt the Industry." Accessed September 10, 2020. https://blog.plangrid.com/2019/02/9-innovations-construction-technology-poised-disrupt-industry/

Corporate Finance Institute. 2021. "Durbin Watson Statistic." Accessed March 18, 2021. https://corporatefinanceinstitute.com/resources/knowledge/other/durbin-watson-statistic/

CPWR (Center for Construction Research and Training). 2018. "The construction chart book: The U.S. construction industry and its workers." Silver Spring, MD: CPWR.

Creative Research Systems. 2019. "Sample size calculator." Accessed July 19, 2019. https://www.surveysystem.com/sscalc.htm#two.

Dale, A., and Hamilton, J. 2007. "Sustainable Infrastructure: Implications for Canada's Future." Accessed September 9, 2020. https://www.crcresearch.org/files-crcresearch/File/SI_Final_Report.pdf

Dawood, T., Z. Zhu, and T. Zayed. 2018. "Computer vision–based model for moisture marks detection and recognition in subway networks." J. Comput. Civ. Eng. 32 (2): 04017079. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000728.

de Melo Mendes, B. V., and H. F. Lopes. 2004. "Data driven estimates for mixtures." Comput. Stat. Data Anal. 47 (3): 583–598. https://doi.org/10.1016/j.csda.2003.12.006.

Devi, R. K., and G. Murugaboopathi. 2019. "An efficient clustering and load balancing of distributed cloud data centers using graph theory." Int. J. Commun. Syst. 32 (5): e3896. https://doi.org/10.1002/dac.3896.

Dickey, D. A., and W. A. Fuller. 1979. "Distribution of the estimators for autoregressive time series with a unit root." J. Am. Stat. Assoc. 74 (366): 427–431. https://doi.org/10.1080/01621459.1979.10482531.

Dikmen, I., C. Budayan, M. Talat Birgonul, and E. Hayat. 2018. "Effects of risk attitude and controllability assumption on risk ratings: Observational study on international construction project risk assessment." J. Manage. Eng. 34 (6): 04018037. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000643.

Dillman, D. A. 2011. Mail and Internet surveys: The tailored design method: 2007 update with new Internet, visual, and mixed-mode guide. Hoboken, NJ: Wiley.

Ding, Q., Z. Li, S. Haeri, and L. Trajković. 2018. "Application of machine learning techniques to detecting anomalies in communication networks: Datasets and feature selection algorithms." In Cyber threat intelligence, 47–70. Cham, Switzerland: Springer.

Dong, S., Li, Q., Farahmand, H., Mostafavi, A., Berke, P. R., and Vedlitz, A. 2020. "Institutional connectedness in resilience planning and management of interdependent infrastructure systems." Journal of Management in Engineering, 36(6), 04020075.

Dong, X. S., J. A. Largay, S. D. Choi, X. Wang, C. T. Cain, and N. Romano. 2017. "Fatal falls and PFAS use in the construction industry: Findings from the NIOSH FACE reports." Accid. Anal. Prev. 102 (1): 136–143. https://doi.org/10.1016/j.aap.2017.02.028.

Donkor, S. 2011. "Determinants of business failure: The perspective of SMEs building contractors in the Ghanaian construction industry." Doctoral dissertation, Dept. of Building Technology, College of Architecture and Planning.

Du, J., B. C. Kim, and D. Zhao. 2016. "Cost performance as a stochastic process: EAC projection by Markov Chain simulation." J. Constr. Eng. Manage. 142 (6): 04016009. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001115.

Duffy Group 2020. "Construction Industry Building Momentum into 2020." Accessed September 10, 2020. https://duffygroup.com/construction-industry-building-momentum-2020/.

Dulaimi, M. F., Nepal, M. P., and Park, M. 2005. "A hierarchical structural model of assessing innovation and project performance." Construction Management and Economics, 23(6), 565-577.

Duncheva, T., and Bradley, F. F. 2019. "Multifaceted productivity comparison of off-site timber manufacturing strategies in Mainland Europe and the United Kingdom." Journal of Construction Engineering and Management, 145(8), 04019043.

Durdyev, S., S. Ismail, and N. Kandymov. 2018. "Structural equation model of the factors affecting construction labor productivity." J. Constr. Eng. Manage. 144 (4): 04018007. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001452.

Eisenbraun, R., and L. LaRiviere. 2014. "Dam inspections—Why are they so important?" Accessed September 9, 2019. https://www.waterpowermagazine.com/opinion/opiniondam-inspections-why-so-important-4265954/.

El Mouden, Z. A., A. Jakimi, and M. Hajar. 2019. "An application of spectral clustering approach to detect communities in data modeled by graphs." In Proc., 2nd Int. Conf. on Networking, Information Systems and Security. New York: Association for Computing Machinery.

El-adaway, I. H., G. Ali, R. Assaad, A. Elsayegh, and I. S. Abotaleb. 2019. "Analytic overview of citation metrics in the civil engineering domain with focus on construction engineering and management specialty area and its subdisciplines." J. Constr. Eng. Manage. 145 (10): 04019060. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001705.

El-Gohary, K. M., R. F. Aziz, and H. A. Abdel-Khalek. 2017. "Engineering approach using ANN to improve and predict construction labor productivity under different influences." J. Constr. Eng. Manage. 143 (8): 04017045. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001340.

El-Hoteiby, A. I., O. A. Hosny, and A. F. Waly. 2017. "Particular conditions to cover potential risks of construction projects." J. Leg. Aff. Dispute Resolut. Eng. Constr. 9 (3): 05017002. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000223.

Elliott, G., and E. Pesavento. 2009. "Testing the null of no cointegration when covariates are known to have a unit root." Econom. Theory 25 (6): 1829–1850. https://doi.org/10.1017/S026646660999034X.

Elsayegh, A., El-Adaway, I.H., Assaad, R., Ali, G., Abotaleb, I., Smith, C., Bootwala, M. and Eteifa, S., 2020. Contractual guidelines for management of infrastructure transportation projects. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 12(3), p.04520023.

ElZomor, M., R. Burke, K. Parrish, and G. E. Gibson Jr. 2018. "Front-end planning for large and small infrastructure projects: Comparison of project definition rating index tools." J. Manage. Eng. 34 (4): 04018022. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000611.

Encyclopedia.com. 2019. "Dams." Accessed September 9, 2019. https://www.encyclopedia.com/science-and-technology/technology/technology-terms-and-concepts/dam.

Engmann, S., and D. Cousineau. 2011. "Comparing distributions: The two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test." J. Appl. Quant. Methods 6 (3): 1–17.

Enshassi, M.S., Walbridge, S., West, J.S. and Haas, C.T., 2020. "Dynamic and Proactive Risk-Based Methodology for Managing Excessive Geometric Variability Issues in Modular Construction Projects Using Bayesian Theory." Journal of Construction Engineering and Management, 146(2), p.04019096.

Erica. 2020. "A guide to conducting cointegration tests." Accessed May 5, 2020. https://www.aptech.com/blog/a-guide-to-conducting-cointegration-tests/.

Esmaeili, B., and Hallowell, M. R. 2012. "Diffusion of safety innovations in the construction industry." Journal of Construction Engineering and Management, 138(8), 955-963.

Eteifa, S. O., and I. H. El-adaway. 2017. "Using social network analysis to model the interaction between root causes of fatalities in the construction industry." J. Manage. Eng. 34 (1): 04017045. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000567.

Evia, C. 2011. "Localizing and designing computer-based safety training solutions for Hispanic construction workers." J. Constr. Eng. Manage. 137 (6): 452–459. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000313.

Fagel, M. J. 2011. Principles of emergency management: Hazard specific issues and mitigation strategies. Boca Raton, FL: CRC Press.

Faghih, S. A. M., and H. Kashani. 2018. "Forecasting construction material prices using vector error correction model." J. Constr. Eng. Manage. 144 (8): 04018075. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001528.

Fan, R., T. Ng, and J. Wong. 2010. "Reliability of the Box–Jenkins model for forecasting construction demand covering times of economic austerity." Constr. Manage. Econ. 28 (3): 241–254. https://doi.org/10.1080/01446190903369899.

Fayek, S., Xia, X., and Zhang, X. 2020a. "A Least Square Optimization Approach for Determining the Soil Boundary and Absolute Volume of Unsaturated Soils." In Geo-Congress 2020: Geo-Systems, Sustainability, Geoenvironmental Engineering, and Unsaturated Soil Mechanics (pp. 394-401). Reston, VA: American Society of Civil Engineers.

Fayek, S., Xia, X., and Zhang, X. 2021. "Validation of Least Square Optimization Method for Determining the Absolute Volume of Unsaturated Soils." In 4th International Conference on Transportation Geotechnics (ICTG). Reston, VA: American Society of Civil Engineers.

Fayek, S., Xia, X., Li, L., and Zhang, X. 2020b. "Photogrammetry-Based Method to Determine the Absolute Volume of Soil Specimen during Triaxial Testing." Transportation Research Record, 2674(8), 206-218. Reston, VA: American Society of Civil Engineers.

Fellows, R. F., and A. M. Liu. 2015. "Research methods for construction." New York: Wiley.

FEMA (Federal Emergency Management Agency). 2004a. "Federal guidelines for dam safety: Glossary of terms." Accessed September 9, 2019. https://www.fema.gov/media-library-data/20130726-1516-20490-9730/fema-148.pdf.

FEMA (Federal Emergency Management Agency). 2004b. "Federal guidelines for dam safety: Hazard potential classification system for dams." Accessed September 9, 2019. https://www.ferc.gov/industries/hydropower/safety/guidelines/fema-333.pdf.

FEMA (Federal Emergency Management Agency). 2013. "Dam safety in the United States." Accessed September 15, 2019. https://www.fema.gov/media-library-data/1402876995238-1c041ca9a4489ea27152c515ed72e38f/DamSafetyintheUnitedStates.pdf.

FEMA (Federal Emergency Management Agency). 2015. "FEMA national dam safety program fact sheet." Accessed September 9, 2019. https://www.fema.gov/media-library-data/1486735320675-8b0597aca8b23c7e2df293310e248bee/NDSPFlashFactSheet2015.pdf.

FEMA (Federal Emergency Management Agency). 2016. "The National Dam Safety Program: Biennial report to the United States Congress, fiscal years 2014-2015." Accessed September 9, 2019. https://www.fema.gov/media-library-data/1470749866373-5de9234b8a02a3577c2646ffdf6eb087/FEMAP1067.pdf.

FHWA (Federal Emergency Management Agency). 2020. "Employment Impacts of Highway Infrastructure Investment." Accessed September 9, 2020. https://www.fhwa.dot.gov/policy/otps/pubs/impacts/

Field, A. 2013. Discovering statistics using IBM SPSS statistics. New York: SAGE.

Flanagan, R., and G. Norman. 1993. Risk management and construction. Oxford, UK: Blackwell.

Fleiss, J. L. 1981. "Statistical methods for rates and proportions" Second Edition. New York: Wiley

Fleming, Q. W., and J. M. Koppelman. 2002. "Earned value management: Mitigating the risks associated with construction projects." Program Manager 31 (2): 90–95.

Flintsch, G. W. 2002. "Soft computing applications in transportation infrastructure asset management." In Applications of Advanced Technologies in Transportation (2002) (pp. 449-456).

Florez, L., P. Armstrong, and J. C. Cortissoz. 2020. "Does compatibility of personality affect productivity? Exploratory study with construction crews." J. Manage. Eng. 36 (5): 04020049. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000807.

FMI (Fails Management Institute) 2018. "New day, new mindset: rethinking offsite construction." Accessed November 17, 2019. https://www.fminet.com/special-reports/2018-fmi-curt-cii-owner-survey/

Fournier-Viger, P., J. C. W. Lin, T. Truong-Chi, and R. Nkambou. 2019. "A survey of high utility itemset mining." In High-utility pattern mining, 1–45. New York: Springer.

Fowler, F. J. 1995. "Improving survey questions: Design and evaluation". Thousand Oaks, CA: SAGE.

Frost & Sullivan 2019. "Global Modular and Prefabricated Building Market Set for Robust CAGR of 6.3% from 2018 to 2025." Accessed February 24, 2021. https://ww2.frost.com/news/press-releases/global-modular-and-prefabricated-building-market-set-for-robust-cagr-of-6-3-from-2018-to-2025/

Fu, Z., W. Hu, and T. Tan. 2005. "Similarity based vehicle trajectory clustering and anomaly detection." In Vol. 2 of Proc., IEEE Int. Conf. on Image Processing 2005, II-602. New York: IEEE.

Gallen, C. L., and M. D'Esposito. 2019. "Brain modularity: A biomarker of intervention-related plasticity." Trends Cognitive Sci. 23 (4): 293–304. https://doi.org/10.1016/j.tics.2019.01.014.

Gao, Z. K., M. Small, and J. Kurths. 2017. "Complex network analysis of time series." Europhys. Lett. 116 (5): 50001. https://doi.org/10.1209/0295-5075/116/50001.

Geisser, S., and W. O. Johnson. 2006. Vol. 529 of Modes of parametric statistical inference. New York: Wiley.

Ghodrati, N., T. Wing Yiu, S. Wilkinson, and M. Shahbazpour. 2018. "Role of management strategies in improving labor productivity in general construction projects in New Zealand: Managerial perspective." J. Manage. Eng. 34 (6): 04018035. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000641.

Giles, D. 2011. "Testing for Granger Causality." Accessed March 18, 2021. https://davegiles.blogspot.com/2011/04/testing-for-granger-causality.html

Giles, D. 2012. "Cointegration Analysis With I(2) & I(1) Data." Accessed March 18, 2021. https://davegiles.blogspot.com/2012/01/cointegration-analysis-with-i2-i1-data.html

Glen, S. 2016a. "Serial correlation/autocorrelation: Definition, tests." Accessed June 14, 2020. https://www.statisticshowto.com/serial-correlation-autocorrelation/.

Glen, S. 2016b. "Unit root: Simple definition, unit root tests." Accessed June 14, 2020. https://www.statisticshowto.com/unit-root/.

Global Modular Construction Market Research Report. 2017. "Forecast to 2023: Market scenario." Accessed October 10, 2019. https://www.reportbuyer.com/product/5376942/global-modular-construction-market-research-report-forecast-to-2023.html.

GlobalData. 2018. "Infrastructure Insight: The US." Accessed September 9, 2020. https://store.globaldata.com/report/gdif0022ii--infrastructure-insight-the-us-2/

Goldberg, A. T. 2003. "Rethinking the chain of events analogy for incidents." In Proc., ASSE Professional Development Conf. and Exposition. Des Plaines, IL: American Society of Safety Engineers.

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

Goodrum, P. M., C. T. Haas, C. Caldas, D. Zhai, J. Yeiser, and D. Homm. 2011. "Model to predict the impact of a technology on construction productivity." Journal of Construction Engineering and Management, 137 (9): 678–688.

Goodrum, P.M., 2003. "Worker satisfaction and job preferences in the US construction industry." In Construction Research Congress: Wind of Change: Integration and Innovation (pp. 1-8).

Goswami, B., Mahajan, R., Koner, B., and Jain, A. 2020. "Team Idea Mapping Method: A Brain Storming Session for Enhancing Problem Solving Skills in Postgraduate Medical Biochemistry Students as Assessed by Self-Efficacy". JMIR Medical Education Preprints, 1(1), 22426.

Gou, J., H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang. 2019. "A generalized mean distance-based k-nearest neighbor classifier." Expert Syst. Appl. 115 (Jan): 356–372. https://doi.org/10.1016/j.eswa.2018.08.021.

Granger, C. W. 1969. "Investigating causal relations by econometric models and cross-spectral methods." Econometrica: J. Econ. Soc. 37 (3): 424–438. https://doi.org/10.2307/1912791.

Gunopulos, D., G. Das, and G. Das. 2001. "Time series similarity measures and time series indexing." ACM SIGMOD Rec. 30 (2): 624. https://doi.org/10.1145/376284.375808.

Guo, J., Z. Gao, and Y. Wang. 2020. "Forecast of civil aviation unsafe events rate using grey-buffer operator-Markov chain method." In Proc., Int. Conf. on Transportation and Development 2020, 71–82. Reston, VA: ASCE.

Gupta, M., A. Hasan, A. K. Jain, and K. N. Jha. 2018. "Site amenities and workers' welfare factors affecting workforce productivity in Indian construction projects." J. Constr. Eng. Manage. 144 (11): 04018101. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001566.

Gurmu, A. T. 2019. "Tools for measuring construction materials management practices and predicting labor productivity in multistory building projects." J. Constr. Eng. Manage. 145 (2): 04018139. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001611.

Gurmu, A. T., and C. S. Ongkowijoyo. 2020. "Predicting construction labor productivity based on implementation levels of human resource management practices." J. Constr. Eng. Manage. 146 (3): 04019115. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001775.

Hahsler, M., and S. Chelluboina. 2011. "Visualizing association rules: Introduction to the R-extension package arulesViz." In R project module, 223–238. Seattle: Semantic Scholar.

Haider, H., R. Sadiq, and S. Tesfamariam. 2016. "Risk-based framework for improving customer satisfaction through system reliability in small-sized to medium-sized water utilities." J. Manage. Eng. 32 (5): 04016008. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000435.

Hammerla, N. Y., and T. Plötz. 2015. "Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition." In Proc., 2015 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing, 1041–1051. New York: Association for Computing Machinery.

Hanna, A.S., Mikhail, G. and Iskandar, K.A., 2017. "State of prefab practice in the electrical construction industry: Qualitative assessment." Journal of Construction Engineering and Management, 143(2), p.04016097.

Hassanzadeh, T. and Meybodi, M.R., 2012. "A new hybrid approach for data clustering using firefly algorithm and K-means." In The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012) (pp. 007-011). IEEE.

Hatamleh, M. T., M. Hiyassat, G. J. Sweis, and R. J. Sweis. 2018. "Factors affecting the accuracy of cost estimate: Case of Jordan." Eng. Constr. Archit. Manage. 25 (1): 113–131. https://doi.org/10.1108/ECAM-10-2016-0232.

Hazır, Ö. 2015. "A review of analytical models, approaches and decision support tools in project monitoring and control." Int. J. Project Manage. 33 (4): 808–815. https://doi.org/10.1016/j.ijproman.2014.09.005.

He, M., S. J. Horng, P. Fan, R. S. Run, R. J. Chen, J. L. Lai, and K. O. Sentosa. 2010. "Performance evaluation of score level fusion in multimodal biometric systems." Pattern Recognit. 43 (5): 1789–1800. https://doi.org/10.1016/j.patcog.2009.11.018.

Health and Safety Executive Construction Division. 2009. "Phase 1 report: Underlying causes of construction fatal accidents—A comprehensive review of recent work to consolidate and summarise existing knowledge." New York: Crown Publishing.

Heaton, J. 2008. "Introduction to neural networks with Java." St. Louis: Heaton Research.

Hendrickson, C. 2005. "Discussion of 'Is construction labor productivity really declining?' by Eddy M. Rojas and Peerapong Aramvareekul." J. Constr. Eng. Manage. 131 (2): 269–270. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:2(269).

Hillson, D. 2004. "Earned value management and risk management: A practical synergy." In Proc., PMI 2004 Global Congress, 1–7. Prague, Czech Republic: PMI.

Hinze, J., M. Hallowell, and K. Baud. 2013. "Construction-safety best practices and relationships to safety performance." J. Constr. Eng. Manage. 139 (10): 04013006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000751.

Hoffman, G. J., A. E. Thal Jr., T. S. Webb, and J. D. Weir. 2007. "Estimating performance time for construction projects." J. Manage. Eng. 23 (4): 193–199. https://doi.org/10.1061/(ASCE)0742-597X(2007)23:4(193).

Holt, G. D. 2015. "British construction business 1700-2000: proactive innovation or reactive evolution?" Construction innovation. 15 (3): 258–277.

Hong, H., J. Zhu, M. Chen, P. Gong, C. Zhang, and W. Tong. 2018. "Quantitative structure–activity relationship models for predicting risk of drug-induced liver injury in humans." In Drug-induced liver toxicity, 77–100. New York: Humana Press.

Horrison, O. 2018. "Machine learning basics with the k-nearest neighbors algorithm." Accessed November 12, 2019. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

Hosseini, M. R., I. Martek, E. Papadonikolaki, M. Sheikhkhoshkar, S. Banihashemi, and M. Arashpour. 2018. "Viability of the BIM manager enduring as a distinct role: Association rule mining of job advertisements." J. Constr. Eng. Manage. 44 (9): 04018085. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001542.

Htet, T. Z. 2019. "Prediction of web usages using apriori algorithm." Int. J. Res. 6 (3): 663–669.

Hu, L. Y., M. W. Huang, S. W. Ke, and C. F. Tsai. 2016. "The distance function effect on k-nearest neighbor classification for medical datasets." SpringerPlus 5 (1): 1304. https://doi.org/10.1186/s40064-016-2941-7.

Hwang, B. G., X. Zhao, and S. Y. Ong. 2015. "Value management in Singaporean building projects: Implementation status, critical success factors, and risk factors." J. Manage. Eng. 31 (6): 04014094. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000342.

Ibadov, N. 2016. "Determination of the risk factors impact on the construction projects implementation using fuzzy sets theory." Acta Phys. Pol. A 130 (1): 107–111. https://doi.org/10.12693/APhysPolA.130.107.

IBM (International Business Machines). 2019. "Lift in an association rule." Accessed November 25, 2019. https://www.ibm.com/support/knowledgecenter/en/SSEPGG_11.1.0/com.ibm.im.model.doc/c_lift_in_an_association_rule.html.

Ibrahim, I. A., J. Hossain, and B. C. Duck. 2019. "An optimized offline random forests-based model for ultra-short-term prediction of PV characteristics." IEEE Trans. Ind. Inf. 16 (1): 202–214. https://doi.org/10.1109/TII.2019.2916566.

IHS (Information Handling Services). 2020. "Economics and country risk." Accessed May 7, 2020. https://www.ihs.com/industry/economics-country-risk.html.

Ilbeigi, M., B. Ashuri, and A. Joukar. 2017. "Time-series analysis for forecasting asphalt-cement price." J. Manage. Eng. 33 (1): 04016030. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000477.

Incident Prevention. 2016. "Dedicated to utility safety and operations professionals." Accessed October 4, 2020. http://online.incident-prevention.com/publication/?m=19389&i=290033&p=1&pp=1.

ISO. 2018. "ISO 31000: Risk management: Guidelines." Accessed July 21, 2019. https://www.iso.org/standard/65694.html.

Issuu 2018. "Leveraging data science and artificial intelligence at the University of Auckland." Accessed September 10, 2020. https://issuu.com/uniservices/docs/ai_uoa_email

Jaillon, L., and C. S. Poon. 2008. "Sustainable construction aspects of using prefabrication in dense urban environment: A Hong Kong case study." Construction management and Economic, 26 (9): 953–966.

Jain, A., K. Nandakumar, and A. Ross. 2005. "Score normalization in multimodal biometric systems." Pattern Recognit. 38 (12): 2270–2285. https://doi.org/10.1016/j.patcog.2005.01.012.

Jamali, A. A., T. O. Randhir, and J. Nosrati. 2018. "Site suitability analysis for subsurface dams using Boolean and fuzzy logic in arid watersheds." J. Water Resour. Plann. Manage. 144 (8): 04018047. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000947.

Janani, R., and S. Vijayarani. 2019. "Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization." Expert Syst. Appl. 134 (Dec): 192–200. https://doi.org/10.1016/j.eswa.2019.05.030.

Jang, S., and Lee, G. 2018. "Process, productivity, and economic analyses of BIM–based multi-trade prefabrication—A case study." Automation in Construction, 89, 86-98.

Jarkas, A. M. 2016a. "Effect of buildability on labor productivity: A practical quantification approach." J. Constr. Eng. Manage. 142 (2): 06015002. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001062.

Jarkas, A. M. 2016b. "Predicting contract duration for building construction: Is Bromilow's time-cost model a panacea?" J. Manage. Eng. 32 (1): 05015004. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000394.

Jebelli, H., B. Choi, and S. Lee. 2019. "Application of wearable biosensors to construction sites. I: Assessing workers' stress." J. Constr. Eng. Manage. 145 (12): 04019079. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001729.

Jia, H., S. Ding, X. Xu, and R. Nie. 2014. "The latest research progress on spectral clustering." Neural Comput. Appl. 24 (7–8): 1477–1486. https://doi.org/10.1007/s00521-013-1439-2.

Jia, Y., G. Wu, and S. Chang. 2011. "Principal component clustering analysis method applied to road traffic safety." In Proc., Int. Conf. on Transportation Engineering 2011, 2832–2837. Reston, VA: ASCE.

Johansen, S., 1991. "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models." Econometrica: journal of the Econometric Society, pp.1551-1580.

Johansen, S., 1995. "Likelihood-based inference in cointegrated vector autoregressive models." Oxford University Press on Demand.

Johari, S., and K. N. Jha. 2020a. "How the aptitude of workers affects construction labor productivity." J. Manage. Eng. 36 (5): 04020055. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000826.

Johari, S., and K. N. Jha. 2020b. "Impact of work motivation on construction labor productivity." J. Manage. Eng. 36 (5): 04020052. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000824.

Jones, K. 2018. "How technology is reshaping the construction industry." Accessed February 14, 2020. https://www.constructconnect.com/blog/construction-technology/technology-reshaping-construction-industry/.

Joseph, R.D. and Arockiamary, A., 2014. "Collective creativity by team idea mapping technique." IMPACT: International Journal of Research in Applied, Natural and Social Sciences, 2(7), pp.161-165.

Joshi, A., A. Bansal, A. S. Sabitha, and T. Choudhury. 2018. "An efficient way to find frequent patterns using graph mining and network analysis techniques on United States airports network." In Smart computing and informatics, 301–316. New York: Springer.

Jun, H., H. Xu, Y. Pei, and H. Ji. 2009. "Influencing factors contributing to rear-end collisions on highways: PCA based investigation and findings." In Proc., Int. Conf. of Chinese Transportation Professionals 2009: Critical Issues in Transportation Systems Planning, Development, and Management, 1–6. Reston, VA: ASCE.

Kadilar, G. Ö., and C. Kadilar. 2017. "Assessing air quality in Aksaray with time series analysis." In Proc., AIP Conf., 020112. College Park, MD: American Institute of Physics.

Kadimisetty, A. 2018. "Association rule mining in R." Accessed November 25, 2019. https://towardsdatascience.com/association-rule-mining-in-r-ddf2d044ae50.

Kajabi 2021. "What is Brainstorming, How to Brainstorm & 15 Effective Techniques". Accessed February 26, 2021. https://kajabi.com/blog/what-is-brainstorming-how-to-brainstorm-15-effective-techniques

Kamali, M., and K. Hewage. 2017. "Development of performance criteria for sustainability evaluation of modular versus conventional construction methods." Journal of Cleaner Production, 142, pp.3592-3606.

Kansas Department of Agriculture—Division of Water Resources. 2019. "Dam safety inspection report requirements and guidelines." Accessed November 15, 2019. https://agriculture.ks.gov/docs/default-source/dwr-ws-forms/dam-safety-inspection-report-requirements-and-guidelines.pdf?sfvrsn=8edfc11b_6.

Karimi, H., and H. Taghaddos. 2019. "The influence of craft workers' educational attainment and experience level in fatal injuries prevention in construction projects." Saf. Sci. 117 (1): 417–427. https://doi.org/10.1016/j.ssci.2019.04.022.

Karimi, H., T. R. Taylor, G. B. Dadi, P. M. Goodrum, and C. Srinivasan. 2018. "Impact of skilled labor availability on construction project cost performance." J. Constr. Eng. Manage. 144 (7): 04018057. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001512.

Katseff, J., Peloquin, S., and Wintner, T. 2020. "Reimagining infrastructure in the United States: How to build better." Accessed September 9, 2020. https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/reimagining-infrastructure-in-the-united-states-how-to-build-better

Kaushik, S. 2016. "Introduction to feature selection methods with an example (or how to select the right variables?)." Accessed September 20, 2019. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/.

Keating, A. 2019. "The Demographics of Washington's Workers." Accessed February 4, 2020. http://www.opportunityinstitute.org/research/post/workforce-demographics-2019/

Keller, J. M., D. Liu, and D. B. Fogel. 2016. Fundamentals of computational intelligence: Neural networks, fuzzy systems, and evolutionary computation. Hoboken, NJ: Wiley.

Kestel, S. 2013. "Vectorautoregressive-VAR Models and Cointegration Analysis." Accessed March 18, 2021. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.379.3191&rep=rep1&type=pdf

Khalef, R., El-adaway, I.H., Assaad, R. and Kieta, N., 2021. Contract Risk Management: A Comparative Study of Risk Allocation in Exculpatory Clauses and Their Legal Treatment. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 13(1), p.04520036.

Khan, S. A., A. D. Joy, S. M. Asaduzzaman, and M. Hossain. 2019. "An efficient sign language translator device using convolutional neural network and customized ROI segmentation." In Proc., 2019 2nd Int. Conf. on Communication Engineering and Technology (ICCET), 152–156. New York: IEEE. https://doi.org/10.1109/ICCET.2019.8726895.

Khasnabis, S., and Ramiz-Al-Assar. 1989. "Analysis of heavy truck accident data—Exposure based approach." J. Transp. Eng. 115 (3): 298–304. https://doi.org/10.1061/(ASCE)0733-947X(1989)115:3(298).

KhodaBandehLou, A., A. Parvishi, R. Taghifam, M. Lotfi, and A. Taleei. 2016. "Integrating earned value management with risk management to control the time-cost of the project." IIOAB J. 7 (S4): 114–119.

Khodeir, L. M., and A. H. M. Mohamed. 2015. "Identifying the latest risk probabilities affecting construction projects in Egypt according to political and economic variables. From January 2011 to January 2013." HBRC J. 11 (1): 129–135. https://doi.org/10.1016/j.hbrcj.2014.03.007.

Khuriwal, N., and N. Mishra. 2018. "Breast cancer diagnosis using deep learning algorithm." In Proc., 2018 Int. Conf. on Advances in Computing, Communication Control and Networking (ICACCCN), 98–103. New York: IEEE. https://doi.org/10.1109/ICACCCN.2018.8748777.

Kim, B. C., and K. F. Reinschmidt. 2011. "Combination of project cost forecasts in earned value management." J. Constr. Eng. Manage. 137 (11): 958–966. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000352.

Kim, C. W., and D. Shin. 2020. "Improved classification of fire accidents and analysis of periodicity for prediction of critical fire accidents." J. Korean Inst. Gas 24 (1): 56–65.

Kim, E., W. G. Wells Jr, and M. R. Duffey. 2003. "A model for effective implementation of earned value management methodology." Int. J. Project Manage. 21 (5): 375–382. https://doi.org/10.1016/S0263-7863(02)00049-2.

Kingma, D. P., and J. Ba. 2014. "Adam: A method for stochastic optimization." Preprint, submitted December 22, 2014. https://arxiv.org/abs/1412.6980.

Kisi, K. P., N. Mani, E. M. Rojas, and E. T. Foster. 2018. "Estimation of optimal productivity in labor-intensive construction operations: Advanced study." J. Constr. Eng. Manage. 144 (10): 04018097. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001551.

Kivilä, J., M. Martinsuo, and L. Vuorinen. 2017. "Sustainable project management through project control in infrastructure projects." Int. J. Project Manage. 35 (6): 1167–1183. https://doi.org/10.1016/j.ijproman.2017.02.009.

Ko, C. H., and M. Y. Cheng. 2007. "Dynamic prediction of project success using artificial intelligence." J. Constr. Eng. Manage. 133 (4): 316–324. https://doi.org/10.1061/(ASCE)0733-9364(2007)133:4(316).

Koch, M. J. 2017. Hiring practices and labor productivity. Abingdon, UK: Taylor & Francis.

Koo, T.K. and Li, M.Y., 2016. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research." Journal of chiropractic medicine, 15(2), pp.155-163.

KPMG. 2015. "Global construction survey 2015: Climbing the curve. KPMG international." Accessed July 22, 2019. https://assets.kpmg/content/dam/kpmg/pdf/2015/05/construction-survey-201502.pdf.

Kumar, D. 2019. "Top 5 advantages and disadvantages of decision tree algorithm." Accessed November 12, 2019. https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a.

Kuznetsova, A., A. Askarov, R. Gusmanov, A. Askarova, and P. Pypłacz. 2019. "Differentiation of labor productivity level and wages as a basis for changes in labor market." Pol. J. Manage. Stud. 2019 (20): 345–357.

Kwak, Y. H., and F. T. Anbari. 2010. Project management in government: An introduction to earned value management. IBM Center for the Business of Government Report. Washington, DC: IBM Center for the Business of Government.

Kwiatek, C. 2018. "Impact of spatial cognitive abilities on the effectiveness of augmented reality in construction and fabrication." Master's thesis, Department of Civil and Environmental Engineering, University of Waterloo.

Larsson, J., P. E. Eriksson, T. Olofsson, and P. Simonsson. 2014. "Industrialized construction in the Swedish infrastructure sector: Core elements and barriers." Construction Management and Economics, 32 (1–2): 83–96.

Lee, C., J. Won, and E. B. Lee. 2019a. "Method for predicting raw material prices for product production over long periods." J. Constr. Eng. Manage. 145 (1): 05018017. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001586.

Lee, D. E., Yi, C. Y., Lim, T. K., and Arditi, D. 2010. "Integrated simulation system for construction operation and project scheduling." Journal of computing in civil engineering, 24(6), 557-569.

Lee, K. Z.-Z., N. Jensen, D. R. Gillette, and D. T. Wittwer. 2019b. "Seismic deformation analysis of embankment dams using simplified total-stress approach." J. Geotech. Geoenviron. Eng. 145 (10): 04019076. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002135.

Lee, S. H., S. R. Thomas, and R. L. Tucker. 2004. "Effective practice utilization using performance prediction software." J. Constr. Eng. Manage. 130 (4): 576–585. https://doi.org/10.1061/(ASCE)0733-9364(2004)130:4(576).

Lee, S., S. Tae, S. Yoo, and S. Shin. 2016. "Impact of business portfolio diversification on construction company insolvency in Korea." J. Manage. Eng. 32 (3): 05016003. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000413.

Lee, W., K. Y. Lin, E. Seto, and G. C. Migliaccio. 2017. "Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction." Autom. Constr. 83 (Aug): 341–353. https://doi.org/10.1016/j.autcon.2017.06.012.

Leon, H., H. Osman, M. Georgy, and M. Elsaid. 2018. "System dynamics approach for forecasting performance of construction projects." J. Manage. Eng. 34 (1): 04017049. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000575.

Leung, I., M. Song, and C. Kam. 2017. "AEC project performance prediction and validation using the artificial neural network." In Computing in Civil Engineering 2017, 360–367. Reston, VA: ASCE.

Li, K., D. Qian, S. Huang, and X. Liang. 2016. "Analysis of traffic accidents on highways using latent class clustering." In Proc., COTA Int. Conf. of Transportation Professionals 2016, 1800–1810. Reston, VA: ASCE.

Li, L., Y. Chen, and Y. Deng. 2015. "Accident analysis based on traffic system stability in third-tier city of developing country." In Access management theories and practices, 89–107. Reston, VA: ASCE.

Li, Z., X. Feng, Z. Wu, C. Yang, B. Bai, and Q. Yang. 2019. "Classification of atrial fibrillation recurrence based on a convolution neural network with SVM architecture." IEEE Access 7: 77849–77856. https://doi.org/10.1109/ACCESS.2019.2920900.

Liao, C. W., and Y. H. Perng. 2008. "Data mining for occupational injuries in the Taiwan construction industry." Saf. Sci. 46 (7): 1091–1102. https://doi.org/10.1016/j.ssci.2007.04.007.

Liao, P. C., G. Lei, J. Xue, and D. Fang. 2015. "Influence of person-organizational fit on construction safety climate." J. Manage. Eng. 31 (4): 04014049. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000257.

Lijauco, F., Gajendran, T., Brewer, G., and Rasoolimanesh, S. M. 2020. "Impacts of Culture on Innovation Propensity in Small to Medium Enterprises in Construction." Journal of construction engineering and management, 146(3), 04019116.

Lim, J. N., Schultmann, F., and Ofori, G. 2010. "Tailoring competitive advantages derived from innovation to the needs of construction firms." Journal of construction engineering and management, 136(5), 568-580.

Limaye, P. 2019. "Modular Market Predicted to Grow to $215 Billion by 2025". Accessed January 16, 2020. https://www.curt.org/modular-market-predicted-to-grow-to-215-billion-by-2025/

Lin, D.-J., W.-D. Yu, C.-M. Wu, and T-M. Cheng. 2018. "Correlation between intellectual capital and business performance of construction industry–an empirical study in Taiwan." Int. J. Constr. Manage. 18 (3): 232–246. https://doi.org/10.1080/15623599.2017.1315528.

Lin, Y., Y. Ye, and J. Niu. 2011. "PCA-CA method for road traffic safety macroscopic evaluation and safety level classification." In Proc., Int. Conf. of Chinese Transportation Professionals 2011: Towards Sustainable Transportation Systems, 933–940. Reston, VA: ASCE.

Ling, F. Y. Y., S. L. Chan, E. Chong, and L. P. Ee. 2004. "Predicting performance of design-build and design-bid-build projects." J. Constr. Eng. Manage. 130 (1): 75–83. https://doi.org/10.1061/(ASCE)0733-9364(2004)130:1(75).

Ling, F. Y., S. P. Low, S. Wang, and T. Egbelakin. 2008. "Models for predicting project performance in China using project management practices adopted by foreign AEC firms." J. Constr. Eng. Manage. 134 (12): 983–990. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:12(983).

Lingard, H., M. Hallowell, R. Salas, and P. Pirzadeh. 2017. "Leading or lagging? Temporal analysis of safety indicators on a large infrastructure construction project." Saf. Sci. 91 (Jan): 206–220. https://doi.org/10.1016/j.ssci.2016.08.020.

Lingard, H., P. Pirzadeh, and D. Oswald. 2019. "Talking safety: Health and safety communication and safety climate in subcontracted construction workgroups." J. Constr. Eng. Manage. 145 (5): 04019029. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001651.

Lipke, W., O. Zwikael, K. Henderson, and F. Anbari. 2009. "Prediction of project outcome: The application of statistical methods to earned value management and earned schedule performance indexes." Int. J. Project Manage. 27 (4): 400–407. https://doi.org/10.1016/j.ijproman.2008.02.009.

Liu, B., T. Huo, J. Meng, J. Gong, Q. Shen, and T. Sun. 2016a. "Identification of key contractor characteristic factors that affect project success under different project delivery systems: Empirical analysis based on a group of data from China." J. Manage. Eng. 32 (1): 05015003. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000388.

Liu, G., Li, K., Zhao, D., and Mao, C. 2017. "Business model innovation and its drivers in the Chinese construction industry during the shift to modular prefabrication." Journal of management in engineering, 33(3), 04016051.

Liu, H. J., P. E. Love, M. C. Sing, and J. Smith. 2019a. "Ex post evaluation of economic infrastructure assets: Significance of regional heterogeneities in Australia." J. Infrastruct. Syst. 25 (2): 05019005. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000485.

Liu, J., X. Zhao, and P. Yan. 2016b. "Risk paths in international construction projects: Case study from Chinese contractors." J. Constr. Eng. Manage. 142 (6): 05016002. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001116.

Liu, M., H. Y. Chong, P. C. Liao, and L. Xu. 2019b. "Probabilistic-based cascading failure approach to assessing workplace hazards affecting human error." J. Manage. Eng. 35 (3): 04019006. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000690.

Liu, Q., G. Feng, N. Wang, and G. K. Tayi. 2018. "A multi-objective model for discovering high-quality knowledge based on data quality and prior knowledge." Inf. Syst. Front. 20 (2): 401–416. https://doi.org/10.1007/s10796-016-9690-6.

Loosemore, M., and Richard, J. 2015. "Valuing innovation in construction and infrastructure." Engineering, construction and architectural management. 22 (1): 38–53.

Lütkepohl, H., M. Krätzig, and P. C. B. Phillips. 2004. Applied time series econometrics. Cambridge, UK: Cambridge University Press.

Mahadevan, R., and J. Asafu-Adjaye. 2007. "Energy consumption, economic growth and prices: A reassessment using panel VECM for developed and developing countries." Energy Policy 35 (4): 2481–2490. https://doi.org/10.1016/j.enpol.2006.08.019.

Mahamid, I. 2013. "Contractors perspective toward factors affecting labor productivity in building construction." Eng. Constr. Archit. Manage. 2013 (Sep): 2. https://doi.org/10.1108/ECAM-08-2011-0074.

Manfreda, K. L., M. Bosnjak, J. Berzelak, I. Haas, and V. Vehovar. 2008. "Web surveys versus other survey modes: A meta-analysis comparing response rates." Int. J. Market Res. 50 (1): 79–104. https://doi.org/10.1177/147078530805000107.

Mao, C., Shen, Q., Pan, W. and Ye, K., 2015. "Major barriers to off-site construction: the developer's perspective in China." Journal of Management in Engineering, 31(3), p.04014043.

Market Research Future. 2017."Modular construction market research report—Forecast to 2023." Accessed February 1, 2021. https://www.marketresearchfuture.com/reports/modular-construction-market-1682.

Markets and Markets. 2019. "Modular Construction Market Worth $157.19 Billion by 2023 - Exclusive Report by MarketsandMarkets." Accessed December 11, 2019. https://www.prnewswire.com/news-releases/modular-construction-market-worth-157-19-billion-by-2023--exclusive-report-by-marketsandmarkets-300884022.html

Markov, S. 2018. "Team Idea mapping". Accessed Feb. 25, 2021. https://geniusrevive.com/en/team-idea-mapping/

Marle, F., and L. A. Vidal. 2016. Managing complex, high-risk projects. London: Springer.

Mateos, E. Y. M., M. A. L. Garrido, J. A. H. Aguilar, C. A. O. Ortiz, O. A. G. González, and A. C. Ferreira. 2019. "Visual association rules on the psychological connection of university students with their studies." Res. Comput. Sci. 148 (6): 263–276. https://doi.org/10.13053/rcs-148-6-20.

McGowan, J. 2019. "How has the growth of E-commerce sales affected retail real estate?" CMC Senior theses, The Robert Day School of Economics and Finance, Claremont McKenna College.

Mckinsey & Company 2019. "Modular construction: From projects to products." Accessed March 2, 2020. http://modular.org/documents/document_publication/mckinsey-report-2019.pdf

McManamay, R. A., C. O. Oigbokie, S. C. Kao, and M. S. Bevelhimer. 2016. "Classification of US hydropower dams by their modes of operation." River Res. Appl. 32 (7): 1450–1468. https://doi.org/10.1002/rra.3004.

Medina, J. C., S. Shen, and R. F. Benekohal. 2014. "Microscopic analysis for accident data at railroad grade crossings." In Proc., T&DI Congress 2014: Planes, Trains, and Automobiles, 366–375. Reston, VA: ASCE.

Meisels 2020. "2020 Engineering and Construction Industry Outlook: A midyear update." Accessed September 9, 2020. https://www2.deloitte.com/us/en/pages/energy-and-resources/articles/engineering-and-construction-industry-trends.html#

Meng, W. L., S. Shen, and A. Zhou. 2018. "Investigation on fatal accidents in Chinese construction industry between 2004 and 2016." Nat. Hazards 94 (2): 655–670. https://doi.org/10.1007/s11069-018-3411-z.

Miller, K., K. Costa, and D. Cooper. 2012. "Ensuring public safety by investigating in our nation's critical dams and levees." Accessed September 9, 2019. https://www.americanprogress.org/issues/economy/reports/2012/09/20/38299/ensuring-public-safety-by-investing-in-our-nations-critical-dams-and-levees/.

Mills, A. 2001. "A systematic approach to risk management for construction." Struct. Surv. 19 (5): 245–252. https://doi.org/10.1108/02630800110412615.

Miner, L., P. Bolding, J. Hilbe, M. Goldstein, T. Hill, R. Nisbet, N. Walton, and G. Miner. 2015. Practical predictive analytics and decisioning systems for medicine: Informatics accuracy and cost-effectiveness for healthcare administration and delivery including medical research. Cambridge, MA: Academic.

Mirhadi, M. 2018. "Adverse effects of shift work on labor productivity." Accessed May 5, 2020. https://www.adroitprojectconsultants.com/2018/03/24/adverse-effects-shiftwork-labor-productivity/.

Mirza, S., and Ali, M. S. 2017. "Infrastructure crisis—a proposed national infrastructure policy for Canada." Canadian Journal of Civil Engineering, 44(7), 539-548.

Mishra, P. N., S. Surendran, V. K. Gadi, R. A. Joseph, and D. N. Arnepalli. 2017. "Generalized approach for determination of thermal conductivity of buffer materials." J. Hazard. Toxic Radioact. Waste 21 (4): 04017005. https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000357.

Montella, A. 2011. "Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types." Accid. Anal. Prev. 43 (4): 1451–1463. https://doi.org/10.1016/j.aap.2011.02.023.

Moran, R. D. 1977. "The developing law of occupational safety and health." Okla. L. Rev. 30 (1): 354.

Mortaji, S. T. H., R. Noorossana, and M. Bagherpour. 2015. "Project completion time and cost prediction using change point analysis" J. Manage. Eng. 31 (5): 04014086. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000329.

Mostafavi, A., D. Abraham, S. Noureldin, G. Pankow, J. Novak, R. Walker, and B. George. 2013. "Risk-based protocol for inspection of transportation construction projects undertaken by state departments of transportation." J. Constr. Eng. Manage. 139 (8): 977–986. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000664.

Mousavi, S. M., Z. Zhang, S. A. Parr, A. Pande, and B. Wolshon. 2019. "Identifying high crash risk highway segments using jerk-cluster analysis." In Proc., Int. Conf. on Transportation and Development 2019: Smarter and Safer Mobility and Cities, 112–123. Reston, VA: ASCE.

Mukherjee, K. 2019. "Delineating fringe by global diversity value: A case study on Barasat municipality, India." In Proc., 2019 Int. Conf. on Big Data and Computational Intelligence (ICBDCI), 1–12. New York: IEEE. https://doi.org/10.1109/ICBDCI.2019.8686100.

Mulenga, S. 2020. "Mathematical modelling using gray markov SCGM (1, 1) c of Zambia's fatal mining accidents between 2001 and 2015." Accessed October 3, 2020. https://osf.io/wygve.

Nahmens, I. and Ikuma, L.H., 2012. "Effects of lean construction on sustainability of modular homebuilding." Journal of architectural engineering, 18(2), pp.155-163.

Nasirian, A., Arashpour, M., Abbasi, B., and Akbarnezhad, A. 2019. "Optimal Work Assignment to Multiskilled Resources in Prefabricated Construction." Journal of Construction Engineering and Management, 145(4), 04019011.

Nasirian, A., Arashpour, M., Abbasi, B., Zavadskas, E.K. and Akbarnezhad, A., 2019. "Skill set configuration in prefabricated construction: Hybrid optimization and multicriteria decision-making approach." Journal of Construction Engineering and Management, 145(9), p.04019050.

NCCER (National Center for Construction Education and Research). 2020. "Search Titles and Disciplines." Accessed March 2, 2020. https://www.nccer.org/workforce-development-programs/disciplines

NCSS. 2006. "Multiple regression with serial correlation." Accessed June 14, 2020. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Multiple_Regression_with_Serial_Correlation.pdf.

Negnevitsky, M. 2005. Artificial intelligence a guide to intelligent systems. Boston: Addison Wesley.

Neves, C., Matos, J. C., and Neves, L. 2016. "SUSTIMS–Sustainable Infrastructure Management System." In 1st European Road Infrastructure Congress.

Ng, A. Y., M. I. Jordan, and Y. Weiss. 2002. "On spectral clustering: Analysis and an algorithm." In Advances in neural information processing systems, 849–856. Cambridge, MA: MIT Press.

NIST/SEMATECH. 2018. "e-handbook of statistical methods." Accessed November 5, 2018. http://www.itl.nist.gov/div898/handbook/.

Noktehdan, M., Shahbazpour, M., Zare, M. R., and Wilkinson, S. 2019. "Innovation management and construction phases in infrastructure projects." Journal of Construction Engineering and Management, 145(2), 04018135.

Ntzeremes, P., K. Kirytopoulos, and G. Filiou. 2020. "Quantitative risk assessment of road tunnel fire safety: Improved evacuation simulation model." ASCE-ASME J. Risk Uncertainty Eng. Syst. Part A: Civ. Eng. 6 (1): 04019020. https://doi.org/10.1061/AJRUA6.0001029.

O'Connor, J. T., W. J. O'Brien, and J. O. Choi. 2013. "Industrial modularization how to optimize; how to maximize." Austin, TX: University of Texas at Austin.

O'Connor, J. T., W. J. O'Brien, and J. O. Choi. 2016. "Industrial project execution planning: Modularization versus stick-built." Practice periodical on structural design and construction, 21 (1): 04015014.

O'Connor, J.T., O'Brien, W.J. and Choi, J.O., 2014. "Critical success factors and enablers for optimum and maximum industrial modularization." Journal of Construction Engineering and Management, 140(6), p.04014012.

O'Connor, J.T., O'Brien, W.J. and Choi, J.O., 2015. "Standardization strategy for modular industrial plants." Journal of Construction Engineering and Management, 141(9), p.04015026.

O'Connor, J. T., and B. D. Mock. 2019. "Construction, commissioning, and startup execution: Problematic activities on capital projects." Journal of Construction Engineering and Management, 145 (4): 04019009.

Ogaji, D. S., E. O. Mabel, and A. D. Adesina. 2018. "Situational analysis of patient safety culture in public health institutions in South-South Nigeria." SM J. Public Health Epidemiol. 4 (1): 1049.

Oh, E. H., N. Naderpajouh, M. Hastak, and S. Gokhale. 2016. "Integration of the construction knowledge and expertise in front-end planning." J. Constr. Eng. Manage. 142 (2): 04015067. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001050.

Olawale, Y. A., and M. Sun. 2010. "Cost and time control of construction projects: Inhibiting factors and mitigating measures in practice." Constr. Manage. Econ. 28 (5): 509–526. https://doi.org/10.1080/01446191003674519.

Olawale, Y., and M. Sun. 2013. "PCIM: Project control and inhibiting-factors management model." J. Manage. Eng. 29 (1): 60–70. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000125.

OSHA (Occupational Safety and Health Administration). 2016. "The importance of root cause analysis during incident investigation." Accessed October 4, 2020. https://www.osha.gov/Publications/OSHA3895.pdf.

OSHA (Occupational Safety and Health Administration). 2017. "Top 10 most frequently cited standards." Accessed October 4, 2020. https://www.osha.gov/top10citedstandards.

OSHRC (Occupational Safety and Health Review Commission). 2020. "Secretary of labor v. summit Contractors, Inc." Accessed October 4, 2020. https://www.oshrc.gov/assets/1/18/05-0839.htm?7879.

Oskouie, P., Becerik-Gerber, B. and Soibelman, L., 2017. "Automated recognition of building façades for creation of As-Is Mock-Up 3D models." Journal of Computing in Civil Engineering, 31(6), p.04017059.

Ozorhon, B., and Oral, K. 2017. "Drivers of innovation in construction projects." Journal of construction engineering and management, 143(4), 04016118.

Ozturk, M., S. Durdyev, O. N. Aras, S. Ismail, and N. Banaitienė. 2020. "How effective are labor wages on labor productivity?: An empirical investigation on the construction industry of New Zealand." Technol. Econ. Dev. Economy. 26 (1): 258–270. https://doi.org/10.3846/tede.2020.11917.

Palisade. 2018. "RISK for risk analysis." Accessed October 5, 2018. http://www.palisade.com/risk/.

Pan, W. and Sidwell, R., 2011. "Demystifying the cost barriers to offsite construction in the UK." Construction Management and Economics, 29(11), pp.1081-1099.

Pan, W., L. Chen, and W. Zhan. 2019. "PESTEL analysis of construction productivity enhancement strategies: A case study of three economies." J. Manage. Eng. 35 (1): 05018013. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000662.

Papana, A., Kyrtsou, C., Kugiumtzis, D. and Diks, C., 2014. "Identifying causal relationships in case of non-stationary time series." Department of Economics of the University of Macedonia, Thessaloniki.

Park, H.M. and Jung, H.W., 2003. "Evaluating interrater agreement with intraclass correlation coefficient in SPICE-based software process assessment." In Third International Conference on Quality Software, 2003. Proceedings. (pp. 308-314). IEEE.

Patel, K. 2018. "What is frequent pattern mining (association) and how does it support business analysis?" Accessed November 24, 2019. https://www.dataversity.net/frequent-pattern-mining-association-support-business-analysis/#.

Pellicer, E., Yepes, V., Correa, C. L., and Alarcón, L. F. 2014. "Model for systematic innovation in construction companies." Journal of Construction Engineering and Management, 140(4), B4014001.

Pereira, E., S. Ahn, S. Han, and S. Abourizk. 2018. "Identification and association of high-priority safety management system factors and accident precursors for proactive safety assessment and control." Journal of Management in Engineering 34 (1): 04017041.

Perrenoud, A., B. C. Lines, J. Savicky, and K. T. Sullivan. 2017. "Using best-value procurement to measure the impact of initial risk-management capability on qualitative construction performance." J. Manage. Eng. 33 (5): 04017019. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000535.

Pirzadeh, P., and H. Lingard. 2017. "Understanding the dynamics of construction decision making and the impact on work health and safety." J. Manage. Eng. 33 (5): 05017003. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000532.

PMI (Project Management Institute). 2005. Practice standard for earned value management. Newton Square, PA: PMI.

PMI (Project Management Institute). 2013. A guide to the project management body of knowledge. 5th ed. Newtown Square, PA: PMI.

PMI (Project Management Institute). 2017a. "Success rates rise: Transforming the high cost of low performance." Accessed July 24, 2019. https://www.pmi.org/-/media/pmi/documents/public/pdf/learning/thought-leadership/pulse/pulse-of-the-profession-2017.pdf.

PMI (Project Management Institute). 2017b. A guide to the project management body of knowledge. 6th ed. Newtown Square, PA: PMI.

PMI (Project Management Institute). 2018. "Success in disruptive time: Expanding the value delivery landscape to address the high cost of law performance." Accessed July 22, 2019. https://www.pmi.org/-/media/pmi/documents/public/pdf/learning/thought-leadership/pulse/pulse-of-the-profession-2018.pdf.

Poch, M., and F. Mannering. 1996. "Negative binomial analysis of intersection-accident frequencies." J. Transp. Eng. 122 (2): 105–113. https://doi.org/10.1061/(ASCE)0733-947X(1996)122:2(105).

Prabhakaran, S. 2019. "Vector autoregression (VAR)—Comprehensive guide with examples in Python." Accessed February 16, 2020. https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/.

Qiao, L., L. Zhang, S. Chen, and D. Shen. 2018. "Data-driven graph construction and graph learning: A review." Neurocomputing 312 (1): 336–351. https://doi.org/10.1016/j.neucom.2018.05.084.

Qualtrics. 2019. "Sample size calculator." Accessed July 19, 2019. https://www.qualtrics.com/blog/calculating-sample-size/.

Rachev, S. T., S. Mittnik, F. J. Fabozzi, and S. M. Focardi. 2007. Financial econometrics: From basics to advanced modeling techniques. Hoboken, NJ: Wiley.

Ragaventhiran, J., and M. K. Kavithadevi. 2019. "Map-optimize-reduce: CAN tree assisted FP-growth algorithm for clusters based FP mining on Hadoop." Future Gener. Comput. Syst. 103 (Feb): 111–122.

Rah, J. E., R. P. Manger, A. D. Yock, and G. Y. Kim. 2016. "A comparison of two prospective risk analysis methods: Traditional FMEA and a modified healthcare FMEA." Med. Phys. 43 (12): 6347–6353. https://doi.org/10.1118/1.4966129.

Rahimi, A., G. Azimi, H. Asgari, and X. Jin. 2019. "Clustering approach toward large truck crash analysis." Transp. Res. Rec. 2673 (8): 73–85. https://doi.org/10.1177/0361198119839347.

Rahman, M. M., C. F. Ahmed, and C. K. S. Leung. 2019. "Mining weighted frequent sequences in uncertain databases." Inf. Sci. 479 (1): 76–100. https://doi.org/10.1016/j.ins.2018.11.026.

Ramaji, I.J., Memari, A.M. and Messner, J.I., 2017. "Product-oriented information delivery framework for multistory modular building projects." Journal of computing in civil engineering, 31(4), p.04017001.

Randiwela, P., and S. T. Wijayaratne. 2017. "Determinants of perceived brand globalness: FMCG and airline services." In Oxford Business and Economics Conf. 2017. Oxford, UK: Global Conference on Business & Economics (GCBE).

Randiwela, P., and S. T. Wijayaratne. 2017. "Determinants of perceived brand globalness: FMCG and airline services." In Oxford Business and Economics Conference 2017. Oxford, UK: Global Conference on Business & Economics (GCBE)

Raouf, A. 2020. "Theory of accident causes." Accessed October 5, 2020. http://www.ilocis.org/documents/chpt56e.htm.

Rasul, N., M. S. A. Malik, B. Bakhtawar, and M. J. Thaheem. 2019. "Risk assessment of fast-track projects: A systems-based approach." Int. J. Constr. Manage. https://doi.org/10.1080/15623599.2019.1602587.

Remold, L. E. 1989. "Simulation of nonsteady construction processes." Journal of Construction Engineering and Management, 115(2), 163-178.

Research and Markets 2018. "Global $150+ Billion Modular Construction Market Forecast to 2023". Accessed November 7, 2019. https://www.globenewswire.com/news-release/2018/12/20/1677222/0/en/Global-150-Billion-Modular-Construction-Market-Forecast-to-2023.html

Richman, J. S. 2011. "Multivariate neighborhood sample entropy: A method for data reduction and prediction of complex data." In Vol. 487 of Methods in enzymology, 397–408. Cambridge, MA: Academic.

RiskWise 2020. "Theories of accident causation." Accessed October 5, 2020. https://riskwise.biz/wp-content/uploads/2017/10/Theories-of-Accident-Causation.pdf.

Rojas, E. M., and P. Aramvareekul. 2003. "Is construction labor productivity really declining?" J. Constr. Eng. Manage. 129 (1): 41–46. https://doi.org/10.1061/(ASCE)0733-9364(2003)129:1(41).

Rousseeuw, P. J. 1987. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics, 20 (1): 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rudeli, N., A. Santilli, I. Puente, and E. Viles. 2017. "Statistical model for schedule prediction: Validation in a housing-cooperative construction database." J. Constr. Eng. Manage. 143 (11): 04017083. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001396.

Ruiz, A., and Guevara, J. 2020. "Environmental and economic impacts of road infrastructure development: Dynamic considerations and policies." Journal of Management in Engineering, 36(3), 04020006.

Russell, M. M., G. Howell, S. M. Hsiang, and M. Liu. 2013. "Application of time buffers to construction project task durations." J. Constr. Eng. Manage. 139 (10): 04013008. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000735.

Ryal-Net, M. B., and L. A. Kaduma. 2015. "Assessment of building information modelling (BIM) knowledge in the Nigerian construction industry." International Journal of Civil and Environmental Engineering, 15 (5): 1–10.

Ryu, J., J. Seo, H. Jebelli, and S. Lee. 2019. "Automated action recognition using an accelerometer-embedded wristband-type activity tracker." J. Constr. Eng. Manage. 145 (1): 04018114. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001579.

Sadafi, N., M. F. M. Zain, and M. Jamil. 2012. "Adaptable industrial building system: Construction industry perspective." J. Archit. Eng. 18 (2): 140–147. https://doi.org/10.1061/(ASCE)AE.1943-5568.0000075.

Saichi, T., S. Renaud, N. Bouaanani, and B. Miquel. 2019. "Effects of rock foundation roughness on the sliding stability of concrete gravity dams based on topographic surveys." J. Eng. Mech. 145 (7): 04019043. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001604.

Said, H., 2016. "Modeling and likelihood prediction of prefabrication feasibility for electrical construction firms." Journal of Construction Engineering and Management, 142(2), p.04015071.

Samuel, J. 2017. "Construction industry still struggles to bridge the performance gap: KPMG survey." Accessed July 24, 2019. https://home.kpmg/xx/en/home/media/press-releases/2017/10/construction-industry-still-struggles-to-bridge-the-performance-gap.html.

Santos, J. R. A. 1999. "Cronbach's alpha: A tool for assessing the reliability of scales." Journal of extension, 37(2), 1-5.

Savage, S., K. Douglas, R. Fell, W. Peirson, and R. Berndt. 2019. "Modeling the erosion and swelling of the sides of transverse cracks in embankment dams." J. Geotech. Geoenviron. Eng. 145 (5): 04019015. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002040.

Sayed, T., W. Abdelwahab, and F. Navin. 1995. "Identifying accident-prone locations using fuzzy pattern recognition." J. Transp. Eng. 121 (4): 352–358. https://doi.org/10.1061/(ASCE)0733-947X(1995)121:4(352).

Schalit, N., and J. Christie. 2011. "Maine's high-hazard dams lack inspection." Accessed September 10, 2019. https://bangordailynews.com/2011/08/24/news/state/half-of-high-hazard-dams-lack-state-inspection/.

Scikit-learn. 2019. "User guide." Accessed September 13, 2019. https://scikit-learn.org/stable/user_guide.html.

Setiani, Y., and M. Z. Abd Majid. 2019. "Safety practices and labour productivity in construction projects." Covenant J. Res. Built Environ. 7 (1): 84–95.

Sexton, D., I. H. El-adaway, M. Abdul Nabi, and A. H. El Hakea. 2020. "Using the simplified acquisition of base engineer requirements in construction projects: Contract administration guidelines." J. Leg. Aff. Dispute Resolut. Eng. Constr. 12 (3): 04520018. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000394.

Shafeeq, A. and Hareesha, K.S., 2012. "Dynamic clustering of data with modified k-means algorithm." In Proceedings of the 2012 conference on information and computer networks (pp. 221-225).

Shahandashti, S. M., and B. Ashuri. 2016. "Highway construction cost forecasting using vector error correction models." J. Manage. Eng. 32 (2): 04015040. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000404.

Shahandashti, S.M. and Ashuri, B., 2013. "Forecasting engineering news-record construction cost index using multivariate time series models." Journal of Construction Engineering and Management, 139(9), pp.1237-1243.

Shao, B., Z. Hu, Q. Liu, S. Chen, and W. He. 2019. "Fatal accident patterns of building construction activities in China" Saf. Sci. 111 (2): 253–263. https://doi.org/10.1016/j.ssci.2018.07.019.

Shi, Y., B. Wu, N. Chen, A. Chen, J. Li, and H. Li. 2019. "Determination of effective management strategies for scenic area emergencies using association rule mining." Int. J. Disaster Risk Reduct. 39 (Oct): 101208. https://doi.org/10.1016/j.ijdrr.2019.101208.

Shifera, A. 2019. Modeling and forecasting volatility of coffee and gold export price in Ethiopia using GARCH model. Nekemte, Ethiopia: Wollega Univ.

Shiha, A., E. M. Dorra, and K. Nassar. 2020. "Neural networks model for prediction of construction material prices in Egypt using macroeconomic indicators." J. Constr. Eng. Manage. 146 (3): 04020010. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001785.

Shin, M., H. S. Lee, M. Park, M. Moon, and S. Han. 2014. "A system dynamics approach for modeling construction workers' safety attitudes and behaviors." Accid. Anal. Prev. 68 (Oct): 95–105. https://doi.org/10.1016/j.aap.2013.09.019.

Shinnou, H., and M. Sasaki. 2008. "Spectral clustering for a large data set by reducing the similarity matrix size." In LREC. Paris: European Language Resources Association.

Shohet, I. M., H. H. Wei, M. J. Skibniewski, B. Tak, and M. Revivi. 2019. "Integrated communication, control, and command of construction safety and quality." J. Constr. Eng. Manage. 145 (9): 04019051. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001679.

Shrestha, B. K., Choi, J. O., Shrestha, P. P., Lim, J., and Nikkhah Manesh, S. 2020. "Employment and Wage Distribution Investigation in the Construction Industry by Gender." Journal of Management in Engineering, 36(4), 06020001.

Shrout, P.E. and Fleiss, J.L., 1979. "Intraclass correlations: uses in assessing rater reliability." Psychological bulletin, 86(2), p.420.

Singh, A. 2018. "A multivariate time series guide to forecasting and modeling (with Python codes)." Accessed February 17, 2020. https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/.

Singh, B. K. 2019a. "Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm." Biocybern. Biomed. Eng. 39 (2): 393–409. https://doi.org/10.1016/j.bbe.2019.03.001.

Singh, R. P., R. Kataria, and S. Singhal. 2019. "Decision-making in real-life industrial environment through graph theory approach." In Computer architecture in industrial, biomechanical and biomedical engineering. London: IntechOpen.

Singh, V. K. 2019b. "Proposing pattern growth methods for frequent pattern mining on account of its comparison made with the candidate generation and test approach for a given data set." In Software engineering, 203–209. New York: Springer.

Slaughter, E. S. 1998. "Models of construction innovation." Journal of Construction Engineering and management, 124(3), 226-231.

Sodiq Olawale, A. 2019. Impact of corruption on economic growth in Nigeria. Lagos, Nigeria: Lagos State Univ.

Soekiman, A., K. S. Pribadi, B. W. Soemardi, and R. D. Wirahadikusumah. 2011. "Factors relating to labor productivity affecting the project schedule performance in Indonesia." Procedia Eng. 14 (Jan): 865–873. https://doi.org/10.1016/j.proeng.2011.07.110.

Solorio-Fernández, S., J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad. 2020. "A review of unsupervised feature selection methods." Artif. Intell. Rev. 53 (2): 907–948. https://doi.org/10.1007/s10462-019-09682-y.

Srour, F.J., Srour, I. and Lattouf, M.G., 2017. "A survey of absenteeism on construction sites." International Journal of Manpower, 38 (4): 533–547.

Stamatelatos, M., Dezfuli, H., Apostolakis, G., Everline, C., Guarro, S., Mathias, D., Mosleh, A., Paulos, T., Riha, D., Smith, C. and Vesely, W., 2011. Probabilistic risk assessment procedures guide for NASA managers and practitioners. Rep. No. NASA/SP-2011-3421. Washington, DC: NASA Headquarters.

Stannard, L. 2020. "Construction Technology to Watch in 2020." Accessed September 10, 2020. https://www.bigrentz.com/blog/construction-technology

Stanslaus, V. 2017. "The causality effects of macroeconomic factors on economic growth in tanzania." Doctoral dissertation, Dept. of Economics, Open Univ. of Tanzania.

Sterman, J. D. 2000. Business dynamics: Systems thinking and modeling for a complex world. No. HD30. 2 S7835. Boston: McGraw-Hill.

Sudman, S. 1983. "Applied sampling." In Handbook of survey research, edited by P. H. Rossi, J. D. Wright, and A. B. Anderson. Bingley, UK: Emerald Publishing Group.

Sundukovskiy, S. 2018. "Using technology to deal with the construction labor shortage." Accessed February 14, 2020. https://consupt.com/2018/03/using-technology-to-deal-with-construction-labor-shortage/.

Sunindijo, R. Y., and I. Kamardeen. 2017. "Work stress is a threat to gender diversity in the construction industry." J. Constr. Eng. Manage. 143 (10): 04017073. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001387.

Sunindijo, R. Y., and P. X. Zou. 2012. "Political skill for developing construction safety climate." J. Constr. Eng. Manage. 138 (5): 605–612. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000482.

Suprayitno, H., and Soemitro, R. A. A. 2018. "Preliminary Reflexion on Basic Principle of Infrastructure Asset Management." Jurnal Manajemen Aset Infrastruktur & Fasilitas, 2(1).

Susymary, J., and R. Lawrance. 2017. "Graph theory analysis of protein-protein interaction network and graph based clustering of proteins linked with Zika virus using MCL algorithm." In Proc., 2017 Int. Conf. on Circuit, Power and Computing Technologies (ICCPCT), 1–7. New York: IEEE.

Sveikauskas, L., S. Rowe, J. D. Mildenberger, J. Price, and A. Young. 2018. "Measuring productivity growth in construction." Mon. Lab. Rev. Accessed May 5, 2020. https://www.bls.gov/opub/mlr/2018/article/measuring-productivity-growth-in-construction.htm.

Sveikauskas, L., S. Rowe, J. Mildenberger, J. Price, and A. Young. 2016. "Productivity growth in construction." J. Constr. Eng. Manage. 142 (10): 04016045. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001138.

Swei, O. 2020. "Forecasting infidelity: Why current methods for predicting costs miss the mark." J. Constr. Eng. Manage. 146 (2): 04019100. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001756.

Swei, O., J. Gregory, and R. Kirchain. 2017. "Probabilistic approach for long-run price projections: Case study of concrete and asphalt." J. Constr. Eng. Manage. 143 (1): 05016018. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001211.

Tabish, S. Z. S., and K. N. Jha. 2018. "Beyond the iron triangle in public construction projects." J. Constr. Eng. Manage. 144 (8): 04018067. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001517.

Taherdoost, H. 2017. "Determining sample size: How to calculate survey sample size." Int. J. Econ. Manage. Syst. 2: 201.

Tai, H.W., Chen, J.H., Cheng, J.Y., Wei, H.H., Hsu, S.C. and Liu, H.C., 2021. "Determining Worker Training Time for Precast Component Production in Construction: Empirical Study in Taiwan." Journal of Construction Engineering and Management, 147(1), p.05020023.

Tan, Y., L. Shen, C. Langston, W. Lu, and M. CH Yam. 2014. "Critical success factors for building maintenance business: A Hong Kong case study." Facilities 32 (5/6): 208–225.

Tasneem, T., T. Tasneem, and M. M. J. Kabir. 2019. "Performance analysis of classical and evolutionary algorithms for mining association rules. In Proc., 2019 Int. Conf. on Electrical, Computer and Communication Engineering (ECCE), 1–6. New York: IEEE.

Tata and Howard. 2016. "Dam safety and the criticality of emergency action plans." Accessed September 15, 2019. https://tataandhoward.com/dam-safety-and-the-criticality-of-emergency-action-plans/.

Tatum, C. B. 1989. "Organizing to increase innovation in construction firms." Journal of construction engineering and management, 115(4), 602-617.

Tatum, C. B. 2005. "Building better: technical support for construction." Journal of construction engineering and management, 131(1), 23-32.

Tereso, A., P. Ribeiro, and M. Cardoso. 2018. "An automated framework for the integration between EVM and risk management." J. Inf. Syst. Eng. Manage. 3 (1): 03.

Terry, S. B., and G. Lucko. 2012. "Algorithm for time-cost trade/off analysis in construction projects by aggregating activity-level singularity functions." In Proc., 2012 Construction Research Congress, 226–235. Reston, VA: ASCE.

The American Institute of Architects and the National Institute of Building Sciences 2019. "Design For Modular Construction: An Introduction For Architects." Accessed March 2, 2020. content.aia.org/sites/default/files/2019-03/Materials_Practice_Guide_Modular_Construction.pdf

The National Academies Press 2018. "Data Science for Undergraduates: Opportunities and Options." Accessed September 10, 2020. https://www.nap.edu/catalog/25104/data-science-for-undergraduates-opportunities-and-options

Tixier, A. J. P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016. "Application of machine learning to construction injury prediction." Autom. Constr. 69 (May): 102–114. https://doi.org/10.1016/j.autcon.2016.05.016.

Too, E. G. 2010. "A framework for strategic infrastructure asset management." In Definitions, concepts and scope of engineering asset management (pp. 31-62). Springer, London.

Topak, F., M. K. Pekeriçli, and A. M. Tanyer. 2018. "Technological viability assessment of Bluetooth low energy technology for indoor localization." J. Comput. Civ. Eng. 32 (5): 04018034. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000778.

Trinajstic, N. 2018. Chemical graph theory. Abingdon, UK: Routledge.

Tuulberg, L. 2018. "The future of construction will be offsite and wood will be an important part of the change." Accessed October 10, 2019. https://medium.com/welement/the-future-of-construction-will-be-offsite-and-wood-will-be-an-important-part-of-the-change-76d305e417f2.

US Bureau of Economic Analysis. 2018a. "GDP by industry." Accessed November 22, 2019. https://apps.bea.gov/iTable/iTable.cfm?ReqID=51&hx0026%3Bstep=1#reqid=51&hx0026;step=51&hx0026;isuri=1&hx0026;5114=a&hx0026;5102=1.

US Bureau of Economic Analysis. 2018b. "Interactive access to industry economic accounts data: GDP by industry." Accessed November 22, 2019. https://apps.bea.gov/iTable/iTable.cfm?ReqID=51&hx0026%3Bstep=1#reqid=51&hx0026;step=51&hx0026;isuri=1&hx0026;5114=a&hx0026;5102=1.

US Bureau of Labor Statistics. 2019. "Employment projection: Employment by major industry sector." Accessed November 17, 2019. https://www.bls.gov/emp/tables/employment-by-major-industry-sector.htm.

US Bureau of Labor Statistics. 2020a. "Industries at a glance: Construction: NAICS 23." Accessed March 18, 2020. https://www.bls.gov/iag/tgs/iag23.htm.

US Bureau of Labor Statistics 2020b. "Occupational Employment Statistics." Accessed March 2, 2020. https://www.bls.gov/oes/2018/may/oes_stru.htm

USACE (United States Army Corps of Engineers). 2019. "National inventory of dams." Accessed September 14, 2019. http://nid.usace.army.mil.

Vaagen, H., M. Kaut, and S. W. Wallace. 2017. "The impact of design uncertainty in engineer-to-order project planning." Eur. J. Oper. Res. 261 (3): 1098–1109. https://doi.org/10.1016/j.ejor.2017.03.005.

Vanhoucke, M. 2012. Project management with dynamic scheduling. Berlin: Springer.

Vereen, S. C. 2013. "Forecasting skilled labor demand in the US construction industry." Ph.D. dissertation, Dept. of Civil, Construction, and Environmental Engineering, North Carolina State Univ.

Vereen, S. C., W. Rasdorf, and J. E. Hummer. 2016. "Development and comparative analysis of construction industry labor productivity metrics." J. Constr. Eng. Manage. 142 (7): 04016020. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001112.

Verma, A., S. D. Khan, J. Maiti, and O. B. Krishna. 2014. "Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports." Saf. Sci. 70 (Jun): 89–98. https://doi.org/10.1016/j.ssci.2014.05.007.

Vincent, D. R., N. Deepa, D. Elavarasan, K. Srinivasan, S. H. Chauhdary, and C. Iwendi. 2019. "Sensors driven AI-based agriculture recommendation model for assessing land suitability." Sensors 19 (17): 3667. https://doi.org/10.3390/s19173667.

Viswanathan, S., N. Damodaran, A. Simon, A. George, M. A. Kumar, and K. P. Soman. 2019. "Detection of duplicates in Quora and Twitter corpus." In Advances in big data and cloud computing, 519–528. Singapore: Springer.

Von Luxburg, U. 2007. "A tutorial on spectral clustering." Stat. Comput. 17 (4): 395–416. https://doi.org/10.1007/s11222-007-9033-z.

Vose, D. 1996. Quantitative risk analysis: A guide to Monte Carlo simulation modelling. Chichester, UK: Wiley.

Wang, H., and S. Sun. 2011. "Accident causation chain analysis of ship collisions based on Bayesian networks." In Proc., Int. Conf. of Chinese Transportation Professionals 2011: Towards Sustainable Transportation Systems, 3944–3953. Reston, VA: ASCE.

Wang, L., and S. Ding. 2018. "Vector auto regression and envelope model." Stat 7 (1): e203. https://doi.org/10.1002/sta4.203.

Wang, L., Q. F. Lin, Z. Y. Wu, and B. Yu. 2020a. "A data-driven estimation of driving style using deep clustering." In Proc., COTA Int. Conf. of Transportation Professionals 2020, 4183–4194. Reston, VA: ASCE.

Wang, M., J. Chen, L. Wu, and B. Song. 2018. "Hydrodynamic pressure on gravity dams with different heights and the Westergaard correction formula." Int. J. Geomech. 18 (10): 04018134. https://doi.org/10.1061/(ASCE)GM.1943-5622.0001257.

Wang, S. Q., C. Gao, Q. Zhang, H. Y. Zeng, and J. Bai. 2020b. "The latest research on clustering algorithms used for radar signal sorting." In Recent trends in intelligent computing, communication and devices, 799–805. New York: Springer.

Wang, S., G. Hua, G. Hao, and C. Xie. 2017a. "A cycle deep belief network model for multivariate time series classification." Math. Prob. Eng. 2017 (Jan): 1–7.

Wang, X., D. Gui, H. Li, and H. Gui. 2019a. "Automatic construction of coal mine accident ontology." In Proc., Int. Conf. on Applications and Techniques in Cyber Security and Intelligence, 1366–1374. New York: Springer.

Wang, X., N. Xia, Z. Zhang, C. Wu, and B. Liu. 2017b. "Human safety risks and their interactions in China's subways: Stakeholder perspectives." J. Manage. Eng. 33 (5): 05017004. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000544.

Wang, X., X. Wang, and D. M. Wilkes. 2019b. Machine learning-based natural scene recognition for mobile robot localization in an unknown environment. New York: Springer.

Wang, Y., and Q. Chen. 2013. "Study on traffic safety evaluation based on traffic conflict technique and gray clustering at signalized intersection." In Proc., ICTIS 2013: Improving Multimodal Transportation Systems-Information, Safety, and Integration, 1311–1317. Reston, VA: ASCE.

Wei, H. H., M. Liu, M. J. Skibniewski, and V. Balali. 2016. "Prioritizing sustainable transport projects through multicriteria group decision making: Case study of Tianjin Binhai New Area, China." J. Manage. Eng. 32 (5): 04016010. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000449.

Wei, W. 2016. "Achieving inclusive growth in China through vertical specialization." Cambridge, UK: Chandos.

Wen, L., J. Chai, Z. Xu, Y. Qin, and Y. Li. 2019. "Comparative and numerical analyses of response of concrete cutoff walls of earthen dams on alluvium foundations." J. Geotech. Geoenviron. Eng. 145 (10): 04019069. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002132.

Wilks, D. S. 2011. Vol. 100 of Statistical methods in the atmospheric sciences. Oxford: Academic Press.

Wilson, J. 2020. "10 effective brainstorming techniques for teams". Accessed February 26, 2021. https://www.wework.com/ideas/worklife/effective-brainstorming-techniques

Wilson, R. J. 1972. "Introduction to graph theory, 1972." In Oliver and Boyd. Upper Saddle River, NJ: Prentice Hall.

Wong, J.M. and Ng, S.T., 2010. "Forecasting construction tender price index in Hong Kong using vector error correction model." Construction management and Economics, 28(12), pp.1255-1268.

Wong, L., Y. Wang, T. Law, and C. T. Lo. 2016. "Association of root causes in fatal fall-from-height construction accidents in Hong Kong." J. Constr. Eng. Manage. 142 (7): 04016018. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001098.

Wu, C. L., D. P. Fang, P. C. Liao, J. W. Xue, Y. Li, and T. Wang. 2015. "Perception of corporate social responsibility: The case of Chinese international contractors." J. Cleaner Prod. 107 (Nov): 185–194. https://doi.org/10.1016/j.jclepro.2015.04.143.

Wu, H., D. Wu, and J. Zhao. 2019. "An intelligent fire detection approach through cameras based on computer vision methods." Process Saf. Environ. Protect. 127 (1): 245–256. https://doi.org/10.1016/j.psep.2019.05.016.

Wu, W. 2020. "Construction of ecological monitoring and early warning system in coal mine based on big data analysis." Feb Fresenius Environ. Bull. 29 (5): 3564.

Wuellner, T., S. Feuerstack, and A. Hahn. 2019. "Clustering environmental conditions of historical accident data to efficiently generate testing sceneries for maritime systems." In Proc., Int. Symp. on Model-Based Safety and Assessment, 349–362. New York: Springer.

Xia, N., Zou, P. X., Liu, X., Wang, X., and Zhu, R. 2018. "A hybrid BN-HFACS model for predicting safety performance in construction projects." Safety science, 101, 332-343.

Xie, S., S. Dong, and G. Zhang. 2019. "Identification of key factors of fire risk of oil depot based on fuzzy clustering algorithm." In Vol. 59001 of Proc., Pressure Vessels and Piping Conf. New York: ASME.

Xie, S., X. Ji, W. Yang, and C. Hu. 2020. "Analysis of traffic accident characteristic and difference in two-lane plateau mountain highways." In Proc., COTA Int. Conf. of Transportation Professionals 2020, 4408–4419. Reston, VA: ASCE.

Xu, B., and B. Lin. 2016. "Assessing CO2 emissions in China's iron and steel industry: A dynamic vector auto regression model." Appl. Energy 161 (Jan): 375–386. https://doi.org/10.1016/j.apenergy.2015.10.039.

Xu, C., J. Bao, C. Wang, and P. Liu. 2018. "Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China" J. Saf. Res. 67 (Dec): 65–75. https://doi.org/10.1016/j.jsr.2018.09.013.

Xu, Y., L. Shi, F. Huang, L. Zhang, Y. Lu, and Y. Wang. 2019. "Recognition of stores' relationship based on constrained spectral clustering." In Proc., 2019 3rd Int. Conf. on Innovation in Artificial Intelligence, 111–115. New York: Association for Computing Machinery.

Yan, C., J. Liang, M. Zhao, X. Zhang, T. Zhang, and H. Li. 2019. "A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy." Anal. Chim. Acta 1080 (Nov): 35–42. https://doi.org/10.1016/j.aca.2019.07.012.

Yang, H., H. Ding, and G. Li. 2014. "Orderly clustering division of highways considering the accident four indicators." In Proc. ICLEM 2014: System Planning, Supply Chain Management, and Safety, 212–218. Reston, VA: ASCE.

Yang, X., and B. Liu. 2019. "Uncertain time series analysis with imprecise observations." Fuzzy Optim. Decis. Making 18 (3): 263–278. https://doi.org/10.1007/s10700-018-9298-z.

Yang, Z., Y. Yuan, M. Zhang, X. Zhao, and B. Tian. 2019. "Assessment of construction workers' labor intensity based on wearable smartphone system." J. Constr. Eng. Manage. 145 (7): 04019039. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001666.

Yao, L., Z. Xu, X. Zhou, and B. Lev. 2019. "Synergies between association rules and collaborative filtering in recommender system: An application to auto industry." In Data science and digital business, 65–80. Cham, Switzerland: Springer.

Yates, J. K. 2014. "Design and construction for sustainable industrial construction." J. Constr. Eng. Manage. 140 (4): B4014005. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000673.

Yi, J. S., Y. W. Kim, K. A. Kim, and B. Koo. 2012. "A suggested color scheme for reducing perception-related accidents on construction work sites." Accid. Anal. Prev. 48 (Sep): 185–192. https://doi.org/10.1016/j.aap.2011.04.022.

Yi, W., and A. P. Chan. 2014. "Critical review of labor productivity research in construction journals." J. Manage. Eng. 30 (2): 214–225. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000194.

Yuan, C., H. Cui, W. Wang, and S. Ma. 2019. "Cause factors in emergency process of fire accidents for oil and gas storage and transportation systems based on ISM and AHP." J. Hazard. Toxic Radioact. Waste 23 (2): 04018038. https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000432.

Yuan, C., S. Ma, Y. Hu, Y. Zhang, and T. Zuo. 2020. "Scenario deduction on fire accidents for oil–gas storage and transportation based on case statistics and a dynamic bayesian network." J. Hazard. Toxic Radioact. Waste 24 (3): 04020004. https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000495.

Zeb, A., A. Qudoos, and H. Hanif. 2015. "Identification and analysis of factors affecting machinery in the construction industry of Pakistan." Int. J. Sci.: Basic Appl. Res. 19 (1): 269–278.

Zhang, L. L., J. Lu, and Y. F. Ai. 2014. "Analysis and prediction on combination patterns of human factors for maritime accidents." In Proc., COTA Int. Conf. of Transportation Professionals 2014: Safe, Smart, and Sustainable Multimodal Transportation Systems, 2313–2322. Reston, VA: ASCE.

Zhang, M., and D. Fang. 2013. "A continuous behavior-based safety strategy for persistent safety improvement in construction industry." Autom. Constr. 34 (Jan): 101–107. https://doi.org/10.1016/j.autcon.2012.10.019.

Zhang, Q.-L., D.-Y. Li, L. Hu, and C. Hu. 2019a. "Viscous damping and contraction joint friction in underwater explosion resistant design of arch dams." J. Perform. Constr. Facil. 33 (3): 04019020. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001274.

Zhang, S., Rong, X., Bakhtawar, B., Tariq, S. and Zayed, T., 2021. "Assessment of Feasibility, Challenges, and Critical Success Factors of MiC Projects in Hong Kong." Journal of Architectural Engineering, 27(1), p.04020047.

Zhang, W., Z. Zhou, L. Li, and R. Huang. 2019b. "Identifying significant injury severity risk factors in traffic accidents based on the machine learning methods." In Proc., COTA Int. Conf. of Transportation Professionals 2019, 3759–3770. Reston, VA: ASCE.

Zhang, Y., J. Hou, V. Towhidlou, and M. Shikh-Bahaei. 2019c. "A neural network prediction based adaptive mode selection scheme in full-duplex cognitive networks." In Proc., IEEE Trans. on Cognitive Communications and Networking. New York: IEEE. https://doi.org/10.1109/TCCN.2019.2911005.

Zhang, Y., Lei, Z., Han, S., Bouferguene, A. and Al-Hussein, M., 2020. "Process-oriented framework to improve modular and offsite construction manufacturing performance." Journal of Construction Engineering and Management, 146(9), p.04020116.

Zhao, L., and G. Shi. 2019. "Maritime anomaly detection using density-based clustering and recurrent neural network." J. Navig. 72 (4): 894–916. https://doi.org/10.1017/S0373463319000031.

Zhao, T., and J. M. Dungan. 2018. "Quantifying lost labor productivity in domestic and international claims." J. Leg. Aff. Dispute Resolut. Eng. Constr. 10 (3): 04518013. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000269.

Zhao, Y., and X. Jiang. 2019. "Long-short memory neural network for short-term high-speed rail passenger flow forecasting." In Proc., RailNorrköping 2019: 8th Int. Conf. on Railway Operations Modelling and Analysis (ICROMA), 1264–1278. Linköping, Sweden: Linköping University Electronic Press.

Zheng, W. and Jin, M., 2020. "The effects of class imbalance and training data size on classifier learning: an empirical study." SN Computer Science, 1(2), pp.1-13.

Zhong, L., Y. Chen, X. Sun, X. Liu, and Y. He. 2007. "Research on section division of freeway with ordinal clustering method." In Proc., Int. Conf. on Transportation Engineering 2007, 4171–4177. Amsterdam, Netherlands: Elsevier.

Zhou, X., S. Chi, and Y. Jia. 2019. "Wetting deformation of core-wall rockfill dams." Int. J. Geomech. 19 (8): 04019084. https://doi.org/10.1061/(ASCE)GM.1943-5622.0001444.

Zhou, Z., and J. Irizarry. 2016. "Integrated framework of modified accident energy release model and network theory to explore the full complexity of the Hangzhou subway construction collapse." J. Manage. Eng. 32 (5): 05016013. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000431.

VITA

Rayan Hassane Assaad was born in Brazil. He earned a master's degree in Engineering Management (GPA 4.0) in 2018, a bachelor's degree in Civil Engineering (GPA 4.0) in 2016, and a Minor in Mathematics in 2016 from the ABET-accredited American University of Beirut, Lebanon. During his undergraduate studies, he did his internship as a Research Assistant at the geotechnical research lab within the Civil and Environmental Engineering department at the University of Illinois Urbana-Champaign, USA. Parallel to his masters, he worked as a Project Manager for Lys Royal General Enterprises—one of the top medium-size contractors in the Pearl City, Qatar—for around two years. After completing his work with Lys Royal General Enterprises, he enrolled in a PhD program at Missouri University of Science and Technology, USA where he received his PhD in Civil Engineering with a concentration in Construction Engineering and Management in July 2021. He also worked as a teaching and research assistant at the American University of Beirut and at Missouri University of Science and Technology.

During his PhD program, he was a recipient of the esteemed College of Engineering and Computing Dean's PhD Scholar Award at Missouri University of Science and Technology. In 2019, he was selected as an outstanding reviewer by the ASCE's Journal of Management in Engineering. In addition, he provided services, volunteering, and extracurricular activities during his academic journey by being actively involved in different organizations, councils, committees, and student clubs. By the date of his defense, he had 25 publications: 17 peer-reviewed journal papers and 8 peer-reviewed conference papers.