Received 13 November 2024
Revised 25 February 2025
Accepted 4 March 2025

The C-BA method: enhancing megaproject forecasting through the "Fifth Hand" principle

Costanza Mariani, Francesco Cellerino and Erik Humberto Araya Aliaga
Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milan, Italy
Edison Atencio
Department of Civil Engineering, Pontificia Universidad Catolica de Valparaiso, Valparaiso, Chile, and
Mauro Mancini
Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milan, Italy

Abstract

Purpose – Current theoretical viewpoints regarding the performance trends of megaprojects endorse the notion that incorporating an outside perspective during the forecasting phase can be advantageous for the overall progress of the megaproject. This paper aims to propose a novel approach, the clustering-behavior analysis (C-BA), that leverages unsupervised machine learning to integrate an outside perspective as support for megaproject forecasting.

Design/methodology/approach – Employing a database of 90 megaprojects, we demonstrated the application of C-BA. By utilizing unsupervised machine learning techniques, the method uncovers unforeseen patterns among past megaprojects, clusters them based on these patterns and allows for conducting a performance comparison with current megaprojects.

Findings – The findings reveal that the proposed C-BA method offers an effective alternative for supporting megaproject forecasting, aligning with the Fifth Hand principle. For practitioners, this would facilitate efficient benchmarking and has the potential to serve as a learning system within megaproject organizations.

Originality/value – The originality of this work lies in introducing a novel method that integrates an outside perspective into megaproject, with forecasts based on unsupervised machine learning. This approach aligns with the Fifth Hand principle and highlights the potential of artificial intelligence to serve as a learning system, offering a new avenue for efficient benchmarking in megaproject management. The paper adds complex network theory by giving the possibility of analyzing the uniqueness and unpredictable nature of megaprojects.

Keywords Forecasting, Methodologies, Megaprojects

Paper type Research paper

1. Introduction

Megaprojects, characterized by their large scale and complexity, often find themselves mired in challenges that lead to cost and time overrun (Flyvbjerg *et al.*, 2009; Flyvbjerg, 2014). While historical definitions of megaprojects have consistently emphasized their economic impact, typically exceeding 1 billion US dollars (Flyvbjerg, 2014), recent discussions have increasingly highlighted the scale of their complexity (Van Marrewijk *et al.*, 2008; Brookes and Locatelli, 2015; He *et al.*, 2021; Vukomanović *et al.*, 2021). Their magnitude and

International Journal of Managing Projects in Business
Vol. 18 No. 8, 2025
pp. 50-78
Emerald Publishing Limited
e-ISSN: 1753-8386
p-ISSN: 1753-8378
DOI 10.1108/IJMPB-11-2024-0281

entangled nature dramatically reduce the accuracy of the estimations of cost and time, causing unanticipated delays and overruns. The extant literature has long debated the causes of these overruns, attributing them to two principles: the Planning Fallacy and the Hiding Hand. Authors supporting the first principle (*such as* Ansar *et al*., 2014; Flyvbjerg, 2014) argue that individuals making estimates for a bid under conditions of uncertainty are prone to underestimate the cost and time required to complete a task, while simultaneously overestimating the associated benefits (Kahneman and Tversky, 1979). Planners and project managers frequently adopt an excessively optimistic and biased perspective, as also noted by Roy *et al*. (2008), leading to the underestimation of project's costs and duration. These estimation errors persist even when planners acknowledge that similar projects have historically deviated from predictions.

On the other hand, proponents of the Hiding Hand principle - such as Ika (2018) and Ika and Söderlund (2016) propose that inaccuracies in estimates during the forecasting stage may not always be detrimental to a project. They contend that, in numerous instances, initial forecasting underestimates the challenges, difficulties, and costs associated with a megaproject. However, they argue that this lack of awareness of the true nature of the situation encourages companies to optimistically embark on projects they might otherwise avoid if fully informed. Therefore, adopting an external perspective and acquiring additional insights from past projects may not necessarily contribute to a project's success and might even impede its initiation.

A substantial amount of academic research has critiqued both Planning Fallacy and Hiding Hand (Love *et al*., 2019; Themsen, 2019; Delise *et al*., 2023; Pinto, 2023), aiming to determine which phenomenon is more commonly observed in real-world projects and which more effectively explains the actual behavior of project managers and planners. Evidence suggests that both the Planning Fallacy and the Hiding Hand are present in real-world projects and that no single perspective completely dominates project-related behavior (Sassano, 2025).

The concept of the Fifth Hand, introduced by Ika *et al*. (2022) inspired by the work of Anheier's (2016), builds on the existing theories of the Planning Fallacy and the Hiding Hand by incorporating an external perspective that enhances the understanding and management of megaprojects. This perspective encourages viewing project preconditions more openly and comprehensively, facilitating both research and practical applications in project management. Thus, this concept asserts that such a perspective enhances the capacity to learn from past megaproject experiences by incorporating external insights from past endeavors.

The literature addressing the incorporation of an external perspective into megaproject forecasting predominantly revolves around (1) the storage of lessons learned in megaproject management (Davies and Brady, 2000; Brady and Davies, 2004), (2) the methods of integrating an outside view into megaproject forecasting (Batselier and Vanhoucke, 2016; Flyvbjerg *et al*., 2016; Fridgeirsson, 2016), and (3) the underlying reasons behind success or failure in assimilating insights from past megaproject experiences (Andranovich *et al*., 2001; Han *et al*., 2009).

In this paper, our focus is on the literature concerning point (2), which predominantly proposes the use of Reference Class Forecasting (RCF) to integrate an outside view in megaproject planning (Flyvbjerg, 2008; Flyvbjerg, 2016). The literature acknowledges its limitations, although it is recognized as a valid method for direct forecasting. The final values produced by this method need adjustment for optimism bias, which typically involves an arbitrary uplift determined by the decision maker (Pinto, 2023).

Recent developments in the field of Artificial Intelligence (AI) have expanded its application across various domains of project forecasting, including enhancing estimations of project timelines (Wang *et al*., 2012), optimizing cost predictions (Cheng *et al*., 2010; Dursun and Stoy, 2016), and predicting potential project risks (Mariani and Mancini, 2023; Mancini *et al*., 2023). In particular, machine learning techniques have proved to be effective in analyzing complex data patterns and trends, thus offering more precise and dynamic forecasting models (Wei and Rana, 2019; Gondia *et al*., 2020; Mariani *et al*., 2023).For this

reason, we propose the idea that these advances can be a valuable support for megaproject forecasting, and we put forward the following research question:

> How can project decision-makers use machine learning to compare current project performance with past outcomes to improve forecasting?

To answer this question, we introduce Clustering-Behavior Analysis (C-BA) as a method that leverages unsupervised machine learning for data-driven insights to support megaproject forecasting. This novel method can effectively collect insights from past megaproject experiences and offer robust support to decision-makers operating under uncertainty (Gentleman and Carey, 2008; Hahne and Gentlemann, 2008; Alikhani and Jeong, 2021).

The theoretical contribution of this work is to suggest an additional method (C-BA) aligned with the implementation of the Fifth Hand principle proposed by Ika *et al*. (2022). The C-BA employs an external data-driven perspective that allows for the comparison of a current focal project's performance with the behaviors and outcomes of past endeavors. For practitioners, we propose that unsupervised machine learning can serve as an alternative method to support megaproject performance forecasting by efficiently benchmarking it against past behaviors. This method can serve as a learning system within megaproject organizations, leveraging past data to inform future learning, and contributing to a more informed and data-driven approach to megaproject management.

The paper is structured as follows: the background section introduces the key constructs of the paper, namely the debate between the Hiding Hand and Planning Fallacy principles, the presentation of the most well-known forecasting methods, and the logic underlying unsupervised clustering algorithms. Subsequently, we introduce and test C-BA using a dataset of 90 megaprojects, analyzed at the initial stage, at 50% progress and upon completion. The results of this dynamic analysis, which examines how project evolution influences changes in group membership, are detailed in the results section. The paper concludes by discussing the implications for both theory and practice, summarizing the findings, and outlining the main limitations.

2. Background

2.1 The Planning Fallacy debate

In literature, the tendency of projects and megaprojects to be late and to go over budget has been studied from different points of view, taking into account a plethora of dimensions of interest (Flyvbjerg, 2014; Denicol *et al*., 2020; Love *et al*., 2022a, b, 2023a, b, c, d). Key dimensions such as initial expectations, level of ignorance, creativity, and approach to risk play pivotal roles in the debates emerging from the literature (Denicol *et al*., 2020). In the last ten years, the discussion around the rationale underpinning unreliable initial estimates, culminated in a debate between two principles, namely the Hiding Hand, and the Planning Fallacy (Lepenies, 2018; Kreiner, 2020; Delise *et al*., 2023; McLeod, 2023).

The Planning Fallacy principle posits that project planners and managers often forecast project costs, timelines, and benefits with undue optimism, aligning their estimates predominantly with best-case scenarios (Flyvbjerg, 2014). This tendency is particularly pronounced in projects where knowledge uncertainty is low but complexity is unexpectedly high (Ika *et al*., 2022). In such environments—typically well-regulated industries—there is abundant information available from past projects. However, this wealth of data can lead to misinterpretations by project sponsors, who may selectively focus on favorable outcomes, overlook potential setbacks, and overemphasize the project's prospective benefits, thus skewing the planning process toward overly optimistic forecasts (Flyvbjerg *et al*., 2002). Conversely, the "Hiding Hand" principle suggests that, since actions and effects are projected into a future that cannot be properly predicted by forecasting, projects of all types are inevitably affected by a high level of uncertainty (Kreiner, 2020). This is especially relevant in contexts where little external knowledge is accessible, making it difficult to fully grasp or

accurately represent the complex nature of megaprojects. Consequently, planners rely on creativity and problem-solving to tackle the challenges that may emerge, possibly leading to benefit overruns alongside cost overruns. Each perspective – the Hiding Hand and Planning Fallacy - present advantages and disadvantages. The former can incentivize active problem-solving when facing uncertainties, where creative solutions may emerge and lead to more benefits than expected. In the case of the latter, the knowledge of the planners is usually sufficient to achieve the desired outcome despite the issues generated by the misrepresentation (Ika, 2018; Love *et al.*, 2023a, b, c, d). However, this might result in behavior characterized by profiteering and corruption. In both instances, lack of knowledge plays a pivotal role. In particular, the Hiding Hand enables the realization of many endeavors that would probably not even begin if the adversities were known antecedently (Kreiner, 2020). This is one of the main flaws of the approach: in cases of extreme uncertainty, the level of risk could be too high to be handled, causing the failure of the project (Eren, 2019).

To overcome this dualistic view, Ika *et al.* (2022) provide a more balanced theoretical explanation of project behavior, the so-called Fifth Hand. This principle highlights the importance of adopting an inclusive perspective when assessing project preconditions. Such an approach is instrumental in assisting both researchers and practitioners as they seek to gain deeper insights into the factors influencing project management performance and the achievement of project objectives (Love *et al.*, 2012a, b; Ika and Pinto, 2022a, b). Ika *et al.* (2022) specifically highlight that this viewpoint improves the capacity to derive valuable lessons by incorporating external insights from previous projects. This perspective has the potential to be relevant about a wide array of projects carried out in diverse circumstances, allowing for taking into account both optimism and pessimism bias (Ika and Pinto, 2022a, b; Love *et al.*, 2022a, b).

2.2 RCF: the main method supporting the outside view in forecasting

The Fifth Hand principle aims to facilitate learning from past experiences, providing a comprehensive understanding of the behaviors associated with previous megaprojects, applicable not only within the same organization's project domains but also across various industries (Alikhani and Jeong, 2021). Current literature indicates that the principal methodology for extracting insights from external data in megaproject forecasting involves the application of RCF (Flyvbjerg, 2008; Batselier and Vanhoucke, 2016). This technique entails identifying a reference class of similar past projects and constructing a probability distribution for the forecasted parameter within the selected reference class. It then compares a current focal project with the reference class distribution to establish its most likely outcomes (Flyvbjerg, 2016). There are numerous papers that have adopted RCF, applying this technique across various fields. Table 1 lists some of them, highlighting the method used, the content of the paper, the industry of application, and the limitations:

The technique has been primarily employed to assess the likelihood of cost and schedule overruns in extensive projects such as those found in transportation (Fridgeirsson, 2016), infrastructure (Suh and Ryerson, 2019; Steininger *et al.*, 2021), construction sector (Bayram and Al-Jibouri, 2016; Zani and Adey, 2025), oil and gas (Natarajan, 2022) and hydropower sectors (Ansar *et al.*, 2014; Awojobi and Jenkins, 2015; Callegari *et al.*, 2018; Awodi *et al.*, 2021). While often regarded as one of the most effective methods for incorporating external data into estimations, using RCF alone remains a relevant point of discussion, as literature has highlighted several structural limitations (highlighted in Table 1) that restrict its effectiveness in practice (Rajabi Asadabadi and Zwikael, 2024).

Furthermore, drawing on Gigerenzer (2013), Ika *et al.* (2022) challenge the applicability of statistical thinking in megaprojects, emphasizing that well-defined risks do not encapsulate the unpredictable nature of such projects, which are often fraught with unknown factors. The discussion further argues that while exact methods can mitigate known risks, they leave unaddressed uncertainties in project planning. Ika *et al.* (2022) suggest that when uncertainty

Table 1. Limitations of reference class forecasting

Reference	Method	Content	Industry	Limitations
Zani *et al.* (2024)	RCF	Proposing an alternative method for class selection before RCF application	Infrastructure	Lack of large samples of similar projects; difficulty in gathering enough data with the accurate information necessary to form the reference classes; the selection of reference projects is a biased process; the method becomes the worst-performing contingency estimating method when the reference class is not specific enough
Zani and Adey (2025)	RCF + alternative stratified approach	Employs RCF for Swiss Highway project cost forecasting	Construction industry	RCF generates a subjective single uplift value rather than providing a spectrum of uplifts that reflect varying degrees of certainty
Themsen (2019)	RCF	Case study on the application of RCF in an infrastructure project	Infrastructure	The application of RCF did not prevent the experts from applying their own biased judgment when selecting the reference class of projects
Salling and Leleur (2015)	RSF	Propose the use of RSF – Integration of RCF and Quantitative Risk Analysis	Transport	RCF initial input is often wrong and biased. Try to solve this by integrating Monte Carlo Simulation and Risk Analysis
Bayram and Al-Jibouri (2016)	RCF	Application RCF to construction projects cost estimate in Turkey	Construction industry	RCF relies on a single uplift value. This approach does not account for varying levels of risk that different projects might entail. The paper proposes an improvement by integrating a range of uplift values that correspond to different risk levels
Leleur *et al.* (2015)	SIMRISK (RCF + OT + EJ)	Apply reference class forecasting (RCF) in association with risk simulation tools	Infrastructure	RCF can be effectively applied but must be used in a flexible way with the other tools to cope with possible mistakes in the sample selection

(*continued*)

Table 1. Continued

Reference	Method	Content	Industry	Limitations
Kaiser and Snyder (2012)	RCF + Regression model	Offshore wind capital cost estimation	Energy infrastructure	Inconsistent reporting standards and varying detail levels in data sources can introduce biases in the forecasting; fluctuating exchange rates and specific inflation rates introduce; conversion errors and biases; Variabilities in project conditions necessitate normalization, limiting the effectiveness of cost comparisons
Lovallo *et al.* (2012)	RCF and Similarity Based Forecasting (SBF)	Examine model of analogy and using empirical test compare it with RCF	Strategic management	Limits related to subjective expected utility (selection bias and anchoring effect)

Source(s): Authors' own creation

cannot be reduced to risk, a heuristic approach, which deliberately disregards certain information to make decisions faster, more economically and with comparable accuracy to a more intricate approach, can be a valuable tool for coping with uncertainty. In this paper, in line with the Fifth Hand, we outline that framing exact methods for estimation and heuristics as a binary choice, an "either/or" approach, does not necessarily enhance the precision of project planning. C-BA stands at the crossroads between the two perspectives by relying on past megaproject data for generating insights without providing an exact forecasting output.

2.3 Machine learning based methods for incorporating external data in forecasting: K-means as an innovative alternative

Traditional forecasting methods, rely heavily on the expertise and intuition of experts or groups of experts to make predictions. Thus, these approaches can lead to biased decisions and overly optimistic results due to their heavy dependence on the subjective perceptions of experts (Litsiou *et al.*, 2022). RCF aims to mitigate these biases by using data from similar past projects to inform predictions. Despite this, it also has limitations, as it still requires careful selection and interpretation of the reference class, which can introduce subjectivity into the process (Baerenbold, 2023).

To address these shortcomings, artificial intelligence has been introduced as a data-driven alternative to support decision-making in complex project scenarios. Table 2 summarizes a selection of papers that have employed AI techniques for forecasting project timelines and costs.

As it can be noted, a widely studied area is the use of feedforward Artificial Neural Networks (ANNs), also known as multilayer perceptrons (MLPs), to predict project metrics such as time, cost, or effort. ANN have the ability to identify key data features and patterns, and they have been shown to outperform traditional supervised linear regression models (López-Martín and Abran, 2015; Hsu *et al.*, 2021).

However, the success of these methods is highly dependent on the availability and quality of the data (Pospieszny *et al.*, 2018), making it challenging to develop a model that performs well in a given scenario. Further, if the ANN training data primarily consists of projects that don't capture the wide variability or the unique constraints and opportunities of new projects,

Table 2. AI for forecasting

Type of AI	Use/type of forecasting	Industry	Reference	Limitation
Neural networks + supervector machines	Predict project performances (cost and schedule)	Construction industry	Wang *et al*. (2012)	ANNs and SVMs act as "black boxes," their decision-making processes are not transparent; SVMs are effective in classification tasks but they can struggle with generalizability when applied to new projects that differ from those in the training set
Neural networks + Support vector machines	Improve Cost and Duration Prediction Accuracy	Construction industry	Darko *et al.* (2023)	The introduction of Deep Neural Networks (DNN) and Support Vector Regression (SVR) introduces complexity in terms of model configuration, training, and optimization
Neural networks	Risk prediction in tunnel construction frastructure	Luo *et al.* (2024)	Feature selection does not consider interrelationships between variables	
Neural networks	Predict construction cost of large sport field facilities	Construction industry	Juszczyk *et al*. (2019)	Do not possible to update on different time frame the database, limiting analysis effectiveness
Neural networks	Predict waste generation rate of building demolitions	Construction industry	Cha *et al.* (2023)	ANNs are sensitive to the input data variations and might not perform well if the data is not representative of the typical scenarios encountered during demolition projects; ANN requires accurately labeled data for training

(*continued*)

Table 2. Continued

Type of AI	Use/type of forecasting	Industry	Reference	Limitation
Neural networks	Predict project success	Construction industry	Ko and Cheng (2007)	The model used Fuzzy Lofic, Neural Networks and Genetic Algorithm. It requires extensive computational resources and time for training and optimizing hyper-parameters through Bayesian inference and Particle Swarm Optimization
Neural networks	Cost and time forecasting of megaprojects	Megaprojects and Infrastructure	Natarajan (2022)	The method cannot quantify all the projects risks and uncertainties; Outliers megaprojects cannot be predicted
Long short-term memory neural networks (LSTM) + ARIMA and ARIFMA	Predict the Volatility of Highway Construction Cost Index	Megaprojects and Infrastructure	Cao and Ashuri (2020)	If the change is in the testing period, ARIMA and ARIFMA can only detect the periodic ones, and cannot catch unhappened ones. If the change is in the training period, the time series model is insensitive to it when change happens near the end of the training sample or distant to the end

Source(s): Authors' own creation

the algorithm might not perform well. It could either overfit to the similarities of the projects in the training set or fail to recognize crucial instances that differentiate one project from another, leading to predictions that don't accurately reflect real-world scenarios (Darko *et al.*, 2023).

Furthermore, ANNs are often criticized for their lack of transparency, as they function as "black boxes", because they learn by adjusting internal weights based on input examples, rendering it nearly impossible to interpret the internal processes that lead to their predictions (Berlin *et al.*, 2009; López-Martín and Abran, 2015). Given these limitations, our paper proposes the implementation of an innovative use of unsupervised machine learning to support complex project forecasting.

Unsupervised clustering is a Machine Learning technique that operates without the need for predefined labels or categories, making it particularly suitable for discovering patterns that might not be immediately evident through manual analysis (Madhulatha, 2012). It has been extensively utilized in fields such as marketing (Volkmar *et al.*, 2022), stakeholder

classification (Pérez Vera, 2018; Mariani *et al*., 2023), and demand forecasting (Huber *et al*., 2017; Seyedan and Mafakheri, 2020).

Among the possible clustering techniques, K-means stands out as the most extensively employed one due to its simplicity and ease of implementation. In addition, its results are highly interpretable, making it a preferred choice for initial exploratory data analysis (Jain, 2010). Additionally, K-means has a relatively low computational cost, which is particularly advantageous when dealing with large datasets (Arthur and Vassilvitskii, 2007; Jain, 2010). The method's efficiency in terms of partitioning data into clusters with minimal computational resources, enhances its suitability for extensive applications in megaproject outcome classification. Additionally, in comparison to other methods such as hierarchical clustering, K-means is better suited for empirical investigations as it is capable of providing meaningful and stable results, making it a reliable method for clustering tasks across various datasets and applications (Madhulatha, 2012).

K-means represent clusters based on their respective centroids, computed as the mean of the objects assigned to the cluster. The sensitivity of centroids can be mitigated with preprocessing and initialization techniques such as Silhouette or Elbow analysis (Singh *et al*., 2011; Celebi, 2015; Ben Salem, Naouali and Chtourou, 2018). The partitioning algorithm then measures the Euclidean distance between the object and the cluster mean, finally ensuring that the resulting clusters are as compact and separate as possible (Soni Madhulatha, 2012; Charu and Chandan, 2013).

In conclusion, this algorithm has the capability to uncover latent recurrent patterns among historical megaprojects, spanning various fields and time periods, allowing the decision-maker to have at their disposal a foundation of information regarding past projects' behavior. This can then serve as an informed overview for future decisions based on similarities among past megaprojects (Invernizzi *et al*., 2018).

3. The new method explained

To demonstrate the application of unsupervised machine learning in megaproject forecasting we conceptualize a new method, named C-BA. The C-BA is an adaptive data-driven decision-making method, which relies on the concept that while a project by itself could be similar to past projects, it is still unique (Ika, 2018; Ika *et al*., 2022). Therefore, it is imprecise to predict the statistical numerical outcome of a project. However, it is possible to study how previous similar projects behaved by analyzing similarities in their past trends and outcomes. The proposed methodology leverages historical data from past megaprojects, organizing it into various stages of progress to observe trends and evolution over their lifetimes. This process enables to perform a cluster analysis at pre-defined progress stages, allowing the identification of distinct patterns in the data regarding megaprojects. This analysis can be performed utilizing customized variables based on the specific requirements of the analysis. After performing the multiple-stage clusterization, the methodology enables to obtain a detailed exploration of the historical behavior of megaprojects, by observing their transitions across clusters during their progression. Finally, this clustering framework can be applied to a new focal megaproject. By comparing its progress against the established cluster behaviors, insights can be obtained about potential outcomes based on historical patterns. These insights provide project managers with information, enabling them to steer the project in alignment with expected behaviors derived from past megaprojects' behavioral trends.

4. C-BA implementation

In this section, we illustrate an application of the C-BA method. As can be seen from Figure 1 the workflow consists of the following four stages: (1) assessment of state-of-the-art megaproject datasets and relative availability; (2) dataset design intended as the deductive

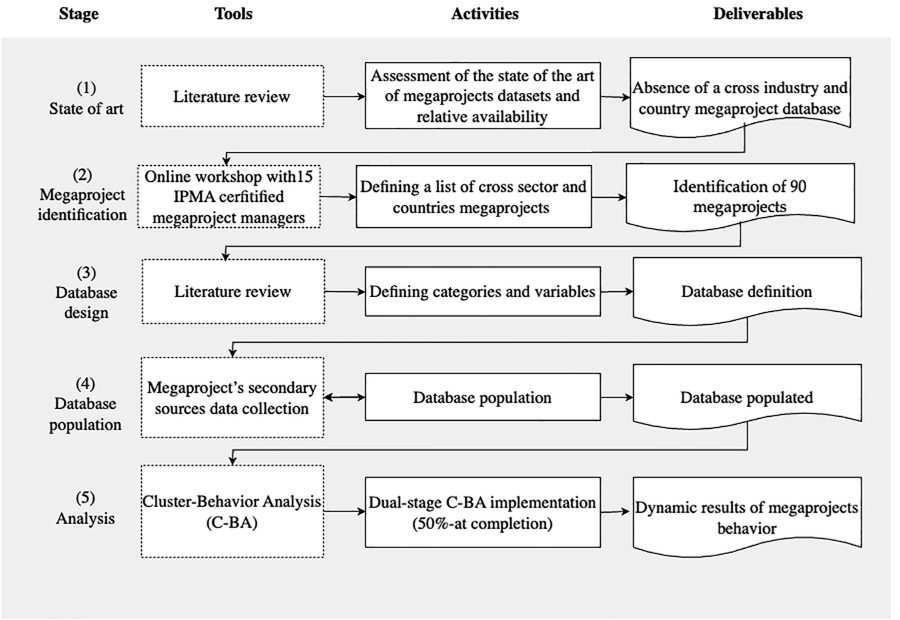

Figure 1. Research steps. Source: Authors' own creation

definition of the variables describing megaprojects; (3) population of the database, data adjustment, and normalization; (4) C-BA method implementation and results evaluation.

4.1 Database design

Figure 1 shows that the first step of the research *(Step 1)* was an extensive review on megaproject datasets, which revealed the absence of any cross-sectoral and international databases for megaprojects. Specifically, we found that (1) the majority of papers referencing megaproject databases do not offer public access to the data, thereby limiting the possibility of conducting further analysis; (2) public databases provided by governments and non-governmental organizations often pertain to specific countries or sectors (for example the Building America list of Energy projects by the U.S. Department of Energy; the Federal and State Department of Transportation project database; the Asian Infrastructure Investment Bank database and others). They provide lists segmented by country and industry, which impedes a comprehensive analysis of megaproject trends. As previously mentioned, this approach falls short because it confines the evaluation to a single reference class rather than exploring broader, cross-sectoral patterns; (3) the existing databases contain heterogeneous variables and lack structured data, making it challenging to perform a comprehensive analysis of megaprojects past data.

To overcome the limitations stated of public datasets, a megaproject dataset was developed to validate and test the Cluster-Behavior Analysis (C-BA) method. The dataset was generated in two stages: (1) Dataset design and (2) Dataset Population.

Dataset design task was focus on the identification of megaprojects from diverse sectors and countries, and for this, an initial list was developed in collaboration with fifteen IPMA certified managers with vast experience and diverse backgrounds *(Step 2)* (Isaac, 2023). Each manager proposed a list of influential megaprojects from the last 20 years according to their background and the definition given to them.

The 20-year threshold was established for mainly two reasons: (1) megaprojects are highly complex, resource-intensive, and lengthy, resulting in a low annual frequency; a shorter time span would yield an insufficient quantity for clustering. (2) Due to the scarcity of public project data, limiting the selection to projects from the past 20 years ensures they were built in the post-internet era, which facilitates data availability.

After gathering individual lists, they were merged into a single comprehensive database. Post processing was applied to standardize the name and remove duplicated records. The depurated list was then filtered by sectors and validated in online meetings with the managers with higher experience in said sectors, assuring the projects meet the requirements to be considered megaprojects according to its characteristics and impact.

The objective of this database is to compile sufficient information for the effective application of k-means clustering, thereby providing decision-makers with clusters of projects that exhibit similar behaviors based on selected parameters. This approach enables historical analysis and offers opportunities to examine how decisions made in previous projects within similar clusters influenced performance outcomes. For this, the aim was not to create a statistically representative database —unnecessary for the effective application of k-means clustering, but rather to assemble a multi-sectorial catalog of megaprojects where data integrity could be assured (Jain, 2010). The selection process was guided by principles of industry sector diversity to ensure a broad representation of megaproject characteristics, thereby overcoming the limitations associated with random cross-sectional sampling (employed in RCF).

The dataset resulted in 117 projects of sectors ranging from infrastructure projects (such as roads, bridges, water security systems, tunnels, and dams), to extractive industries (focusing on oil and minerals), research and development projects (covering areas like software design, biotechnology, and aerospace innovation), and consumption-related projects (including travel and tourism, film festivals, Olympic stadiums, and entertainment complexes). In terms of project stages, we restricted our selection to megaprojects that had already been completed to ensure consistency and comparability in our analysis.

The effectiveness of our unsupervised clustering analysis depended significantly on the availability of complete and detailed data. Thus, after a first selection of the dataset, a secondary filtering process was performed to verify data coverage for each of them in order to minimize data gaps. This resulted in a final dataset of 90 megaprojects with available data, that meet the requirement of a budget higher than US$1bn, a value that is in line with some of the best-known definitions of megaprojects (Love *et al.*, 2022a, b, 2023a, b, c, d).

4.2 Database population

Once the selection of the megaprojects was completed, we deductively identified the megaprojects' variables based on a literature analysis *(Step 3)* Appendix. These variables were selected with the specific aim of capturing relevant characteristics of megaprojects within the study domain, ensuring that they would yield meaningful results during the subsequent C-BA. It is important to note that the selected variables are not restrictive; the method is fully dynamic and allows for modifications to the input variables to adapt to the specific needs and objectives of the study.

Once the variables were identified we populated the database by manually consulting several reputable websites, scientific blogs and government websites *(Step 4)*. These included Science Direct, Scopus, Google Scholar, Web of Science, *The Guardian*, *The New York Times*, *The Seattle Times*, the NASA official website and many others. Once completed the data collection, a post-processing process was implemented to standardize the variables, thus allowing the comparison of the items and generate significant clusters (Figure 2). This was performed in two phases: (1) Context and (2) Temporal corrections. Context corrections (1) involved the transformation of values into a single currency, in this case all values were standardize by converting them into US dollars (USD) with the conversion rate corresponding

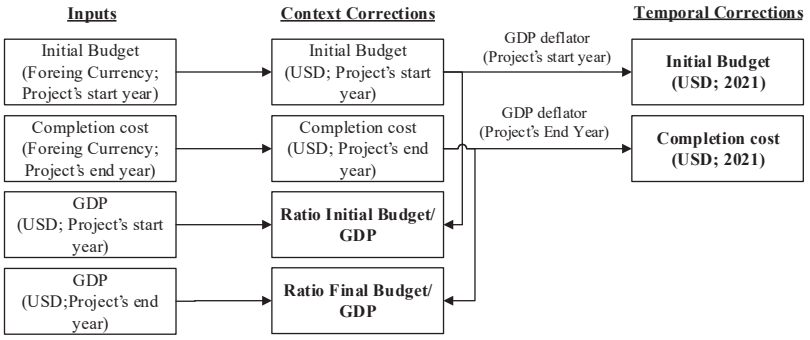

Figure 2. Pre-processing of data. Source: Authors' own creation

to the year of which the data was extracted. Two ratio variables were introduced, calculated as the initial and final budgets divided by the Gross Domestic Product (GDP) of the megaproject's country of origin. This allowed us to represent the absolute impact of the megaproject on the country's GDP. Temporal corrections (2) consisted in adjustments for transforming the cost variables into relative values that could account for both inflation and variations in GDP across different countries. Converting the relative value of a monetary amount from one year to the equivalent value in another year can be accomplished using the GDP deflator (Barro, 2013).Since GDP alone does not consider the influence of inflation or changes in price levels, the GDP deflator addresses this issue by comparing current prices to those in an established base year. The computation of the initial megaproject's budget and the relative final cost was calculated as the proportion between the relative cost at year 2021 and the GDP deflator at year 2021. In doing so, we ensured that the values of megaprojects completed in different countries could be effectively compared, also taking inflation into account.

The final list of variables included in the C-BA procedure are marked with an (*) in Annex 1. Once the dataset was processed, we proceeded to apply the K-means unsupervised clustering algorithm on Python. To prove the C-BA method *(Step 5)*, we performed the analysis considering megaprojects at three progress stages: at their beginning (Initial Clusters), at their half-life (Partial Clusters) and at completion (Full Clusters).

4.3 Definition and analysis of clusters

In this section, we describe the methodology used to define and analyze clusters, using the "at completion" database as an illustrative example. For clarity, this demonstration employs data from megaprojects that have been completed. In the subsequent section, we apply the C-BA to the same megaprojects at a 50% completion stage. We employed Principal Component Analysis (PCA), a statistical technique used to simplify the complexity of high-dimensional data by transforming it into fewer dimensions. This method reduces the dataset to its most significant features, which are called principal components. To determine the number of components we utilized eigenvalues, where each eigenvalue represents the amount of variance captured by its corresponding principal component. The plot in Figure 3 shows the eigenvalues of the components, with a horizontal line indicating the cutoff value of 1.0 for retaining them. In our analysis, four components have eigenvalues above this threshold, suggesting that these four capture the most significant variance in the dataset and should be considered for further analysis.

Table 3 illustrates the rotated components resulting from the PCA. The first component represents the scale of the project relative to the country's GDP, highlighting its economic significance to the nation. The second component focuses on temporal variables, capturing time-related aspects of the project. The third component pertains to the cost variable,

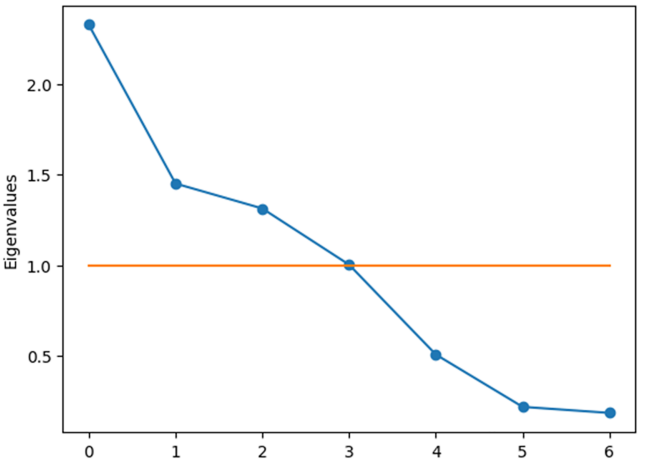

Figure 3. Eigenvalues for PCA analysis. Source: Authors' own creation

Table 3. Rotated components of PCA analysis

Index	Comp1	Comp2	Comp3	Comp4
Relative value 2021 of initial budget [billion $]	−0.05	0.01	0.71	−0.07
Initial budget/GDP	0.70	0.02	−0.01	−0.04
Relative value 2021 of the final cost [billion $]	0.06	−0.01	0.69	0.09
Final cost/GDP	0.70	−0.02	0.01	0.03
Overbudget percentage	−0.00	0.00	0.00	0.99
Years of delay	0.03	0.72	−0.10	0.05
Duration	−0.03	0.69	0.11	−0.05

Source(s): Authors' own creation

specifically the project's budget. Lastly, the fourth component emphasizes variables related to budget overruns, identifying financial exceedances.

The K-means algorithm necessitates the number of clusters *(k)* as an input. To determine *k*, we employed three methods: Silhouette Analysis, the Elbow Method, and Hierarchical Clustering (Figure 4a and b).

Silhouette Analysis evaluates the quality of clustering by determining the silhouette coefficient for each data point (Rousseeuw, 1987). This coefficient quantifies how closely a data point aligns with its cluster relative to its proximity to other clusters. The Elbow Method, on the other hand, plots the total within-cluster sum of squares against the number of clusters (Antunes *et al*., 2018; Mouton *et al*., 2020). The point where the rate of decrease shifts is considered as the optimal number of clusters. The Silhouette coefficient and elbow method values are reported in Table 4.

The optimal number of clusters is identified as five as indicated by the Elbow Method. This method demonstrates a significant leveling off in the rate of decrease in the sum of squared distances at this point. Further supporting this selection, the Silhouette coefficient for this cluster count is notably high at 0.51. This suggests a good balance between intra-cluster cohesion and inter-cluster separation, affirming that five clusters are the optimal choice.

This choice is further confirmed by the dendrogram plotted in Figure 5, which is a visual representation analysis that shows the hierarchical structure of clusters by illustrating how data points or groups of data points merge as the number of clusters decreases.

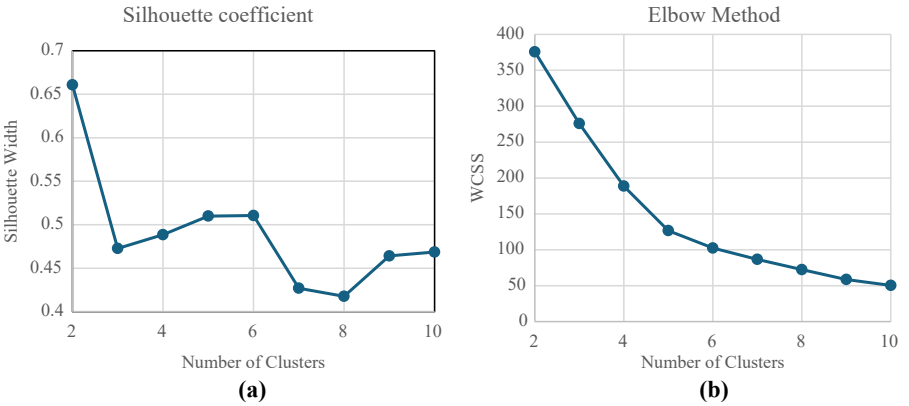

Figure 4. (a) Silhouette coefficient plot; (b) Elbow method plot. Source: Authors' own creation

Table 4. Silhouette coefficient and elbow method values

Clusters	Elbow method	Silhouette coefficient
2	375.73	0.66
3	275.84	0.47
4	188.69	0.49
5	126.67	0.51
6	102.44	0.51
7	86.63	0.43
8	72.35	0.42
9	58.60	0.46
10	50.40	0.47

Source(s): Authors' own creation

This three-step analysis, used to identify PCA components and the optimal number of clusters, can be consistently applied across various stages of the megaprojects.

5. Results

5.1 C-BA at completion

This section provides a comprehensive overview of the results of the application of the C-BA method. The results section will start by providing the application of the C-BA to the megaproject at completion (Final Clusters), proceeding then with the analysis of the trends when examining projects' shifts. Table 5 reports the means of the clusters for the four components discussed in the previous section (influence of the budget over the country's GDP, megaproject duration, cost, and overbudget).

Cluster Final 1 (F1) primarily consists of infrastructure projects characterized by minimal delays, short durations, and relatively low budget overruns. *Cluster F2* includes megaprojects that show very high initial budgets; however, these projects do not show cost overruns, experience delays, or significantly impact the GDP of the country. This suggests that these projects might have been overestimated initially and ended up being less impactful than expected. *Cluster F3* comprises megaprojects that have a substantial impact on the country's GDP, indicating their significant influence, though they are not the most expensive in the dataset and have average durations. *Cluster F4* encompasses projects with long durations

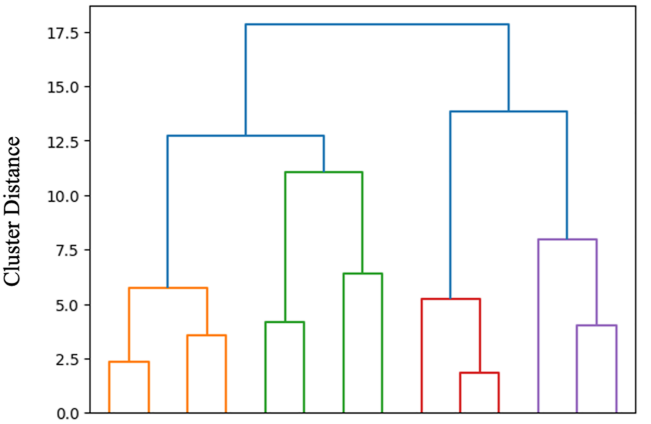

Figure 5. Hierarchical clustering dendrogram. Source: Authors' own creation

Table 5. Clusters obtained performing the cluster analysis on the complete database

Cluster label	Influence	Duration	Cost	Overbudget
F1	−0.25	−0.65	−0.35	−0.17
F2	0.36	0.51	5.23	−0.47
F3	6.58	0.30	1.27	−0.18
F4	−0.25	1.72	−0.18	−0.22
F5	−0.30	0.31	−0.20	4.13

Source(s): Authors' own creation

showing low overbudget figures and relatively modest total budgets. Finally, *Cluster F5* is marked by projects that show significant cost overruns, with average durations, indicating challenges in terms of budget management and planning within these projects.

5.2 C-BA at half-life megaproject

Table 6 shows the results of the K-means cluster analysis performed at fifty percent of completion of the megaprojects. This analysis was performed by generating a synthetic dataset. The PCA analysis was performed, obtaining the same 4 components as the

Table 6. Cluster Analysis at 50 percent of completion

Cluster label	Influence	Duration	Cost	Overbudget
P1	−0.25	−0.65	−0.37	−0.16
P2	12.15	−0.32	5.25	−0.30
P3	0.13	−0.30	6.75	−0.38
P4	−0.27	0.50	−0.23	4.17
P5	−0.10	0.83	−0.26	−0.18
P6	0.02	2.85	−0.25	−0.27
P7	1.15	0.46	2.34	−0.32

Source(s): Authors' own creation

at-completion analysis. For the K-means analysis, the same 3-step verification process (Silhouette, Elbow method, and hierarchical clustering) was applied, and we obtained a total of 7 clusters to analyze.

The analysis of megaproject clusters at 50% completion reveals varying impacts on economic and operational metrics across different categories. *Cluster Partial 1 (P1)* consists of well-balanced megaprojects characterized by the shortest durations and relatively low costs, indicating efficient management and effective progress control. *Cluster P2* includes megaprojects that already have a significant influence on the GDP relative to their costs, highlighting their pivotal role in their respective countries' economies, despite being only halfway completed. *Cluster P3* features high-cost megaprojects that exert a moderate to low impact on GDP. *Cluster P4* is comprised of megaprojects that frequently exceed budget expectations but maintain low costs and medium to high durations, reflecting either optimistic initial cost estimations or unforeseen challenges as they progress. *Cluster P5* projects demonstrate longer durations with minimal over-budget occurrences, indicating steady execution. *Cluster P6* includes long-term megaprojects that manage to keep costs low. Finally, *Cluster P7* includes megaprojects that, while exhibiting low over-budget incidents, still involve a moderate to high influence on GDP, balancing cost, duration, and economic impact effectively.

After gathering the cluster data relating to a variety of projects at different stages of completion, we proceeded to implement the C-BA method by comparing how the different projects dynamically move to different clusters across progress variation. Figure 6 represents the outcome of the C-BA analysis, where the evolution of megaprojects from the clusters identified at the 50% advancement and at completion can be seen. The method captures the dynamism of megaproject progress, and how, depending on management decisions and specific situations, they can end up with many different outcomes. For example, we can see that if a new project at 50% completion enters *Cluster P3*, which is the cluster showing high costs, the C-BA analysis shows that previous projects in this same cluster ended up at their completion either in *Cluster F2*, which also shows high costs, or in *F4*, that is average in the same dimension, showing two possible outcomes. The situation can be further analyzed in terms of the differences between the different paths, and how decisions or specific situations affected the project outcomes. Again, if a new project falls into the *P4 cluster* indicating high duration, similar projects have ended up in two different clusters, showing different degrees of duration. This analysis illustrates the fluid nature of megaproject development, underscoring

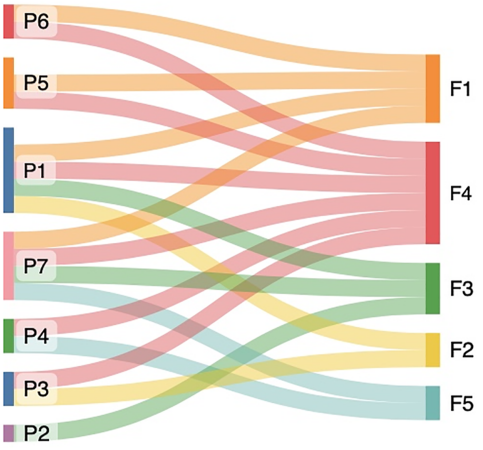

Figure 6. C-BA outcome. Source: Authors' own creation

the C-BA method's ability to adaptively reflect the shifts in project dynamics based on management decisions and varying project conditions. The transitions between clusters, as depicted in, reveal the multifaceted paths that projects can traverse, depending on evolving project conditions and strategic decisions made along the way.

5.3 The holistic application of C-BA

To demonstrate the holistic application of the C-BA method, we extended our analysis beyond the midpoint and completion stages of the megaprojects. This enabled us to conduct a dynamic behavioral analysis across the entire lifecycles of the megaprojects, providing a comprehensive understanding of their evolution from start to finish. To demonstrate this point, we conducted an exemplary analysis using only the initial variables of the projects "Relative value in 2021 of initial budget" and "Initial Budget/GDP". Since there were only two variables under consideration, we did not apply Principal Component Analysis. Instead, we directly applied K-means clustering after standardizing the variables. The three-step method involving the elbow, silhouette, and dendrogram, indicated that there were five clusters for this initial case.

Figure 7 displays the results of clustering the initial variables and illustrates how projects are connected to their partial and full states. In this case, it can be observed that if a project, based on its initial state, is classified as I4, it could be expected to end up in clusters P1 or P3 when it reaches 50% completion. If it falls into P1, there is a higher degree of uncertainty because the possible outcomes may lead to clusters F1, F2, F3, or F4. Conversely, if the project ends up in cluster P3, the most likely scenarios are F4 or F2. This information provides guidance on the multiple potential outcomes depending on the project phase, offering insights for decision-makers in terms of strategic planning, based on the experiences of previous projects. Furthermore, the method could be expanded by adding more phases or even by sub-clustering the obtained clusters, if necessary, although this would introduce a greater degree of uncertainty and require further study.

6. Discussion

The Fifth Hand principle advocates incorporating an external perspective when approaching project forecasting, emphasizing the importance of learning from past projects (Love *et al.*, 2012a, b; Ika *et al.*, 2022; Pinto, 2023). The existing body of literature has predominantly reported RCF as the primary method for incorporating lessons from previous megaprojects and extracting insights to enhance the forecasting of new megaprojects (Flyvbjerg, 2008;

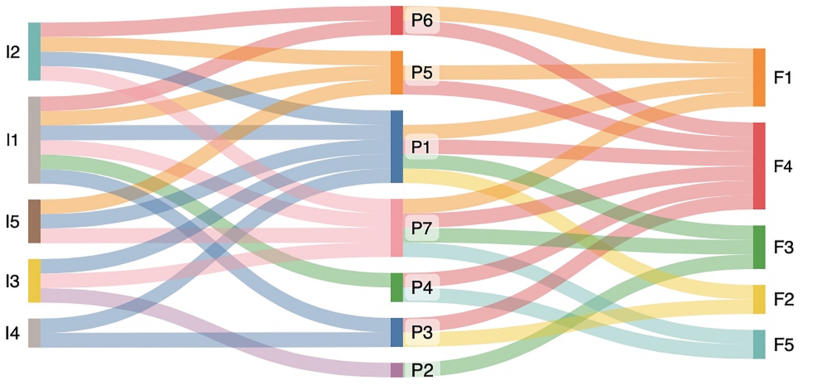

Figure 7. C-BA outcome including initial variables. Source: Authors' own creation

Flyvbjerg *et al*., 2016). However, as previously indicated, RCF exhibits several limitations, primarily due to the static assumptions on which it is based (Love *et al.*, 2022a, b, 2023a, b, c, d). Although it proves effective in megaproject scenarios marked by stability and well-defined risks, exact forecasting methods may not be well-suited for situations characterized by a high degree of uniqueness and uncertainty, which can frequently occur in such projects (Love *et al.*, 2023a, b, c, d).

The C-BA method introduces a new perspective on integrating external viewpoints into megaproject forecasting, enabling the identification of latent patterns in data from previous megaprojects. This approach emphasizes a data-driven method that does not produce exact forecasts, but rather offers a cluster-based analysis of similar behaviors across projects at various stages of their development. This provides project managers with insights that allow them to make informed decisions based on knowledge gleaned from past projects. As indicated in literature, we acknowledge that numerous factors during the planning and implementation phases may lead to a shortfall in megaproject management performance compared to the initial expectations (Ika, 2018). In the forecasting phase, planners might succumb to optimism bias, resulting in an underestimation of the time and cost associated with a specific project (Flyvbjerg, 2016). As previously mentioned, the perspective of Planning Fallacy views these inaccuracies in estimations as a flaw. Human bias or strategic misrepresentation can skew estimates, leading to an overestimation of benefits and an underestimation of costs. Consequently, during the implementation phase, project managers find themselves consistently striving to realign their projects with the original plan (Love *et al.*, 2019). The perspective of a Hiding Hand offers an explanation for inaccurate estimations by attributing them to over-optimism, but in a manner that conceals potentially significant obstacles from view. Companies embark on these projects, and it is only during implementation that the extent of ignorance becomes apparent. However, project teams can leverage their creativity and problem-solving skills to achieve long-term successes that match or overcome the initial expectations, even though they still exceed the anticipated time and cost estimates (Ika *et al*., 2022).

We acknowledge, as emphasized by Ika (2018), that not only optimism bias but also a variety of internal elements within the project team such as leadership, team issues, scope changes and rework, might influence the outcome of megaprojects. Given this, we propose the C-BA as an analytical tool that can present, without optimism bias, the full spectrum of outcomes from past similar projects, thereby providing diagnostic methods for megaproject forecasting.

The analysis conducted through unsupervised clustering highlights similarities with regard to past projects and underscores the potential to derive certain rules that generalize the behavior of megaprojects belonging to a cluster and their evolution. Love *et al.* (2023a, b, c, d) recognize that to cope with uncertainty, decision-makers can effectively revert to heuristics intended as cognitive shortcuts that filter out irrelevant information, enabling the leveraging of fundamental psychological abilities for quick, economical judgments that often lead to efficient and effective outcomes. Differently, the results of our study suggest that employing clustering can be an effective foundation for making informed heuristic-driven choices. Regardless of the decision-making methodology employed by managers, we advocate for the use of the C-BA approach as a tool to facilitate context-aware decision-making. This method is both rapid and effective, offering data-driven support for decision-making in megaprojects and other scenarios.

6.1 Implication for theory

Our paper presents three main implications for theory. The primary contribution of our study is the introduction of an innovative approach to enhance forecasting for megaprojects aligning with the Fifth Hand, a novel concept that needs to be further explored. This paper proposes an effective measures to assist researchers and practitioners in harnessing this perspective to

support project forecasting, by facilitating the extraction of insights from past megaproject datasets. Therefore, contributing to the Planning Fallacy debate by advocating C-BA as a method that supports "moving away from dualisms like Planning Fallacy versus Hiding Hand, or bias versus error, to dualities of Planning Fallacy and Hiding Hand or bias and error" (Ika *et al*., 2022). In fact, our approach involves adopting an external perspective for supporting forecasting that positions itself as a data-oriented foundation for heuristic decision-making. As reported by Love *et al.*, 2022a, b, heuristic decision making derives from reinforcement learning, in other words learning by experience of past projects. In this sense, we might consider the patterns derived from our megaprojects dataset as an effective heuristic selection aid.

The second contribution is related to complexity studies in megaprojects. In the context of megaproject forecasting, the application of Complex Network Theory (CNT) has shown the significant challenge of managing interdependencies and nonlinear interactions that characterize large-scale projects (Pryke *et al.*, 2018; Guo *et al*., 2020). This theory calls for models that can handle the often-unpredictable dynamics of megaprojects, thus our paper enriches this theoretical perspective by advocating for the adoption of dynamic forecasting making use of unsupervised machine learning to continuously adapt and assimilate emerging information. This integration enhances the predictive accuracy of decision makers and enriches the theoretical understanding of how complex systems can be effectively managed and forecasted through continuous updating and learning from the experience, aligning with the Organizational Learning Theory (OLT), stating that preserving institutional memory is crucial for learning from the past (Crossan *et al.*, 1995). We emphasize the significance of broadening the scope of learning beyond individual organizational experiences, particularly in the context of new megaprojects where similar configurations seldom recur. We contend that integrating data from external organizations is critical to enhancing the depth and breadth of learning (Delise *et al*., 2023)

The introduction of clustering-based analyses as a support for megaproject forecasting also extends its influence on another dimension of organizational learning theory, specifically the concept of "double loop learning" (Argyris, 1977; Auqui-Caceres and Furlan, 2023). Organizations engaged in double loop learning continuously reflect on and challenge their core assumptions, leading to a loop of constant improvement. By looking at the C-BA method as an organizational knowledge base, organizations can not only monitor and refine current management strategies, but also proactively shape future approaches for better project outcomes (Auqui-Caceres and Furlan, 2023). Moreover, by implementing the C-BA method, it is possible to perform a comparative performance evaluation of a current focal project against historical trends. For example, a current focal project at a certain stage of progress can be compared with the trajectory and dynamics of a past project to better understand potential future progress perspectives.

6.2 Implication for practice

Project management often involves an intuitive understanding of a project's trajectory, derived from the project manager's past experiences and subjective insights (Litsiou *et al.*, 2022). This paper proposes a novel framework for articulating these subjective perceptions, providing a structured approach to predict project behavior across different stages of the project lifecycle.

The primary insight of this paper for practice lies in recognizing that, despite their high degree of uniqueness, it is possible to learn from past megaprojects. Specifically, in this paper we propose the C-BA approach as a qualitative method that can aid project managers in deriving insights from past or external megaprojects for understanding the possible behavior of new ones. Managers and project team members, on the basis of company's needs, can easily implement this method by selecting both the number of project stages to analyze and the variables to be considered. For example, starting from a dataset of past

completed megaprojects, including their progress data, project managers could decide to conduct a C-BA on a tailored number of critical stages. This analysis is dynamic in two ways: (1) data from different stages of progress can generate a varying number of clusters, allowing for an analysis that dynamically tracks the shift of megaprojects into clusters with different characteristics; (2) when the analysis is repeated, the principal components may change, leading to a dynamic shift in the variables that explain the majority of the variance. This allows to perform a completely dynamic data driven analysis, minimizing the biases arising from reference classes' selection. Once the analysis of historical data is complete, comparing a new focal project with this historical data becomes possible. This comparison involves mapping the progress of the new project against established benchmarks and trends identified in the historical analysis. By using the cluster analyses conducted at different stages, project managers can position the new project within these predefined clusters to see how it aligns with or deviates from past projects. This comparison benefits forecasting by predicting potential outcomes, identifying risk areas, and determining which aspects of the project may require closer monitoring or adjustment. Additionally, the dynamic nature of the analysis allows for adjustments based on shifts in key variables or changes in project characteristics over time, making the comparison adaptive to new developments as the project progresses. To facilitate the implementation of the Clustering-Behavior Analysis (C-BA) method, we recommend utilizing a suite of software tools that are adept at data integration and advanced cluster analysis. For example, Python, with its extensive libraries such as Pandas for data manipulation, Scikit-learn for machine learning, and Matplotlib for data visualization, is particularly well-suited for implementing the C-BA method. Other tools that could also be beneficial include R for statistical computing and MATLAB for handling complex numerical calculations and visualizations. Figure 8 below aims to explicitly demonstrate, step-by-step, how the C-BA method can be implemented in real project settings.

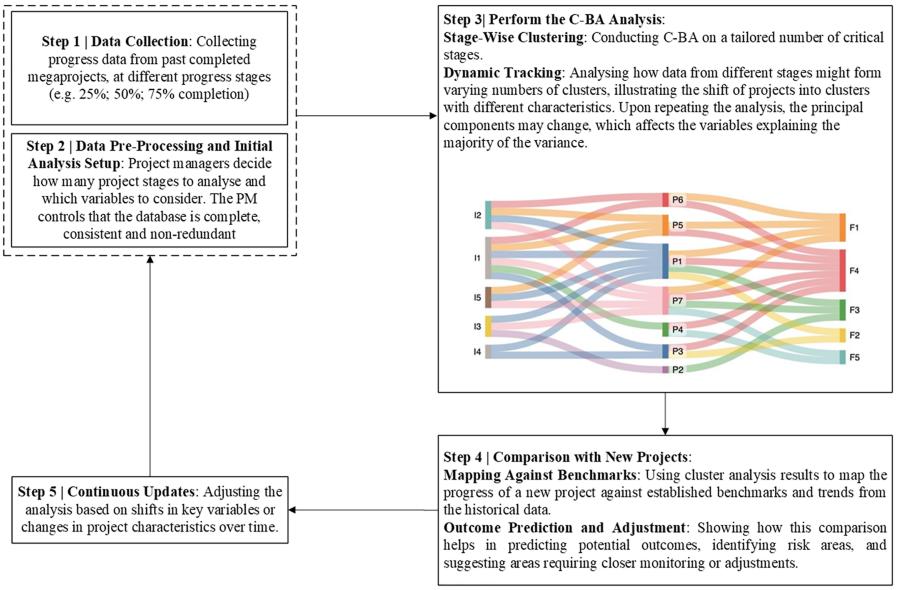

Figure 8. Framework for C-BA application. Source: Authors' own creation

7. Limitations and conclusion

This work contributes to the ongoing megaproject Planning Fallacy debate by introducing C-BA as a possible alternative method to support megaproject performance estimation from a Fifth Hand perspective (Ika *et al*., 2022). By leveraging these machine learning-based techniques, this research offers project professionals a powerful tool when it comes to extracting information from a database of past project data. This resource can then be utilized to conduct comprehensive analyses of lessons learned, identifying trends, patterns and critical success factors. Consequently, this paper fills a void in the existing literature, providing insights into a novel supporting method for megaproject forecasting that enhances the diagnosis of megaproject management performance issues.

We do recognize, however, certain limitations in this study that can be explored in future research. First, we acknowledge that our database and analysis primarily rely on measures of quantitative performance, namely, duration/cost overruns. The decision to focus on these variables, deemed particularly significant for megaprojects, is supported by the literature (Love *et al.*, 2012a, b, 2019; Ika, 2018), and allowed us to build up a reliable database which is built on publicly-accessible information. However, the method is flexible, and allows the inclusion of different and project specific variables, which can be tested in future research. Indeed, recent studies stress that evaluating a megaproject involves more than just considering cost and duration overruns; it is equally important to account for the anticipated benefits to stakeholders and the wider community (Ika and Pinto, 2022a, b). This implies a shift in focus from mere efficiency to more comprehensive, long-term outcomes. In this context, variables such as the ecological footprint, contributions to the local area and employment rates become critical when applying C-BA and might be included in future studies (Ika and Pinto, 2022a, b; Pinto *et al*., 2022).

Secondly, in this paper, the C-BA is performed using only 3 megaproject progress stages as testing grounds. Future research could hypothesize additional scenarios and set up a comparison between the trajectory of a current focal project and a dynamic analysis based on historical data. Furthermore, the intermediate progress data should be obtained from empirical contexts. This approach would enhance the practical applicability of the analysis and provide deeper insights into how our theoretical models perform in real-world settings. Thus, while we have discussed the theoretical application of our model to our database, further research in real organizational settings is necessary to examine how this methodology affects operational practices, shapes learning from past projects, and influences decision-making strategies when addressing focal projects. Finally, the last limitation of this study is the number of megaprojects used for the analysis: 90. While the number is not excessively small, the application of machine learning techniques tends to yield more accurate and meaningful results with an increased sample size (Kinkel *et al*., 2022). Therefore, an avenue for future research could involve expanding the database by incorporating more megaproject items.

References

Alikhani, H. and Jeong, D. (2021), "Highway project clustering using unsupervised machine learning approach", *ASCE International Conference on Computing in Civil Engineering 2021*.

Andranovich, G., Burbank, M.J. and Heying, C.H. (2001), "Olympic cities: lessons learned from mega-event politics", *Journal of Urban Affairs*, Vol. 23 No. 2, pp. 113-131, doi: 10.1111/0735-2166.00079.

Anheier, H. (2016), "Of hiding hands and other ways of coping with uncertainty: a commentary", *Social Research*, Vol. 83 No. 4, pp. 1005-1010, doi: 10.1353/sor.2016.0064.

Ansar, A., Flyvbjerg, B., Budzier, A. and Lunn, D. (2014), "Should we build more large dams? The actual costs of hydropower megaproject development", *Energy Policy*, Vol. 69, pp. 43-56, doi: 10.1016/J.ENPOL.2013.10.069.

Antunes, M., Gomes, D. and Aguiar, R.L. (2018), "Knee/Elbow estimation based on first derivative threshold", *Proceedings - IEEE 4th International Conference on Big Data Computing Service and Applications, BigDataService 2018*, pp. 237-240, doi: 10.1109/BigDataService.2018.00042.

Argyris, C. (1977), "Double loop learning in organizations", *Harvard Business Review*, Vol. 55 No. 5, pp. 115-125.

Arthur, D. and Vassilvitskii, S. (2007), "k-means++: the advantages of careful seeding", *SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*.

Auqui-Caceres, M. and Furlan, A. (2023), "Revitalizing double-loop learning in organizational contexts: a systematic review and research agenda", *European Management Review*, Vol. 21 No. 1, doi: 10.1111/emre.12615.

Awodi, N.J., Liu, Y.K., Ayodeji, A. and Adibeli, J.O. (2021), "Expert judgement-based risk factor identification and analysis for an effective nuclear decommissioning risk assessment modeling", *Progress in Nuclear Energy*, Vol. 136, 103733, doi: 10.1016/J.PNUCENE.2021.103733.

Awojobi, O. and Jenkins, G.P. (2015), "Were the hydro dams financed by the World Bank from 1976 to 2005 worthwhile?", *Energy Policy*, Vol. 86, pp. 222-232, doi: 10.1016/j.enpol.2015.06.040.

Baerenbold, R. (2023), "Reducing risks in megaprojects: the potential of reference class forecasting", *Project Leadership and Society*, Vol. 4, 100103, doi: 10.1016/j.plas.2023.100103.

Barro, R.J. (2013), "Inflation and economic growth *", *Annals of Economics and Finance*, Vol. 14 No. 1, pp. 85-109.

Batselier, J. and Vanhoucke, M. (2016), "Practical application and empirical evaluation of reference class forecasting for project management", *Project Management Journal*, Vol. 47 No. 5, pp. 36-51, doi: 10.1177/875697281604700504.

Bayram, S. and Al-Jibouri, S. (2016), "Efficacy of estimation methods in forecasting building projects' costs", *Journal of Construction Engineering and Management*, Vol. 142 No. 11, doi: 10.1061/(ASCE)CO.1943-7862.0001183.

Ben Salem, S., Naouali, S. and Chtourou, Z. (2018), "A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach", *Computers and Electrical Engineering*, Vol. 68, pp. 463-483, doi: 10.1016/J.COMPELECENG.2018.04.023.

Berlin, S., Raz, T., Glezer, C. and Zviran, M. (2009), "Comparison of estimation methods of cost and duration in IT projects", *Information and Software Technology*, Vol. 51 No. 4, pp. 738-748, doi: 10.1016/j.infsof.2008.09.007.

Brady, T. and Davies, A. (2004), "Building project capabilities: from exploratory to exploitative learning", *Organization Studies*, Vol. 25 No. 9, pp. 1601-1621, doi: 10.1177/0170840604048002.

Brockmann, C. (2020), *Advanced Construction Project Management, Advanced Construction Management*, John Wiley & Sons, doi: 10.1002/9781119554721.ch3.

Brookes, N.J. and Locatelli, G. (2015), "Power plants as megaprojects: using empirics to shape policy, planning, and construction management", *Utilities Policy*, Vol. 36, pp. 57-66, doi: 10.1016/j.jup.2015.09.005.

Callegari, C., Szklo, A. and Schaeffer, R. (2018), "Cost overruns and delays in energy megaprojects: how big is big enough?", *Energy Policy*, Vol. 114, pp. 211-220, doi: 10.1016/j.enpol.2017.11.059.

Cao, Y. and Ashuri, B. (2020), "Predicting the volatility of highway construction cost index using long short-term memory", *Journal of Management in Engineering*, Vol. 36 No. 4, doi: 10.1061/(ASCE)ME.1943-5479.0000784.

Celebi, M. (2015), "Partitional cluster algorithms - introduction", in *Partitional Cluster Algorithms*, Springer, pp. 5-8.

Cha, G.-W., Park, C.W., Kim, Y.C. and Moon, H.J. (2023), "Predicting generation of different demolition waste types using simple artificial neural networks", *Sustainability*, Vol. 15 No. 23, 16245, doi: 10.3390/su152316245.

Charu, A. and Chandan, K R. (2013), "Data clustering algorithms and applications, an introduction to toxicogenomics", doi: 10.1201/9781315373515-1.

Cheng, M.Y., Tsai, H.C. and Sudjono, E. (2010), "Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry", *Expert Systems with Applications*, Vol. 37 No. 6, pp. 4224-4231, doi: 10.1016/j.eswa.2009.11.080.

Crossan, M.M., Lane, H.W., White, R.E. and Djurfeldt, L. (1995), "Organizational learning: dimensions for a theory", *International Journal of Organizational Analysis*, Vol. 3 No. 4, pp. 337-360, doi: 10.1108/eb028835.

Darko, A., Glushakova, I., Boateng, E.B. and Chan, A.P.C. (2023), "Using machine learning to improve cost and duration prediction accuracy in green building projects", *Journal of Construction Engineering and Management*, Vol. 149 No. 8, doi: 10.1061/JCEMD4.COENG-13101.

Davies, A. and Brady, T. (2000), "Organisational capabilities and learning in complex product systems: towards repeatable solutions", *Research Policy*, Vol. 29 Nos 7-8, pp. 931-953, doi: 10.1016/s0048-7333(00)00113-x, available at: www.elsevier.nlrlocatereconbase

Delise, L.A., Lee, B. and Choi, Y. (2023), "Understanding project management performance using a comparative overrun measure", *International Journal of Project Management*, Vol. 41 No. 2, 102450, doi: 10.1016/j.ijproman.2023.102450.

Denicol, J., Davies, A. and Krystallis, I. (2020), "What are the causes and cures of poor megaproject performance? A systematic literature review and research agenda", *Project Management Journal*, Vol. 51 No. 3, pp. 328-345, doi: 10.1177/8756972819896113.

Dursun, O. and Stoy, C. (2016), "Conceptual estimation of construction costs using the multistep ahead approach", *Journal of Construction Engineering and Management*, Vol. 142 No. 9, doi: 10.1061/(ASCE)CO.1943-7862.0001150.

Eren, F. (2019), "Top government hands-on megaproject management: the case of Istanbul's grand airport", *International Journal of Managing Projects in Business*, Vol. 12 No. 3, pp. 666-693, doi: 10.1108/IJMPB-02-2018-0020.

Flyvbjerg, B. (2008), "Curbing optimism bias and strategic misrepresentation in planning: reference class forecasting in practice", *European Planning Studies*, Vol. 16 No. 1, pp. 3-21, doi: 10.1080/09654310701747936.

Flyvbjerg, B. (2014), "What you should know about megaprojects and why: an overview", *Project Management Journal*, Vol. 45 No. 2, pp. 6-19, doi: 10.1002/pmj.21409.

Flyvbjerg, B. (2016), "The fallacy of beneficial ignorance: a test of hirschman's hiding hand", *World Development*, Vol. 84, pp. 176-189, doi: 10.1016/j.worlddev.2016.03.012.

Flyvbjerg, B., Holm, M.S. and Buhl, S. (2002), "Underestimating costs in public works projects: error or lie?", *Journal of the American Planning Association*, Vol. 68 No. 3, pp. 279-295, doi: 10.1080/01944360208976273.

Flyvbjerg, B., Garbuio, M. and Lovallo, D. (2009), "Delusion and deception in large infrastructure projects: two models for explaining and preventing executive disaster", *California Management Review*, Vol. 51 No. 2, pp. 170-193, doi: 10.1225/CMR423.

Flyvbjerg, B., Hon, C.K. and Fok, W.H. (2016), "Reference class forecasting for Hong Kong's major roadworks projects", *Proceedings - Institution of Civil Engineers: Civil Engineering*, Vol. 169 No. 6, pp. 17-24, doi: 10.1680/jcien.15.00075.

Fridgeirsson, T.V. (2016), "Reference class forecasting in Icelandic transport infrastructure projects", *Transport Problems*, Vol. 11 No. 2, pp. 103-115, doi: 10.20858/tp.2016.11.2.10.

Gentleman, R. and Carey, V. (2008), "Unsupervised machine learning", in *Bioconductor Case Studies*, Springer, pp. 137-157.

Gigerenzer, G. (2013), *Risk Savvy: How to Makegood Decisions*, Penguin, Baltimore, MD, USA, 2013. Baltimore, MD, USA: Penguin.

Gondia, A., Siam, A., El-Dakhakhni, W. and Nassar, A.H. (2020), "Machine learning algorithms for construction projects delay risk prediction", *Journal of Construction Engineering and Management*, Vol. 146 No. 1, doi: 10.1061/(ASCE)CO.1943-7862.0001736.

Greiman, V.A. (2013), *Megaproject Management: Lessons on Risk and Project Management from the Big Dig*, 1st ed., Wiley, doi: 10.1002/9781118671092.

Guo, N., Guo, P., Madhavan, R., Zhao, J. and Liu, Y. (2020), "Assessing the vulnerability of megaprojects using complex network theory", *Project Management Journal*, Vol. 51 No. 4, pp. 429-439, doi: 10.1177/8756972820911236.

Hahne, F. and Gentlemann, R. (2008), "The all dataset", in *Bioconductor Case Studies*, Springer.

Han, S.H., Yun, S., Kim, H., Kwak, Y.H., Park, H.K. and Lee, S.H. (2009), "Analyzing schedule delay of mega project: Lesson learned from Korea train express", *IEEE Transactions on Engineering Management*, Vol. 56 No. 2, pp. 243-256, doi: 10.1109/tem.2009.2016042.

He, Q., Wang, T., Chan, A.P.C. and Xu, J. (2021), "Developing a list of key performance indictors for benchmarking the success of construction megaprojects", *Journal of Construction Engineering and Management*, Vol. 147 No. 2, doi: 10.1061/(ASCE)CO.1943-7862.0001957.

Hsu, M., Dacre, N. and Senyo, P. (2021), "Applied algorithmic machine learning for intelligent project prediction: towards an AI framework of project success", *Advanced Project Management*, Vol. 21 No. 4, doi: 10.2139/ssrn.3823900.

Huber, J., Gossmann, A. and Stuckenschmidt, H. (2017), "Cluster-based hierarchical demand forecasting for perishable goods", *Expert Systems with Applications*, Vol. 76, pp. 140-151, doi: 10.1016/j.eswa.2017.01.022.

Ika (2018), "Beneficial or detrimental ignorance: the straw man fallacy of flyvbjerg's test of hirschman's hiding hand", *World Development*, Vol. 103, pp. 369-382, doi: 10.1016/j.worlddev.2017.10.016.

Ika, L.A. and Pinto, J.K. (2022a), "International journal of project management the ' re-meaning ' of project success : updating and recalibrating for a modern project management", *International Journal of Project Management*, Vol. 40 No. 7, pp. 835-848, doi: 10.1016/j.ijproman.2022.08.001.

Ika, L.L.P.E.D. and Pinto, J.K. (2022b), "Moving beyond the planning fallacy: the emergence of a new principle of project behavior", *IEEE Transactions on Engineering Management*, Vol. 69 No. 6, pp. 3310-3325, doi: 10.1109/TEM.2020.3040526.

Ika, L.A. and Söderlund, J. (2016), "Rethinking revisited: insights from an early rethinker", *International Journal of Managing Projects in Business*, Vol. 9 No. 4, pp. 931-954, doi: 10.1108/IJMPB-05-2016-0041.

Ika, L.A., Love, P.E.D. and Pinto, J.K. (2022), "Moving beyond the planning fallacy: the emergence of a new principle of project behavior", *IEEE Transactions on Engineering Management*, Vol. 69 No. 6, pp. 3310-3325, doi: 10.1109/TEM.2020.3040526.

Invernizzi, D.C., Locatelli, G. and Brookes, N.J. (2018), "A methodology based on benchmarking to learn across megaprojects: the case of nuclear decommissioning", *International Journal of Managing Projects in Business*, Vol. 11 No. 1, pp. 104-121, doi: 10.1108/IJMPB-05-2017-0054.

Isaac, E. (2023), "Convenience and purposive sampling techniques: are they the same?", *International Journal of Innovative Social and Science Education Research*, Vol. 11 No. 1, pp. 1-7, available at: www.seahipaj.org

Jain, A.K. (2010), "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, Vol. 31 No. 8, pp. 651-666, doi: 10.1016/j.patrec.2009.09.011.

Juszczyk, M., Zima, K. and Lelek, W. (2019), "Forecasting of sports fields construction costs aided by ensembles of neural networks", *Journal of Civil Engineering and Management*, Vol. 25 No. 7, pp. 715-729, doi: 10.3846/jcem.2019.10534.

Kahneman, D. and Tversky, A. (1979), "Prospect theory: an analysis of decision under risk", *Econometrica*, Vol. 47 No. 2, available at: https://doi.org/10.2307/1914185 (accessed 15 April 2025).

Kaiser, M.J. and Snyder, B. (2012), "Offshore wind capital cost estimation in the U.S. outer Continental shelf—A reference class approach", *Marine Policy*, Vol. 36 No. 5, pp. 1112-1122, doi: 10.1016/j.marpol.2012.02.001.

Kinkel, S., Baumgartner, M. and Cherubini, E. (2022), "Prerequisites for the adoption of AI technologies in manufacturing – evidence from a worldwide sample of manufacturing companies", *Technovation*, Vol. 110 No. July 2020, 102375, doi: 10.1016/j.technovation.2021.102375.

Ko, C.-H. and Cheng, M.-Y. (2007), "Dynamic prediction of project success using artificial intelligence", *Journal of Construction Engineering and Management*, Vol. 133 No. 4, pp. 316-324, doi: 10.1061/(ASCE)0733-9364(2007)133:4(316).

Kreiner, K. (2020), "Conflicting notions of a project: the battle between Albert O. Hirschman and bent flyvbjerg", *Project Management Journal*, Vol. 51 No. 4, pp. 400-410, doi: 10.1177/8756972820930535.

Leleur, S., Salling, K.B., Pilkauskiene, I. and Nicolaisen, M.S. (2015), "Combining reference class forecasting with overconfidence theory for better risk assessment of transport infrastructure investments", *European Journal of Transport and Infrastructure Research*, Vol. 15 No. 3, pp. 362-375, doi: 10.18757/ejtir.2015.15.3.3083.

Lepenies, P.H. (2018), "Statistical tests as a hindrance to understanding: what the controversy around the "Hiding Hand" reveals about research in the social sciences and conceals about project management", *World Development*, Vol. 103, pp. 360-365, doi: 10.1016/j.worlddev.2017.10.017.

Litsiou, K., Polychronakis, Y., Karami, A. and Nikolopoulos, K. (2022), "Relative performance of judgmental methods for forecasting the success of megaprojects", *International Journal of Forecasting*, No. 3, doi: 10.1016/j.ijforecast.2019.05.018.

López-Martín, C. and Abran, A. (2015), "Neural networks for predicting the duration of new software projects", *Journal of Systems and Software*, Vol. 101, pp. 127-135, doi: 10.1016/j.jss.2014.12.002.

Lovallo, D., Clarke, C. and Camerer, C. (2012), "Robust analogizing and the outside view: two empirical tests of case-based decision making", *Strategic Management Journal*, Vol. 33 No. 5, pp. 496-512, doi: 10.1002/smj.962.

Love, P.E.D., Wang, X., Sing, C.P. and Tiong, R.L.K. (2012a), "Determining the probability of project cost overruns", *Journal of Construction Engineering and Management*, Vol. 139 No. 3, pp. 321-330, doi: 10.1061/(ASCE)CO.1943-7862.0000575.

Love, P.E.D., Edwards, D.J. and Irani, Z. (2012b), "Moving beyond optimism bias and strategic misrepresentation: an explanation for social infrastructure project cost overruns", *IEEE Transactions on Engineering Management*, Vol. 59 No. 4, pp. 560-571, doi: 10.1109/TEM.2011.2163628.

Love, P.E.D., Sing, M.C., Ika, L.A. and Newton, S. (2019), "The cost performance of transportation projects: the fallacy of the planning fallacy account", *Transportation Research Part A: Policy and Practice*, Vol. 122, pp. 1-20, doi: 10.1016/j.tra.2019.02.004.

Love, P.E.D., Ika, L.A. and Pinto, J.K. (2022a), "Homo heuristicus: from risk management to managing uncertainty in large-scale infrastructure projects", *IEEE Transactions on Engineering Management*, Vol. 71, pp. 1-10, doi: 10.1109/TEM.2022.3170474.

Love, P.E.D., Pinto, J.K. and Ika, L.A. (2022b), "Hundreds of years of pain, with minimal gain: Capital project cost overruns, the past, present, and optimistic future", *IEEE Engineering Management Review*, Vol. 50 No. 4, pp. 56-70, doi: 10.1109/EMR.2022.3219362.

Love, P.E.D., Ika, L.A., Matthews, J. and Fang, W. (2023a), "Large-scale transport infrastructure project performance: generating a narrative of context and meaning", *IEEE Transactions on Engineering Management*, Vol. 70 No. 10, pp. 3637-3652, doi: 10.1109/TEM.2021.3094511.

Love, P.E.D., Ika, L.A. and Pinto, J.K. (2023b), "Fast-and-frugal heuristics for decision-making in uncertain and complex settings in construction", *Developments in the Built Environment*, Vol. 14, 100129, doi: 10.1016/j.dibe.2023.100129.

Love, P.E.D., Matthews, J. and Ika, L.A. (2023c), "Fast-and-frugal heuristics: an exploration into building an adaptive toolbox to assess the uncertainty of rework", *Production Planning and Control*, Vol. 36 No. 3, pp. 251-266, doi: 10.1080/09537287.2023.2257178.

Love, P.E.D., Matthews, J. and Ika, L.A. (2023d), "Fast-and-frugal heuristics: an exploration into building an adaptive toolbox to assess the uncertainty of rework", *Production Planning and Control*, pp. 1-16, [Preprint], doi: 10.1080/09537287.2023.2257178.

Luo, H., Chen, J., Love, P.E. and Fang, W. (2024), "Explainable transfer learning for modeling and assessing risks in tunnel construction", *IEEE Transactions on Engineering Management*, Vol. 71, pp. 8339-8355, doi: 10.1109/TEM.2024.3369231.

Madhulatha, S. (2012), "An overview on clustering methods", *Journal of Engineering*, Vol. 2 No. 4, pp. 719-725, ISSN 2250-3021.

Mancini, M., Mariani, C. and Manfredi, M. (2023), "Nuclear decommissioning risk management adopting a comprehensive artificial intelligence framework : an applied case in an Italian site", *Progress in Nuclear Energy*, Vol. 158, February, 104589, doi: 10.1016/j.pnucene.2023.104589.

Mariani, C. and Mancini, M. (2023), "Selection of projects' primary and secondary mitigation actions through optimization methods in nuclear decommissioning projects", *Nuclear Engineering and Design*, Vol. 407, 112284, doi: 10.1016/j.nucengdes.2023.112284.

Mariani, C., Navrotska, Y. and Mancini, M. (2023), "Unsupervised machine learning for project stakeholder classification: benefits and limitations", *Project Leadership and Society*, Vol. 4, 100093, doi: 10.1016/j.plas.2023.100093.

McLeod, S. (2023), "Rethinking public infrastructure megaproject performance: theorizing alternative benefits, and the need for open science in project research", *Project Leadership and Society*, Vol. 4, 100080, doi: 10.1016/j.plas.2023.100080.

Merrow, E.W. (2012), "Oil and gas industry megaprojects: our recent track record", *Oil and Gas Facilities*, Vol. 1 No. 2, pp. 38-42, doi: 10.2118/153695-PA.

Mouton, J.P., Ferreira, M. and Helberg, A.S.J. (2020), "A comparison of clustering algorithms for automatic modulation classification", *Expert Systems with Applications*, Vol. 151, 113317, doi: 10.1016/j.eswa.2020.113317.

Natarajan, A. (2022), "Reference class forecasting and machine learning for improved offshore oil and gas megaproject planning: methods and application", *Project Management Journal*, Vol. 53 No. 5, pp. 456-484, doi: 10.1177/87569728211045889.

Pérez Vera, Y. and Bermudez Peña, A. (2018), "Stakeholders classi fi cation system based on clustering techniques", *Lecture Notes in Computer Science*, Vol. 11238, pp. 241-252, doi: 10.1007/978-3-030-03928-8.

Pinto, J.K. (2023), "Is this how big things get done?", *International Journal of Project Management*, Vol. 41 No. 5, 102484, doi: 10.1016/j.ijproman.2023.102484.

Pinto, J.K., Davis, K., Ika, L.A., Jugdev, K. and Zwikael, O. (2022), "Coming to terms with project success: current perspectives and future challenges", *International Journal of Project Management*, Vol. 40 No. 7, pp. 831-834, doi: 10.1016/j.ijproman.2022.09.001.

Pospieszny, P., Czarnacka-Chrobot, B. and Kobylinski, A. (2018), "An effective approach for software project effort and duration estimation with machine learning algorithms", *Journal of Systems and Software*, Vol. 137, pp. 184-196, doi: 10.1016/j.jss.2017.11.066.

Pryke, S., Badi, S. and Addyman, S. (2018), "Self-organizing networks in complex infrastructure projects", *Project Management Journal*, Vol. 49 No. 2, doi: 10.1177/875697281804900202.

Rajabi Asadabadi, M. and Zwikael, O. (2024), "Unrealistic project goals: detection and modification", *Journal of Construction Engineering and Management*, Vol. 150 No. 3, doi: 10.1061/JCEMD4.COENG-13665.

Rousseeuw, P.J. (1987), "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53-65, doi: 10.1016/0377-0427(87)90125-7.

Roy, M.M., Mitten, S.T. and Christenfeld, N.J.S. (2008), "Correcting memory improves accuracy of predicted task duration", *Journal of Experimental Psychology: Applied*, Vol. 14 No. 3, pp. 266-275, doi: 10.1037/1076-898X.14.3.266.

Salling, K.B. and Leleur, S. (2015), "Transport project evaluation: feasibility risk assessment and scenario forecasting", *Transport*, Vol. 32 No. 2, pp. 180-191, doi: 10.3846/16484142.2015.1063003.

Sassano, G. (2025), "The holistic view in forecasting: a conceptual framework to analyze and mitigate cost underestimation arising from optimism bias", *Project Leadership and Society*, Vol. 6, 100177, doi: 10.1016/j.plas.2025.100177.

Seyedan, M. and Mafakheri, F. (2020), "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities", *Journal of Big Data*, Vol. 7 No. 1, 53, doi: 10.1186/s40537-020-00329-2.

Singh, K., Malik, D. and Sharma, N. (2011), "Evolving limitations in K-means algorithm in data mining and their removal", *IJCEM International Journal of Computational Engineering and Management*, available at: www.IJCEM.orgIJCEMwww.ijcem.org

Soni Madhulatha, T. (2012), "An overview on clustering methods", Vol. 2 No. 4, pp. 719-725, available at: www.iosrjen.org

Steininger, B.I., Groth, M. and Weber, B.L. (2021), "Cost overruns and delays in infrastructure projects: the case of stuttgart 21", *Journal of Property Investment and Finance*, Vol. 39 No. 3, pp. 256-282, doi: 10.1108/JPIF-11-2019-0144.

Suh, D.Y. and Ryerson, M.S. (2019), "Forecast to grow: aviation demand forecasting in an era of demand uncertainty and optimism bias", *Transportation Research Part E: Logistics and Transportation Review*, Vol. 128, pp. 400-416, doi: 10.1016/j.tre.2019.06.016.

Themsen, T.N. (2019), "The processes of public megaproject cost estimation: the inaccuracy of reference class forecasting", *Financial Accountability and Management*, Vol. 35 No. 4, pp. 337-352, doi: 10.1111/faam.12210.

Van Marrewijk, A., Clegg, S.R., Pitsis, T.S. and Veenswijk, M. (2008), "Managing public-private megaprojects: paradoxes, complexity, and project design", doi: 10.1016/j.ijproman.2007.09.007.

Volkmar, G., Fischer, P.M. and Reinecke, S. (2022), "Artificial intelligence and machine learning: exploring drivers, barriers, and future developments in marketing management", *Journal of Business Research*, Vol. 149, June, pp. 599-614, doi: 10.1016/j.jbusres.2022.04.007.

Vukomanović, M. , Cerić, A., Brunet, M., Locatelli, G. and Davies, A. (2021), "Editorial: trust and governance in megaprojects", *International Journal of Project Management*, Vol. 39 No. 4, pp. 321-324, doi: 10.1016/j.ijproman.2021.04.004.

Wang, Y.-R., Yu, C.-Y. and Chan, H.-H. (2012), "Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models", *International Journal of Project Management*, Vol. 30 No. 4, pp. 470-478, doi: 10.1016/j.ijproman.2011.09.002.

Wei, W. and Rana, M.E. (2019), "Software project schedule management using machine learning and data mining", *International Journal of Scientific and Technology Research*, Vol. 8 No. 9, pp. 1385-1389.

Zani, D. and Adey, B.T. (2025), "Swiss highway project cost estimate performance: deviations from norms and expected trends", *Case Studies on Transport Policy*, Vol. 19, 101344, doi: 10.1016/j.cstp.2024.101344.

Zani, D., Adey, B.T. and Carroll, S. (2024), "An approach to support reference class forecasting when adequate project data are unavailable", *Results in Engineering*, Vol. 22, 102333, doi: 10.1016/j.rineng.2024.102333.

Zheng, J., Wu, G. and Xie, H. (2017), "Impacts of leadership on project-based organizational innovation performance: the mediator of knowledge sharing and moderator of social capital", *Sustainability*, Vol. 9 No. 10, 1893, doi: 10.3390/su9101893.

Appendix

Table A1. Database variables

Variable name	Type	Description	Possible values	Assumptions and pre-processing	References
Industry type	Categorical	Type of industry to which the megaproject belongs	Infrastructure, Extractive industry, Research and development, Consumption	–	Greiman (2013)
Main sector	Categorical	Specific sector related to the industry type	Category of belonging (e.g.: road and transport infrastructure, aerospace, energy, etc.)	–	Greiman (2013)
Initial budget (IB)	Numerical	Planned and expected cost of construction	Numbers in terms of billions USD	Converted into US dollars (USD)	Merrow (2012), Natarajan (2022)
Relative value in 2021 of initial budget (*)	Numerical	The value that the initial budget has in 2021 considering inflation and GDP	Numbers in terms of billions USD	Converted into USD, adjusted to 2021 value through GDP deflator	
IB/GDP (*)	Numerical	The ratio between the initial value and the GDP of the country in that year	Percentage	GDP converted into USD	
Total cost of completion (TC)	Numerical	The total investments that the megaproject needed	Numbers in terms of billions USD	Converted into USD	Merrow (2012)
Relative value in 2021 of the final cost (*)	Numerical	The value that the final cost of completion has in 2021 taking Into account inflation and GDP	Numbers in terms of billions USD	In case of limited availability of data and negligible time delay, assumed to be equal to the initial budget. Converted into USD, adjusted to 2021 value through GDP deflator	
Final cost/ GDP (*)	Numerical	The ratio between the total cost of completion and the GDP of the country in that year	Percentage	GDP converted into USD	Flyvbjerg (2014)
Overbudget percentage (*)	Numerical	$(TC_{2021}\text{-}IB_{2021})/IB_{2021}*100$	Percentage	–	Merrow (2012)

(*continued*)

Table A1. Continued

Variable name	Type	Description	Possible values	Assumptions and pre-processing	References
Years of delay (*)	Numerical	Number of years of delay	Numbers (integers)	In case of unavailability of data and negligible overbudget, assumed to be 0	Merrow (2012)
Start of construction	Numerical	Year in which the construction of the megaproject begins	Years	–	Greiman (2013)
End of construction	Numerical	Year in which the construction of the megaproject ends	Years	–	Greiman (2013)
Years of delay/ duration	Numerical	The ratio between the number of years of delay and the total duration of the megaproject	Percentage	–	
Duration (*)	Numerical	Number of years in which the megaproject has been realized	Numbers (integers)	–	Greiman (2013)
Owner/ Contractor	Categorical	The one who has the ownership over the project and is accountable for the project's success or failure	Public, private, PPP, governmentetc	–	Brockmann (2020)
Region	Categorical	Geographical location of the project	Countries and continents	–	Greiman (2013), Natarajan (2022)
Extension	Categorical	Geographical extension of the project impact	City, national, multi-national	–	Greiman (2013)
Impact categories	Categorical	The impact of the project (local, national,international)	Military scope, national economic development, social progress, environmental protection, profit interest	–	Zheng *et al.* (2017)

Corresponding author
Francesco Cellerino can be contacted at: francesco.cellerino@polimi.it