Original Research **Open Access**

Fuad Al-Bataineh, Ahmed Ali Khatatbeh, Yazan Alzubi*

Unsupervised machine learning for identifying key risk factors contributing to construction delays

DOI 10.2478/otmcj-2024-0014
Received: August 31, 2023; accepted: June 26, 2024

Abstract: The present study uses unsupervised machine learning capabilities with an emphasis on K-means clustering for addressing the problem of construction delays. The primary objective is to investigate the critical risk factors that contribute to such delays, thereby enabling more efficient risk-management strategies. The study employs a large dataset compiled from contracting firms operating in developing regions. This information is a vital resource for identifying crucial risk variables. These variables are analysed and categorised using the Likert scale into five levels based on their potential influence. This systematic approach permits the development of a comprehensive understanding of the relevant factors. These risk factors are grouped to enhance comprehension of the intricate risk landscape using K-means clustering. This allows for a broader, more comprehensive understanding of the factors contributing to construction delays. The application of K-means clustering demonstrates the potential of machine learning techniques to improve conventional approaches to risk management. This empirical investigation significantly expands the existing body of construction risk-management knowledge. It offers invaluable insights into various project stakeholders, allowing for more informed decision-making. Notably, the clustering analysis results provide a practical, user-friendly tool. This tool can assist project managers in enhancing their risk foresight, drafting more effective plans and developing robust mitigation strategies. Consequently, this research offers the potential for substantial improvements in project timeline adherence, thereby substantially reducing the impact of construction delays in developing nations.

Keywords: Machine learning, Unsupervised learning, Construction management, Project management, Delays

1 Introduction

Indeed, the construction industry is essential to the economy, especially in developing nations. In this industry, delays frequently negatively affect project costs, stakeholder relationships and overall project success. In order to ensure effective risk management and on-time project delivery, it is essential to identify the underlying causes of these delays. The literature on this subject is vast and diverse. Kassem et al. (2020) discussed risk factors affecting Yemen's oil and gas construction projects. Abd Karim et al. (2012) identified significant risk factors from a contractor's perspective. Ahmed et al. (2002) conducted an empirical study of construction delays in Florida, whereas Aibinu and Odeyinka (2006) examined the causes of such delays in Nigeria. Akomah and Jackson (2016) investigated road project delays, and Al Zubaidi and Al Otaibi (2008) identified risk factors for time overruns in Kuwait. Alajmi and Ahmed Memon (2022) analysed factors causing delays in Saudi Arabian projects, whereas Ali et al. (2010) investigated commercial projects in Malaysia. Al-Momani (2000) provided a quantitative analysis, and Alshihri et al. (2022) identified the risk factors contributing to time and cost overruns in Saudi Arabian construction projects. Cheng and Darsa (2021) analysed construction schedule risks in Ethiopia, whereas Buertey et al. (2013) discussed large construction projects in Ghana. Choudhry et al. (2014) established risk guidelines for bridge construction in Pakistan, whereas Do et al. (2023) identified claim causes that result in construction delays. Fashina et al. (2021) investigated significant factors in Hargeisa, whereas Gardezi et al. (2014) analysed factors affecting time extension in Pakistan. Hossen et al. (2015) evaluated construction

***Corresponding author: Yazan Alzubi,** Department of Civil Engineering, Faculty of Engineering Technology, Al-Balqa Applied University, Amman, Jordan. E-mail: yazan.alzubi@bau.edu.jo
Fuad Al-Bataineh and Ahmed Ali Khatatbeh, Department of Civil Engineering, Faculty of Engineering, Al-al Bayt University, Mafraq, Jordan

schedule risks for an international Net Primary Production (NPP) project, whereas Hung and Wang (2016) investigated delay risks in Vietnam. In Qatar, Jarkas and Haupt (2015) and Jarkas and Younes (2014) identified the most significant construction risk factors. Kesavan et al. (2015) analysed civil engineering project delays in Sri Lanka, while Mahamid (2016) discussed factors contributing to poor project performance in Saudi Arabia. Melaku Belay et al. (2021) analysed cost overruns and schedule delays in Ethiopia. Mpofu et al. (2017) analysed the underlying causes in the United Arab Emirates. Prateepasen and Aumpiem (2021) identified welding construction project delays in Thailand, and Rauzana and Dharma (2022) investigated the causes in Indonesia. Salem and Suleiman (2020) identified risk factors in the Jordanian construction industry, and Wuni et al. (2022) systematically reviewed risk factors in modular integrated construction. Lastly, Zafar et al. (2016) evaluated risk factors in terror-affected regions of Pakistan, and Zhao et al. (2022a) analysed risk factors associated with Prefabricated construction projects (PCP) delays. On the other hand, various efforts to adopt machine learning techniques in construction management have been made recently. For instance, Adedokun et al. (2023) utilised random forest and path diagram taxonomies to investigate construction project risks in higher education. Chen et al. (2022) utilised machine learning to investigate factors influencing the success of sustainable development projects. Egwim et al. (2021) applied Artificial intelligence (AI) to predict construction project delays, and Ivanovi et al. (2022) combined machine learning with expert knowledge to investigate the causes of construction project delays. Shoar et al. (2022) used machine learning to predict cost overruns in high-rise residential building projects. Uddin et al. (2022) presented a data-driven framework for project analytics using machine learning. Van and Quoc (2021) performed a thorough scientometric analysis of machine learning trends in construction management. Yu et al. (2019) utilised deep learning to predict flight delays in the aviation industry. Additionally, Zhao et al. (2022b) utilised machine learning to predict delays in prefabricated projects. These ground-breaking studies demonstrate the transformative potential of machine learning techniques to reshape conventional construction risk-management approaches. The primary objective of this study is to utilise the capabilities of unsupervised machine learning, particularly K-means clustering, to investigate the underlying causes of construction delays across multiple projects. The study's objectives are twofold: first, to identify and classify the critical risk factors contributing to these delays, and second, to employ a novel application of K-means clustering to group these risk factors, thereby enhancing understanding of the complex risk landscape. This innovative application of machine learning challenges traditional risk-management paradigms and paves the way for a more integrated approach that bridges the gap between empirical data and actionable insights. Accordingly, this research can improve risk-management strategies by providing stakeholders with a pragmatic tool to proactively address potential delays, ensuring the timely completion of projects in developing regions.

2 Research significance

The importance of minimising construction delays cannot be understated. Such delays increase the cost of projects and can hinder development objectives, negatively affecting the microeconomic and macroeconomic ecosystems. The significance of the research presented here stands out for numerous reasons. First, while many studies investigate the causes of construction delays, few make use of the unrealised potential of machine learning (Egwim et al. 2021; Shoar et al. 2022, Uddin et al. 2022). In this context, using unsupervised machine learning and K-means clustering, as is the case with this study, has been uncommon. This study offers a more accurate data-driven approach than traditional methods that heavily rely on human intuition and experience by employing such advanced analytical techniques. Second, the dataset utilised, which was collected from multiple contracting firms in developing regions, is a large and representative sample. Such depth guarantees that the drawn conclusions are statistically significant and universally applicable to a wider range of projects. In addition, the study's systematic categorisation of risk variables into five levels based on the Likert scale adds a level of granularity that most conventional studies lack. Such a hierarchical breakdown assists stakeholders in prioritising their resources and efforts by focusing on the most significant risk factors first. The application of K-means clustering increases the significance of the study. By categorising the identified risk factors, the research provides a clearer picture of the interrelationships and overlaps of various risks. Understanding these intricate relationships are essential for developing comprehensive risk-mitigation strategies that target the underlying causes as opposed to merely the symptoms. In addition to its academic contributions, this research has important implications for the real world also. The results offer stakeholders the possibility of anticipating potential risks.

This means that project managers and decision-makers will have greater foresight, more efficient planning and more reliable adherence to project deadlines.

3 Materials and methods

3.1 Research methodology

In the present study, the research methodology, depicted in Figure 1, starts with a literature review, in which previous studies are analysed to comprehend the historical context and existing patterns. This analysis extends to methodologies from prior research, highlighting their most salient characteristics and techniques. The phase concludes with the identification of gaps and the determination of the inquiry's focus areas. Thereafter, emphasis is placed on aggregating diverse project data and the subsequent data preprocessing that ensures data quality and machine learning suitability. The segment on identifying risk variables entails a thorough data analysis complemented by data visualisation to facilitate an in-depth evaluation of risk factors. The subsequent phase, the K-means clustering application, integrates the K-means application to identify optimal clusters and provide insights into inherent cluster relationships. The results and discussions section represents the study's

Step	Description	Sub-Step	Description
1	Literature review	Previous studies	Review of prior research studies
		Methodologies	Analysis of methodologies used in prior studies
		Gaps	Identification of gaps in existing literature
2	Data collection	Diverse project data	Collection of diverse project data
		Data preprocessing	Data cleaning and preparation
3	Risk variable identification	Data analysis	Examination of collected data
		Data visualisation	Graphical representation of data
		Risk factors	Identification of risk factors
4	Applications of K-means clustering	K-means applications	Application of K-means clustering technique
		Optimal clusters	Determination of optimal number of clusters
		Cluster relationships	Analysis of relationships within clusters
5	Results and discussions	Results consolidation	Compilation of results
		Recommendations	Formulation of recommendations
		Stakeholder sharing	Sharing findings with stakeholders

Fig. 1: General research methodology adopted in this study.

findings, resulting in informed recommendations and promoting stakeholder sharing for broader collaborative implications.

3.2 Database collection

This research's primary dataset was collected from Rauzana and Dharma (2022), which comprehensively explored risk factors causing construction delays in Indonesia. The selection of this database is justified by its detailed and extensive coverage of construction project delays, particularly within Aceh Province, where the rapid development and inherent project complexities heighten the relevance of understanding these delays. This study identified 60 risk factors, with 30 categorised as having a very high influence and 29 with a high influence, providing a nuanced understanding of the critical elements impacting project timelines. Furthermore, the methodological rigour demonstrated through validity and reliability tests, along with the construction of descriptive statistics, enhances the credibility and robustness of the data. By addressing a broad spectrum of risk factors, from financial and economic issues to project management and consultant delays, this dataset offers a comprehensive view crucial for developing accurate and effective K-means clustering models. The dataset's empirical nature, gathered from structured questionnaires administered to 68 contracting firms in Indonesia, ensures the analysis is deeply rooted in real-world scenarios, making it highly relevant and applicable to the research objectives.

3.3 Database reliability assessment

Cronbach's Alpha is a fundamental metric in psychometrics and scale development. Generally, it measures internal consistency reliability, gauging how closely related a set of items are as a group. Conceptually, if a scale possesses true internal consistency, each scale item should measure the same underlying construct, leading to a high Cronbach's Alpha value. The value of Cronbach's Alpha can range from 0 to 1. A Cronbach's Alpha closer to 1 indicates strong internal consistency, while values near 0 suggest weak consistency. In practice, a threshold of 0.7 is often cited as acceptable, although this benchmark can vary depending on the research context. While a high Cronbach's Alpha can suggest that the items are consistent, it does not necessarily denote unidimensionality. Additionally, Cronbach's Alpha is sensitive to the number of items in a scale; scales with more items can yield inflated Cronbach's Alpha values. Accordingly, a reliability analysis was conducted using Cronbach's Alpha on the dataset derived from the survey to ensure the internal consistency of the questionnaire items. The analysis yielded an overall Cronbach's Alpha of approximately $\alpha = 0.791$ for the entire questionnaire, suggesting good internal consistency among the items. To further explore the contribution of each item to the overall reliability, Cronbach's Alpha was computed for the questionnaire while excluding each item, one at a time. This approach provides insights into whether the reliability of the questionnaire might improve in the absence of any specific item. In general, Figure 2 indicates that the questionnaire exhibits an acceptable level of internal consistency and can be regarded as reliable.

3.4 K-means clustering

Clustering is a critical data segmentation and pattern recognition technique in machine learning and data science. The K-means clustering algorithm is one of the most popular clustering algorithms, recognised for its simplicity and efficiency. K-means aims to partition a set of data points into distinct groups, known as clusters so that data points in the same cluster are more similar than those in other clusters. This similarity is often defined in terms of distance metrics such as the Euclidean distance. The core operation of the K-means algorithm involves assigning data points to the nearest centroid, where each centroid represents the centre of a cluster. Upon the assignment, the centroid of each cluster is recalculated based on the mean of all points within that cluster. This two-step process is repeated iteratively until the centroids stabilise, signalling the algorithm's convergence. However, one challenge that often arises with K-means clustering is determining the optimal number of clusters (K) for a given dataset. Too few clusters can oversimplify the data, leading to a loss of information, while too many clusters can overcomplicate it, resulting in overfitting. On the other hand, the Elbow method is a heuristic approach used to find a dataset's optimal number of clusters. It involves running the K-means clustering on the dataset for a range of values of K, and then for each value of K, compute the sum of squared distances from each point to its assigned centre. As the number of clusters increases, the variance captured by each cluster will diminish, reducing the total sum of squared distances. By plotting the number of clusters against this sum, one typically observes an 'elbow' in the graph. This point, where the rate of decrease sharply changes, represents an optimal

value for K, which is a balance between precision and computational cost. Accordingly, K-means clustering is an important tool for data analysis, and the Elbow method helps determine the number of clusters that best fit the data. These techniques, when combined, pave the way for meaningful insights and data-driven decisions across various domains and applications. The general methodology used during the model development process is shown in Figure 3.

3.5 Principal component analysis

Principal component analysis (PCA) is a well-established statistical technique used primarily in multivariate data analysis for dimensionality reduction. In general, PCA aims to represent high-dimensional data in a lower-dimensional space, preserving as much variance as possible by identifying orthogonal axes, called principal components, which maximise the variance of the

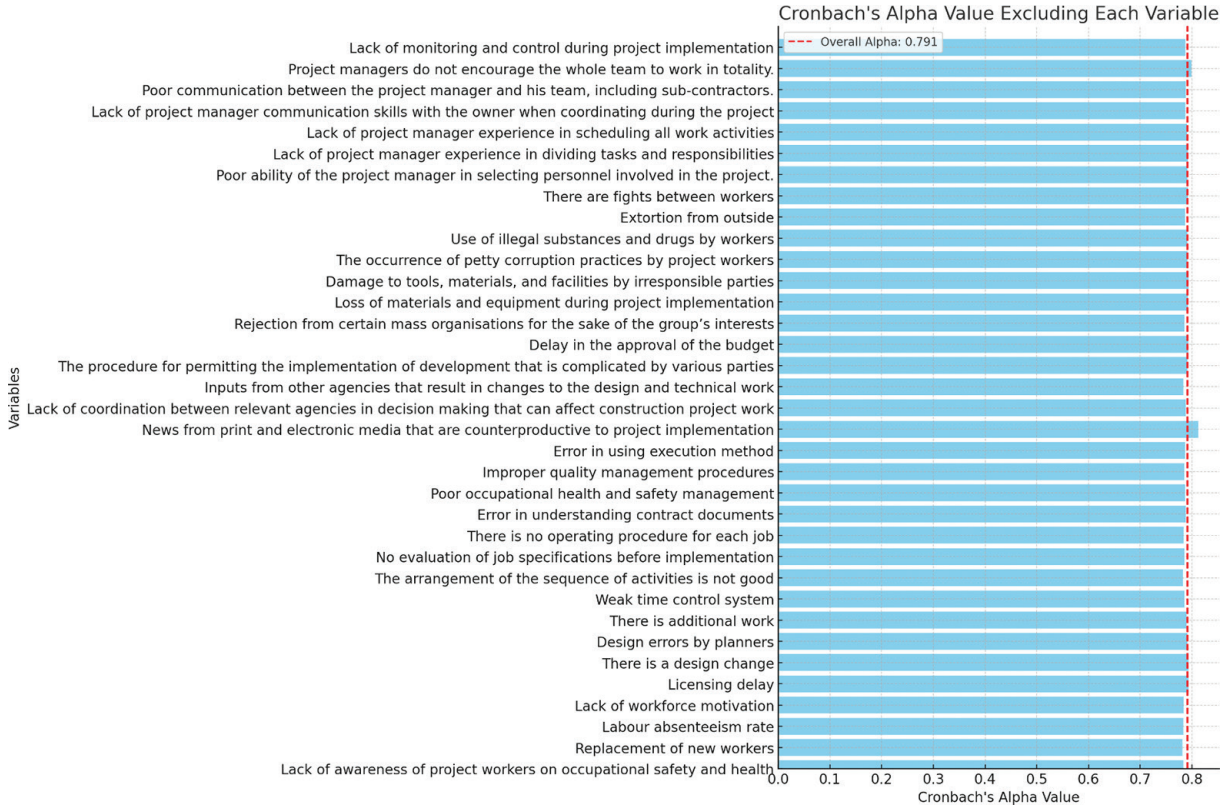

Fig. 2: Reliability assessment using Cronbach's Alpha.

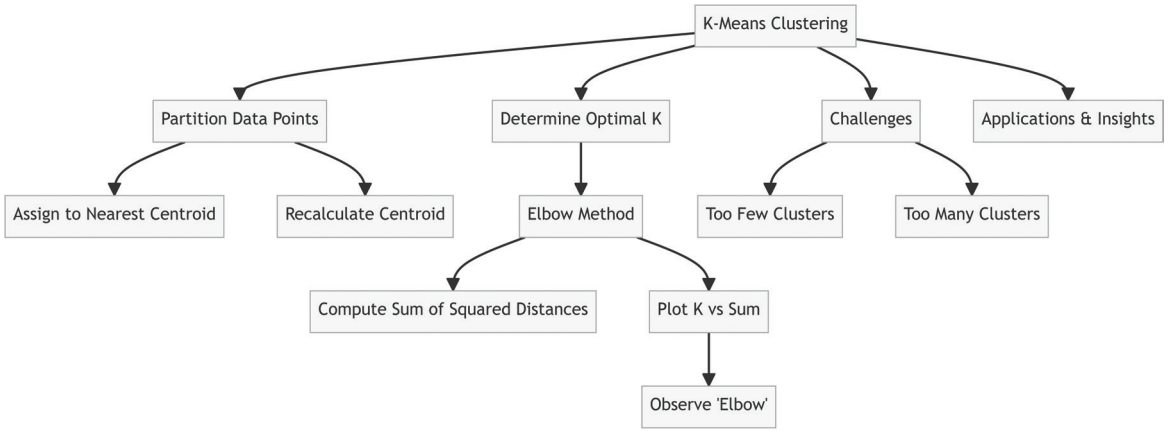

Fig. 3: General overview of the K-means clustering approach.

projected data. The first principal component accounts for the highest variance, while each subsequent component captures decreasing amounts of variance, subject to being orthogonal to the preceding components. By transforming data into this new basis set, PCA allows researchers to identify patterns, reduce noise and visualise complex datasets more interpretably. Additionally, PCA can be important in mitigating the curse of dimensionality and has found applications in various fields, ranging from biology to finance and speech processing to image recognition.

4 Key risk factors contributing to construction delays

The construction industry is an ecosystem with processes and stakeholders that must smoothly interact to achieve project success. Within this context, delays, whether minor or significant, can have high effects on project timelines, financial outcomes and stakeholder satisfaction. The survey data selected in this study were investigated using the traditional technique, and the results, given in Figure A1 in Appendix and Figure 4, are investigated as follows.

4.1 Financial aspects and budgetary concerns

The survey results show an average score of 4.81 for delays in budget approval, as shown in Figure 4. This is a reminder of the bureaucratic challenges many projects face. Approval processes, especially involving multiple checkpoints, can be slow. This is not in waiting for funds to be allocated; the entire construction process can stop during this period. Similarly, the delay in payment processes, 4.69, as depicted in Figure 4, underscores the financial interdependencies in construction projects. A single missed or delayed payment can ripple down the chain, affecting suppliers, labourers, and subcontractors and ultimately stalling the project.

4.2 Design and planning hurdles

The survey results reveal that errors in this phase are common, with a score of 4.63. Such missteps can have a domino effect, necessitating multiple modifications and revisions. Furthermore, changes introduced due to inputs from other agencies, 4.60, point to the challenges of integrating diverse visions into a cohesive plan.

4.3 Operational challenges and equipment reliability

On the operational front, two significant concerns emerge: equipment breakdowns and high operational costs, scoring 4.66. The former causes an immediate disruption, halting specific tasks and potentially derailing the project timeline. Conversely, the latter poses a more systemic challenge, often prompting project managers to revisit and potentially revise their strategies.

4.4 Influence of external entities

The influence of external stakeholders is undeniable. Rejections from mass organisations, with a score of 4.68, can be particularly daunting. Such entities can wield considerable power, and their opposition can introduce unexpected delays.

4.5 Labour and personnel concerns

The survey also explores aspects that, while not top concerns, still hold significant implications. Labour issues, for instance, emerge as a recurring theme. The human element remains pivotal, from labour shortages and fatigue to integrating new workers. Communication, or the lack thereof, also surfaces as a concern. Whether the project manager's interactions with the owner or within the team, the scores of 3.96 and 3.87 highlight the importance of clear, consistent communication.

4.6 External pressures and media influence

The least influential factor, as perceived by respondents, relates to external news and media reports. Scoring 3.44 suggests that while external perceptions matter, they do not significantly disrupt the on-ground progress of construction projects.

The survey's findings serve as a roadmap for stakeholders in the construction industry, pointing them towards areas that need attention. While financial and design issues top the list, it is evident that an approach encompassing operational efficiency, stakeholder management and effective communication is the key to mitigating delays. Thus, in construction, where unpredictability is often the norm, proactive planning and adaptive management can distinguish between a project's success and its descent into a quagmire of delays.

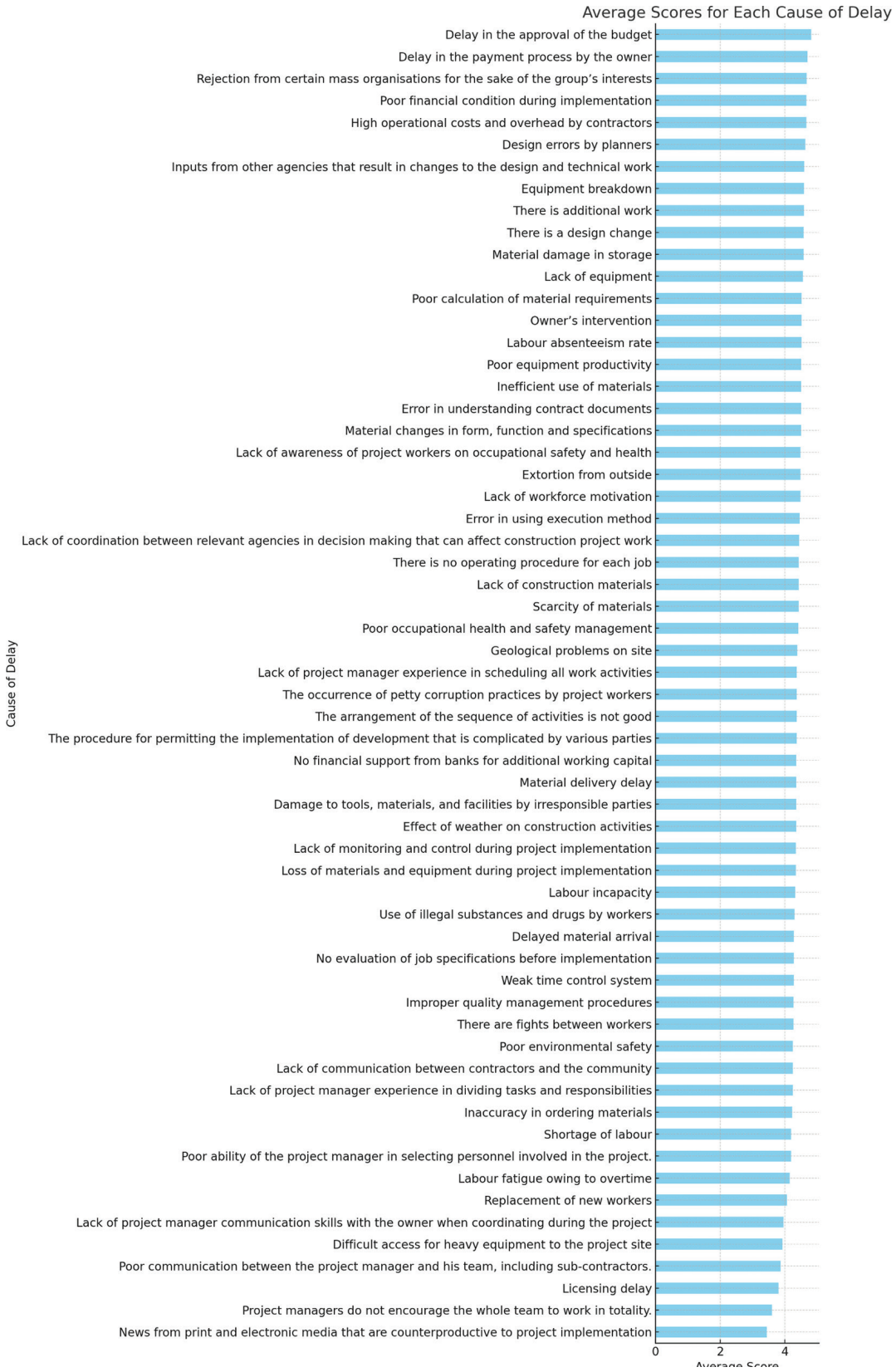

Fig. 4: Average scores of the effects of each risk variable on the delay of construction projects.

5 Unsupervised machine learning-based key risk factors assessment

The advent of unsupervised machine learning, especially K-means clustering, has improved the management of risk factors leading to construction delays in various projects. This study aims to enhance the understanding of these risk factors through data analysis.

5.1 Cluster formation and interpretation

Determining the optimal number of clusters was a pivotal step in the analysis. By using the elbow method, the optimal number of clusters, *k*, was identified as four. This optimal segmentation allowed for a robust categorisation of the risks associated with construction delays. The four clusters were found to represent diverse challenges inherent in construction projects, Figure 5 and Table 1. The first cluster, Cluster 0, includes operational and external challenges faced by projects. Risks in this cluster predominantly centred around external threats, such as damages from external influences or extortion. Concurrently, they also highlighted challenges within the operations, such as financial constraints during the implementation phase or high operational costs. In contrast, Cluster 1 was more centred on projects' financial and managerial aspects. Delays stemming from payment processes, budgetary approvals and specific managerial challenges were the defining characteristics of this cluster. It highlighted the criticality of sound financial processes and managerial efficiencies in meeting project timelines. Moving forward, Cluster 2 stood out with its labour-centric risks. From shortages of labour and related disputes to fatigue due to excessive overtime, this cluster underscored the pivotal role labour plays in project completion timelines. The challenges highlighted here indicate various projects' labour dynamics and management strategies. Lastly, Cluster 3 was the most comprehensive, encompassing a broad spectrum of risks. These ranged from challenges related to the availability and quality of construction materials, equipment breakdowns and specific environmental challenges to more intrinsic issues related to management inefficiencies.

The box plots in Figure 6 show the distribution of average risk perceptions across four distinct clusters. Cluster 1, with the highest median score nearing 4.5, consistently perceives most risk factors as highly influential, as evidenced by its narrow interquartile range (IQR). Contrastingly, Cluster 2, with the lowest median below 3, generally views risk factors as moderately influential, albeit with a slightly broader IQR hinting at greater internal variability. Notably, outliers in Cluster 2 suggest a subset of respondents with more heightened risk

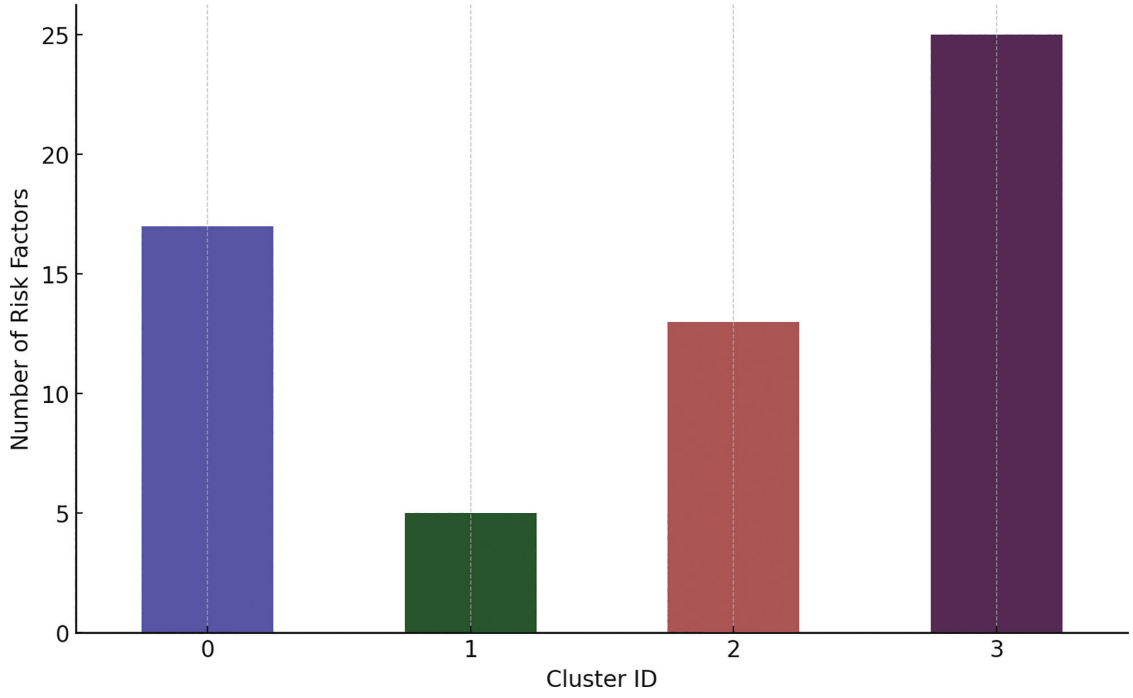

Fig. 5: Number of risk factors associated with each cluster.

Tab. 1. Risk factor categorisation using K-means clustering

Risk Factor Dominant	Cluster ID	Risk Factor Dominant	Cluster ID
Lack of construction materials	3	There is a design change	0
Material changes in form, function and specifications	3	Design errors by planners	0
Material delivery delay	2	There is additional work	0
Material damage in storage	3	Weak time control system	3
Scarcity of materials	3	The arrangement of the sequence of activities is not good	3
Inaccuracy in ordering materials	2	No evaluation of job specifications before implementation	3
Inefficient use of materials	3	There is no operating procedure for each job	3
Delayed material arrival	2	Error in understanding contract documents	3
Poor calculation of material requirements	3	Poor occupational health and safety management	3
Equipment breakdown	3	Improper quality management procedures	3
Lack of equipment	3	Error in using execution method	3
Poor equipment productivity	3	News from print and electronic media that are counterproductive to project implementation	1
Poor financial condition during implementation	0	Lack of coordination between relevant agencies in decision making that can affect construction project work	3
Delay in the payment process by the owner	3	Inputs from other agencies that result in changes to the design and technical work	0
No financial support from banks for additional working capital	1	The procedure for permitting the implementation of development that is complicated by various parties	3
High operational costs and overhead by contractors	0	Delay in the approval of the budget	1
Effect of weather on construction activities	3	Rejection from certain mass organisations for the sake of the group's interests	0
Poor environmental safety	3	Loss of materials and equipment during project implementation	0
Geological problems on site	3	Damage to tools, materials, and facilities by irresponsible parties	0
Lack of communication between contractors and the community	3	The occurrence of petty corruption practices by project workers	0
Difficult access for heavy equipment to the project site	3	Use of illegal substances and drugs by workers	0
Shortage of labour	2	Extortion from outside	0
Labour incapacity	2	There are fights between workers	0
Owner's intervention	2	Poor ability of the project manager in selecting personnel involved in the project.	0
Labour fatigue owing to overtime	2	Lack of project manager experience in dividing tasks and responsibilities	0
Lack of awareness of project workers on occupational safety and health	2	Lack of project manager experience in scheduling all work activities	0
Replacement of new workers	2	Lack of project manager communication skills with the owner when coordinating during the project	1
Labour absenteeism rate	2	Poor communication between the project manager and his team, including sub-contractors.	2
Lack of workforce motivation	2	Project managers do not encourage the whole team to work in totality.	1
Licensing delay	2	Lack of monitoring and control during project implementation	0

perceptions. Clusters 0 and 3, with medians around 3.5, position themselves between the clusters above in terms of perceived risk influence. However, Cluster 3's wider IQR indicates more diverse opinions within its cohort. This diversity underscores the importance of recognising that perceptions regarding construction delay risk factors are heterogeneous. Such insights are essential for stakeholders aiming to develop risk-management strategies that are both effective and resonate with various respondent groups. This analysis emphasises the need to consider the

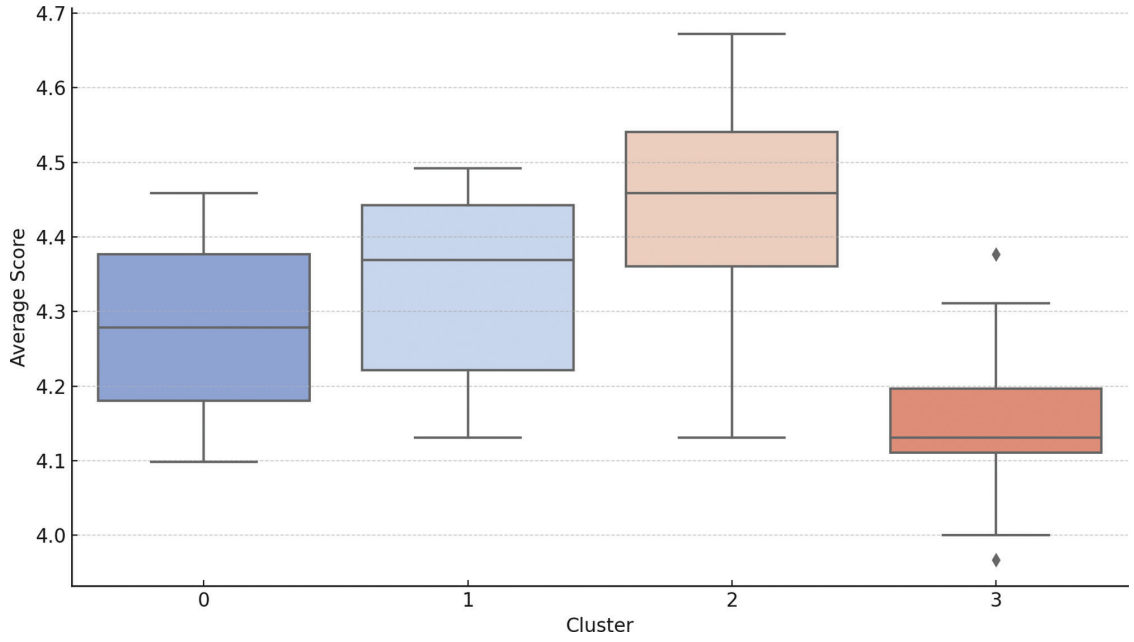

Fig. 6: Distribution of average scores in each cluster.

multifaceted nature of risk perceptions in the construction domain.

5.2 Risk factor categorisation

The visual representation of data, especially through heatmaps, provided a granular perspective on the average ratings for each risk factor, as segregated by clusters. The colour gradients in the heatmap offered intuitive insights, with darker shades indicating higher average ratings, thereby signalling heightened concerns for the associated risk factor. In Figure A2 in the Appendix, each cluster's average values for various risk factors are shown, providing an aggregate perspective on the perceived influence of these factors across different respondent groups. This visualisation approach offers a dual-layered understanding. Firstly, it shows the primary drivers that have led to the formation of each cluster, derived from the K-means clustering algorithm's inherent grouping based on similarities in responses. Secondly, it highlights the broader perception landscape by averaging the scores of all risk factors for respondents within a given cluster. It is pivotal to comprehend that even if a specific risk factor is not the predominant reason for a respondent's allocation to a cluster, it still influences the average value represented in the heatmap. Each respondent, having scored every risk factor, contributes to these averages, thus ensuring that the heatmap does not exclusively highlight the most influential factors. Instead, it offers a view of each cluster's average perceptions across the entire risk spectrum. This comprehensive representation empowers stakeholders with a comprehensive understanding, bridging individual perceptions and collective trends, thereby facilitating more informed decision-making in construction delay risk management.

In general, Cluster 0, focusing on operational challenges, presented a clear picture of projects often beleaguered by external threats and operational inefficiencies. On the other hand, Cluster 1, with its emphasis on financial and managerial aspects, brought forth the intricacies of financial constraints and managerial challenges that often lead to project delays. The labour-centric nature of Cluster 2 was visually evident, highlighting the centrality of labour management in project outcomes. Lastly, Cluster 3, with its broad spectrum of risks, emphasises various challenges projects face, ranging from resource constraints to managerial inefficiencies. A deeper exploration of the data led to the categorisation of each risk factor based on its dominant cluster, as shown in Table 1. This categorisation provided a structured understanding of how different clusters perceive specific risk factors.

For instance, risk factors related to material scarcity, equipment breakdowns and certain environmental challenges were predominantly aligned with Cluster 3. In contrast, labour-related risks, such as labour shortages and fatigue due to overtime, were primarily associated with Cluster 2. Financial constraints during project

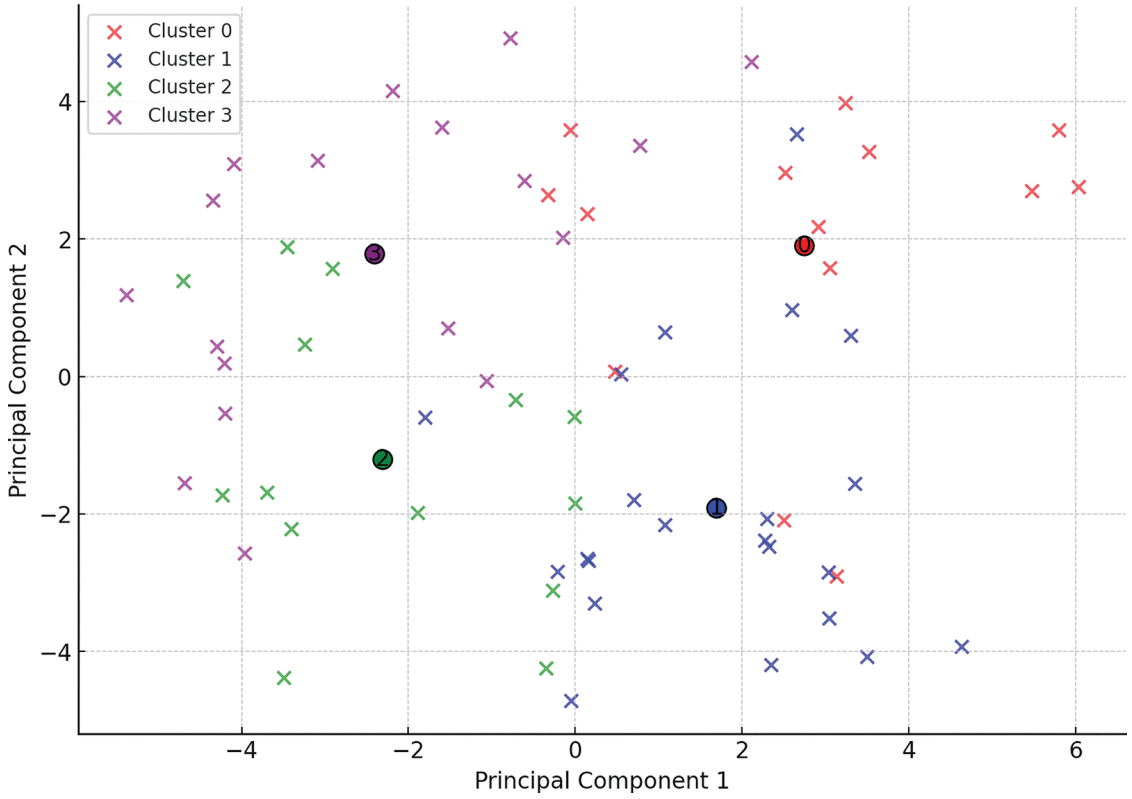

Fig.7: PCA analysis based on the K-means clustering results. PCA, principal component analysis.

implementation and high operational costs were dominantly categorised under Cluster 0, while delays in the payment process and managerial challenges found their predominant association with Cluster 1.

5.3 Insights using principal component analysis

PCA is a dimensionality-reduction technique that transforms a set of possibly correlated variables into a new set of uncorrelated variables known as principal components. Indeed, in construction delays and risk factors, the application of PCA is especially pertinent due to the potentially large number of variables involved, each representing a distinct risk factor. This study applied the PCA to the dataset to reduce its dimensionality to two principal components, allowing for a comprehensive 2D visualisation. The primary motivation behind this reduction was to better visualise and interpret the clustering results obtained from the prior K-means clustering analysis. In general, Figure 7 offers a comprehensive visual representation of the clusters in a two-dimensional space, achieved through applying PCA for dimensionality reduction. Notably, the distinctness between Clusters 2 and 3 as opposed to Clusters 0 and 1 is evident, pointing to divergences in risk perceptions among the respondents of these clusters. Clusters 0 and 1 also exhibit certain overlaps, indicating analogous risk perceptions among their respondents, even though they remain predominantly separable. By going deeper, the overlapping regions of Clusters 0 and 1, despite their inherent unique attributes, suggest that respondents across these clusters consistently perceive a specific subset of risk factors. Addressing these universally recognised risk factors becomes paramount for stakeholders. In contrast, Cluster 2 stands out, representing respondents who overwhelmingly identify labour-associated risks as primary contributors to construction delays, underscoring the necessity for specialised labour-management strategies for projects aligned with this cluster. Cluster 3, another distinct category, poses some concerns, spanning from resource-related impediments to managerial inefficacies, warranting a comprehensive, multifaceted risk-mitigation strategy. Recognising and decoding these clusters empowers stakeholders to craft risk-management strategies tailored to specific clusters. For instance, projects associated with Cluster 2 could immensely benefit from a heightened focus on labour welfare and efficient management, potentially accelerating project completion timelines.

5.4 Implications and discussions

The distinct clusters that emerged from the analysis underscore the multifaceted nature of construction delays. Each cluster, with its unique set of predominant risks, offers a lens through which the intricacies of project delays can be understood. This clustering approach aligns with previous studies that have highlighted the diverse nature of risk factors in construction projects across different regions and project types. Projects associated with Cluster 0, characterised by operational and external challenges, often grapple with threats from within and outside their operational boundaries. This finding resonates with the work of Mpofu et al. (2017) and Al Zubaidi and Al Otaibi (2008), who identified external influences such as unforeseen political events and security threats as critical delays. Internally, financial constraints during the project's implementation or operational inefficiencies can be significant roadblocks, as noted by Abd Karim et al. (2012) and Al-Momani (2000). Addressing these challenges requires a two-pronged approach. Externally, enhancing security measures and ensuring risk assessments account for potential external threats become crucial. Internally, securing financial contingencies and streamlining operations can be pivotal in ensuring projects stay on track (Ali et al. 2010; Hossen et al. 2015). Cluster 1, focusing on financial and managerial challenges, highlights the importance of robust financial processes and managerial efficiencies. Delays from payment processes or budgetary approvals can often have a cascading effect on project timelines (Ahmed et al. 2002; Adedokun et al. 2023). Similarly, managerial inefficiencies or lack of requisite skills can exacerbate these delays. Strengthening financial processes, ensuring timely budgetary approvals and investing in managerial training become essential strategies for projects associated with this cluster (Cheng and Darsa 2021; Alajmi and Ahmed Memon 2022). This cluster's findings are consistent with previous research that underscores the critical role of financial management in preventing delays (Aibinu and Odeyinka 2006; Akomah and Jackson 2016). The challenges highlighted in Cluster 2 provide a window into the labour dynamics. The centrality of labour in construction projects is undeniable, and the risks associated with labour shortages, disputes or fatigue due to overtime can significantly impact project timelines (Gardezi et al. 2014; Fashina et al. 2021). Emphasising worker welfare, fostering a conducive work environment and ensuring efficient labour management can be effective strategies to mitigate these risks (Jarkas and Younes 2014; Do et al. 2023). This cluster aligns with the findings of Kesavan et al. (2015) and Mahamid (2016), who highlighted labour issues as a significant factor in construction delays. Lastly, Cluster 3, with its diverse set of risks, underscores the need for a holistic approach to risk management. The strategies to address these risks must be multifaceted, from ensuring the availability of quality construction materials and timely equipment maintenance to focusing on environmental safety and managerial efficiencies (Choudhry et al. 2014; Zhao et al. 2022b). This comprehensive approach to risk management is supported by the work of Salem and Suleiman (2020) and Wuni et al. (2022), who emphasised the importance of integrated risk-management strategies in the construction industry. The integration of earlier research into the discussion emphasises the importance of contextualising new findings within the established body of knowledge. By doing so, this study not only reinforces the validity of the current findings but also sets a foundation for future research. The clustering of risk factors using K-means provides a novel perspective that can significantly enhance risk-management practices by offering a more nuanced understanding of the risk landscape in construction projects.

6 Conclusion

In this study, the application of unsupervised machine learning, particularly K-means clustering, has pioneered a fresh perspective on understanding the complex and multifaceted risk factors contributing to construction delays. By integrating a dataset from diverse contracting firms with advanced analytical methods, the research provided a granular and structured understanding of the intricate risk landscape, challenging traditional paradigms and setting the stage for a more integrated risk-management approach. Key findings of this study include:

- Construction delays are shaped by intertwined factors, ranging from operational to managerial challenges.
- When effectively applied, the K-means clustering can delineate these factors into clear and distinct clusters, offering unique lenses for understanding and mitigating the associated risks.
- Labor dynamics, highlighted predominantly in Cluster 2, underscore the essentiality of prioritising worker welfare and efficient labour management.
- Clusters 0 and 1 emphasised the importance of addressing external threats and internal inefficiencies, especially in finance and management.
- Cluster 3's broad spectrum of risks underscored the need for comprehensive risk-management strategies that are adaptive and multifaceted.

Finally, the insights from this study offer a new approach to risk management in construction projects. By enabling stakeholders to be proactive rather than reactive and fostering data-driven decision-making, the potential for timely project completion in these regions looks promisingly high. The future beckons for a world where construction delays are significantly minimised, and this research has laid a robust foundation for realising that vision.

References

Abd Karim, N. A., Rahman, I. A., Memmon, A. H., Jamil, N., & Azis, A. A. A. (2012, December). Significant risk factors in construction projects: Contractor's perception. In *2012 IEEE Colloquium on Humanities, Science and Engineering (CHUSER)*. IEEE, pp. 347-350.

Adedokun, O., Egbelakin, T., & Omotayo, T. (2023). Random forest and path diagram taxonomies of risks influencing higher education construction projects. *International Journal of Construction Management*, *24*(1), pp. 1-9.

Ahmed, S. M., Azhar, S., Castillo, M., & Kappagantula, P. (2002). *Construction Delays in Florida: An Empirical Study*. Final report. Department of Community Affairs, Florida, US.

Aibinu, A. A., & Odeyinka, H. A. (2006). Construction delays and their causative factors in Nigeria. *Journal of Construction Engineering and Management, 132*(7), pp. 667-677.

Akomah, B. B., & Jackson, E. N. (2016). Contractors perception of factors contributing to road project delay. *International Journal of Construction Engineering and Management, 5*(3), pp. 79-85.

Al Zubaidi, H., & Al Otaibi, S. (2008). An empirical approach for identifying critical time-ovrerun risk factors in Kuwait's construction projects. *Journal of Economic and Administrative Sciences, 24*(2), pp. 35-53.

Alajmi, A. M., & Ahmed Memon, Z. (2022). A review on significant factors causing delays in Saudi Arabia construction projects. *Smart Cities, 5*(4), pp. 1465-1487.

Ali, A. S., Smith, A., Pitt, M., & Choon, C. H. (2010). Contractors' perception of factors contributing to project delay: Case studies of commercial projects in Klang valley, Malaysia. *Journal of Design and Built Environment, 7*(1), pp. 1-17.

Al-Momani, A. H. (2000). Construction delay: A quantitative analysis. *International Journal of Project Management, 18*(1), pp. 51-59.

Alshihri, S., Al-Gahtani, K., & Almohsen, A. (2022). Risk factors that lead to time and cost overruns of building projects in Saudi Arabia. *Buildings, 12*(7), p. 902.

Buertey, J. I. T., Miezah, A. K., & Adjei-Kumi, T. (2013, August). Delays to large construction projects in Ghana: A risk overview. In *Proceedings: 5th West Africa Built Environment Research (WABER) Conference*. Accra, Ghana: WABER Conference, (pp. 367-380).

Chen, Z. J., Hsieh, T. S., & Mousavi Davoudi, S. M. (2022). Analysis of factors affecting the success of sustainable development projects with the help of machine learning tools. *Discrete Dynamics in Nature and Society, 2022*, pp. 1-9.

Cheng, M. Y., & Darsa, M. H. (2021). Construction schedule risk assessment and management strategy for foreign general contractors working in the Ethiopian construction industry. *Sustainability, 13*(14), p. 7830.

Choudhry, R. M., Aslam, M. A., Hinze, J. W., & Arain, F. M. (2014). Cost and schedule risk analysis of bridge construction in Pakistan: Establishing risk guidelines. *Journal of Construction Engineering and Management, 140*(7), p. 04014020.

Do, S. T., Nguyen, V. T., Tran, C. N., & Aung, Z. M. (2023). Identifying and evaluating the key claim causes leading to construction delays. *International Journal of Construction Management, 23*(12), pp. 1999-2011.

Egwim, C. N., Alaka, H., Toriola-Coker, L. O., Balogun, H., & Sunmola, F. (2021). Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications, 6*, p. 100166.

Fashina, A. A., Omar, M. A., Sheikh, A. A., & Fakunle, F. F. (2021). Exploring the significant factors that influence delays in construction projects in Hargeisa. *Heliyon, 7*(4), pp. 1-9.

Gardezi, S. S. S., Manarvi, I. A., & Gardezi, S. J. S. (2014). Time extension factors in construction industry of Pakistan. *Procedia Engineering, 77*, pp. 196-204.

Hossen, M. M., Kang, S., & Kim, J. (2015). Construction schedule delay risk assessment by using combined AHP-RII methodology for an international NPP project. *Nuclear Engineering and Technology, 47*(3), pp. 362-379.

Hung, M. S., & Wang, J. (2016). Research on delay risks of EPC hydropower construction projects in Vietnam. *Journal of Power and Energy Engineering, 4*(4), pp. 9-16.

Ivanović, M. Z., Nedeljković, Đ., Stojadinović, Z., Marinković, D., Ivanišević, N., & Simić, N. (2022). Detection and in-depth analysis of causes of delay in construction projects: Synergy between machine learning and expert knowledge. *Sustainability, 14*(22), p. 14927.

Jarkas, A. M., & Haupt, T. C. (2015). Major construction risk factors considered by general contractors in Qatar. *Journal of Engineering, Design and Technology, 13*(1), pp. 165-194.

Jarkas, A. M., & Younes, J. H. (2014). Principle factors contributing to construction delays in the State of Qatar. *International Journal of Construction Project Management, 6*(1), p. 39.

Kassem, M., Khoiry, M. A., & Hamzah, N. (2020). Using probability impact matrix (PIM) in analyzing risk factors affecting the success of oil and gas construction projects in Yemen. *International Journal of Energy Sector Management, 14*(3), pp. 527-546.

Kesavan, M., Gobidan, N. N., & Dissanayake, P. B. G. (2015). Analysis of factors contributing civil engineering project delays in Sri Lanka. In *6th International Conference on Structural Engineering and Construction Management 2015*, pp. 40-46.

Mahamid, I. (2016). Factors contributing to poor performance in construction projects: Studies of Saudi Arabia. *Australian Journal of Multi-Disciplinary Engineering, 12*(1), pp. 27-38.

Melaku Belay, S., Tilahun, S., Yehualaw, M., Matos, J., Sousa, H., & Workneh, E. T. (2021). Analysis of cost overrun and schedule delays of infrastructure projects in low income economies: Case studies in Ethiopia. *Advances in Civil Engineering, 2021*, pp. 1-15.

Mpofu, B., Ochieng, E. G., Moobela, C., & Pretorius, A. (2017). Profiling causative factors leading to construction project

delays in the United Arab Emirates. *Engineering, Construction and Architectural Management, 24*(2), pp. 346-376.

Prateepasen, A., & Aumpiem, A. (2021). Problems causing delays and risk factors in welding construction projects of Thailand. *Engineering Journal, 25*(5), pp. 33-44.

Rauzana, A., & Dharma, W. (2022). Causes of delays in construction projects in the Province of Aceh, Indonesia. *PLoS One, 17*(1), e0263337.

Salem, Z. T. A., & Suleiman, A. (2020). Risk factors causing time delay in the Jordanian construction sector. *International Journal of Engineering Research and Technology, 13*(2), pp. 307-315.

Shoar, S., Chileshe, N., & Edwards, J. D. (2022). Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: Application of random forest regression. *Journal of Building Engineering, 50*, p. 104102.

Uddin, S., Ong, S., & Lu, H. (2022). Machine learning in project analytics: A data-driven framework and case study. *Scientific Reports, 12*(1), pp. 15252.

Van, T. N., & Quoc, T. N. (2021). Research trends on machine learning in construction management: A scientometric analysis. *Journal of Applied Science and Technology Trends, 2*(03), pp. 96-104.

Wuni, I. Y., Shen, G. Q., & Mahmud, A. T. (2022). Critical risk factors in the application of modular integrated construction: A systematic review. *International Journal of Construction Management, 22*(2), pp. 133-147.

Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review, 125*, pp. 203-221.

Zafar, I., Yousaf, T., & Ahmed, D. S. (2016). Evaluation of risk factors causing cost overrun in road projects in terrorism affected areas Pakistan–a case study. *KSCE Journal of Civil Engineering, 20*, pp. 1613-1620.

Zhao, Y., Chen, W., Arashpour, M., Yang, Z., Shao, C., & Li, C. (2022b). Predicting delays in prefabricated projects: SD-BP neural network to define effects of risk disruption. *Engineering, Construction and Architectural Management, 29*(4), pp. 1753-1776.

Zhao, Y., Chen, W., Yang, Z., Li, Z., & Wang, Y. (2022a). Analysis on risk factors related delay in PCPs. *Engineering, Construction and Architectural Management*, *30*(10), pp. 4609-4644.

Appendix

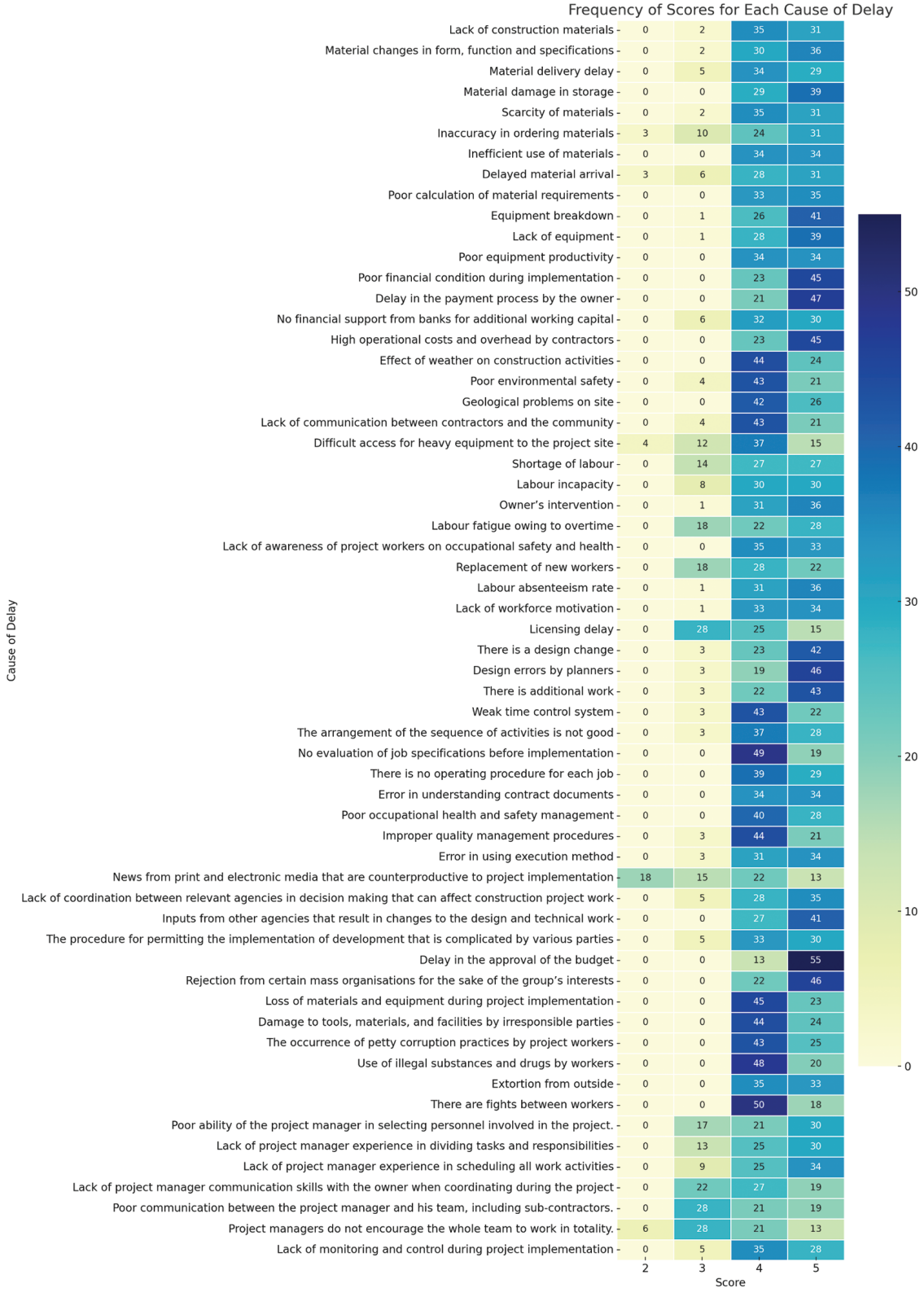

Fig. A1: A statistical representation of the risk factors on the construction delays of projects.

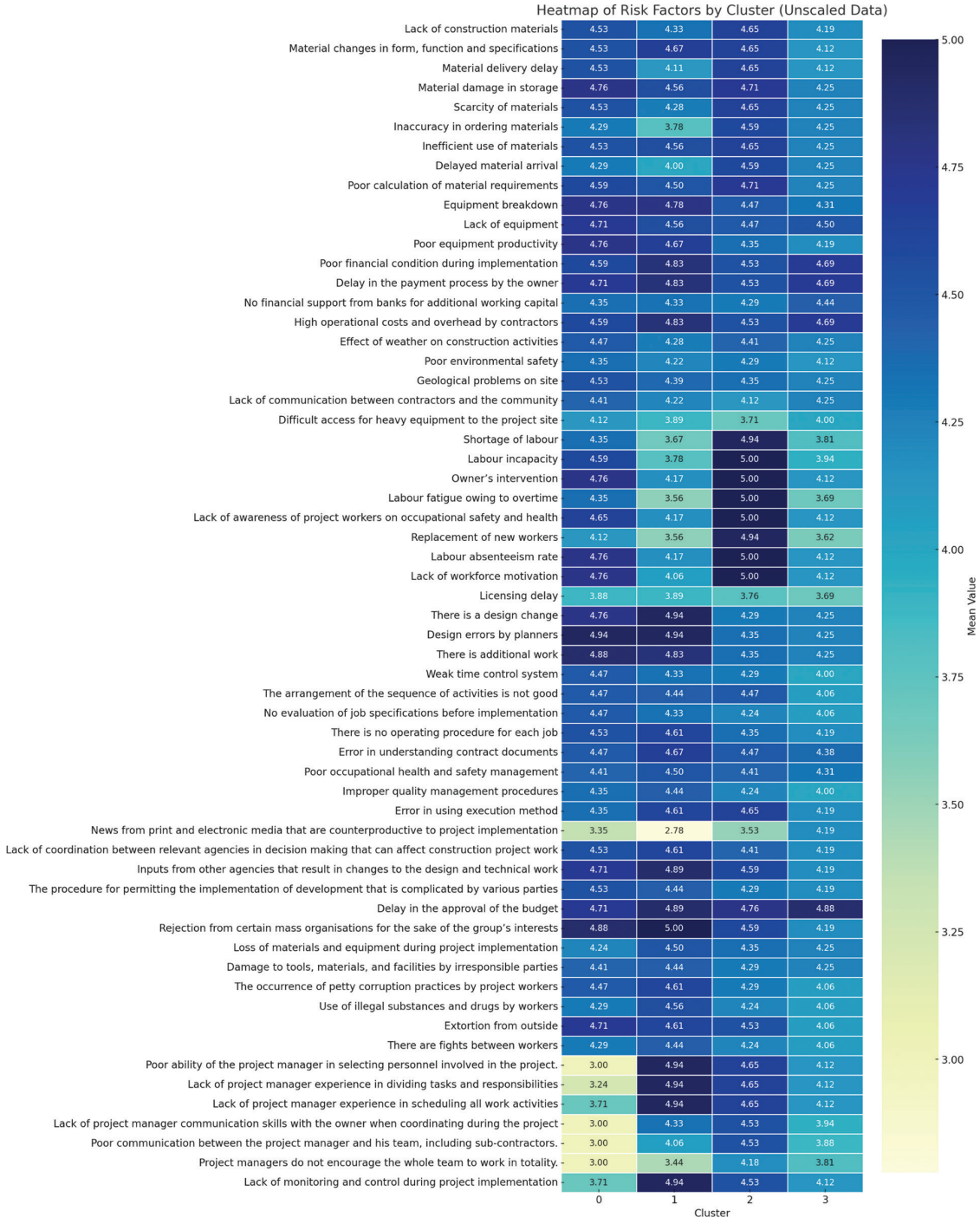

Fig. A2: Heatmap of the interaction between risk factors and cluster.