Article

Project Management Journal
1–29

DOI: 10.1177/87569728211045889
journals.sagepub.com/home/pmx
SAGE

Reference Class Forecasting and Machine Learning for Improved Offshore Oil and Gas Megaproject Planning: Methods and Application

Ananth Natarajan[1]

Abstract

This article develops and describes rigorous oil and gas project forecasting methods. First, it builds a theoretical foundation by mapping megaproject performance literature to these projects. Second, it draws on heuristics and biases literature, using a questionnaire to demonstrate forecasting-related biases and principal-agent issues among industry project professionals. Third, it uses methodically collected project performance data to demonstrate that overrun distributions are non-normal and fat-tailed. Fourth, reference-class forecasting is demonstrated for cost and schedule uplifts. Finally, a predictive approach using machine learning (ML) considers project-specific factors to forecast the most likely cost and schedule overruns in a project.

Keywords

megaproject performance, reference class forecasting, oil and gas projects, machine learning, schedule and cost overruns, megaproject performance forecasting, planning, heuristics, biases

Introduction

This work applies machine learning (ML) to megaproject forecasting to improve cost and time performance planning. Megaprojects account for enormous investments, amounting to a substantial portion of the world's gross domestic product (GDP) (Flyvbjerg, 2014) and their key performance metrics are planned versus actual time, cost, and benefits. High cost and time overruns and benefit shortfalls are pervasive and chronic. Flyvbjerg (2014) estimates that only one in 1,000 megaprojects meets all three targets. Flyvbjerg et al. (2003) studied 258 global projects amounting to about US$90 billion and found an average 28% cost overrun; 9 out of 10 transportation projects were affected. Significant overruns plague infrastructure megaprojects (Flyvbjerg et al., 2018), big dams (Ansar, Flyvbjerg & Budzier, 2014), and big IT projects (Flyvbjerg & Budzier, 2011).

The primary goal of our study was to verify if ML can lead to better forecasting. Accurate cost and schedule estimations are challenged by complexity, principal-agent issues, and behavioral biases. We first investigated if these issues, which result in underestimated schedules and budgets, are present in industrial megaprojects. Having verified this, we verified if ML can be applied to effectively predict corrective project-specific cost and schedule uplifts using project features underlying misestimation.

Our work will contribute to the incipient body of knowledge on ML application to projects. Complexity and limited datasets make ML application to megaprojects challenging. Our ML models address this by building on reference class forecasting (RCF) methodology and predicting expected overruns from identified project features. There are distinct viewpoints on heuristics for decision-making under complexity with implications on the place of ML: one sees them favorably and the other emphasizes cognitive limitations. Along with their practical significance, our results are significant to the balance between expert judgment and ML. In our conclusion, we discuss how to combine their strengths, which can guide ML applications to projects beyond our specific application.

We begin by introducing the chosen case, offshore oil and gas (O&G) projects. We build the theoretical framework, based on which the research questions are formally stated as hypotheses and research methods are derived. We build on the theoretical grounding and demonstrate the presence of bias in industrial megaproject forecasting using a questionnaire sent to 26 O&G project managers, one-half each from O&G

[1] Cybereum, USA

Corresponding Author:
Ananth Natarajan, 1150W 22nd St, Unit B, Houston, Texas-77008, USA.
Email: ananth.natarajan@cybereum.io

companies and offshore EPC contractors. This is an important indicator of the potential and place of ML in advancing industrial megaproject management, which is then demonstrated through application. We methodically collected cost and schedule data for a large sample of commissioned offshore O&G projects from public and secondary data. This dataset was used to establish the extent of cost and schedule overruns in O&G offshore projects, demonstrate their fat-tailed non-normal distribution, and identify features that affect performance. The research results are discussed at length, followed by the conclusion, which summarizes our findings, discusses limitations, and identifies several interesting results for future research.

The Problem

Industrial megaprojects account for substantial and growing investments (Merrow, 2011). O&G megaprojects are challenging industrial megaprojects with high upfront investments for long-term returns in uncertain, complex socioeconomic-technical environments (Raval, 2020), and a track record of time, cost, and benefit underperformance (Merrow, 2012). Project portfolio selection and execution efficiency are core strategies for O&G companies (Singh, 2010), as for any sector where revenue-generating assets are created through megaprojects. Offshore project investment decisions amounted to US$92 billion in 2019, peaking at US$217 billion in 2011 (Rystad Energy, 2020). Underperformance has wider consequences. Cost overruns caused a 38% single-day tumble in an offshore EPC contractor's share price (Upstream, 2002). Schedule overruns cause significant financial losses (Caron & Ruggeri, 2016). In one project, severe overruns toppled both the oil company's board and the main contractor's top management (Upstream, 2000). Costs from a Norwegian project with massive overruns were transferred to taxpayers via tax deductions and state investments.

Theoretical Framework

The theoretical foundation intersects several streams of project performance research requiring extensive literature review. Furthermore, it is at the confluence of megaprojects, statistics, and ML, accentuating the transdisciplinary nature of management and organization studies (Denyer & Tranfield, 2009). Denicol et al. (2020) used a systematic literature review and filtering process to explore the causes of and cures for poor megaproject performance. The concepts they extracted were: (1) decision-making behavior; (2) strategy, governance, procurement; (3) risk and uncertainty; (4) leadership capability; (5) stakeholder engagement/management; and (6) supply-chain integration and coordination. As they discussed, currently, no overarching theory unites them. However, our systematic data collection allowed us to map these concepts, representing the body of megaproject research, to project performance. The framework illustrated in Figure 1 relates these concepts to factors we found to affect forecasting and overruns: decision-making behavior during planning and emergence of unplannable outcomes with inadequate response during execution. These concepts and factors, expatiated below, are systematically connected to our research methods and subsequently to research findings.

Complexity

Megaprojects are temporary organizations (Lundin & Söderholm, 1995) characterized by uncertainty (Denicol et al., 2020) and complexity (Baccarini, 1996). They display the hierarchy, inter-connectedness, emergence, sensitivity to initial conditions, and phase transition associated with complexity. Complex systems can display emergent and chaotic behavior (Hitchins, 2007). Emergent behavior is highly nonlinear, state-dependent, challenging to forecast, and can change rapidly (Warren, 2008). Chaotic behavior from high sensitivity to initial conditions makes theoretically deterministic outcomes practically unpredictable (Werndl, 2009). Chaotic behavior has been related to cost overruns in offshore O&G projects due to complex interactions and high exposure to change (Olaniran et al., 2015). Complexity engendered positive feedback and sudden phase change were observed in projects with extreme overruns from our dataset. Sometimes this results in *black swans*—high impact unpredictable events associated with complexity and characterized by retrospective sensemaking (Taleb, 2008).

Extreme Values

While extreme, outliers are significant to statistical analysis of overruns (Flyvbjerg et al., 2018). Flyvbjerg (2006) showed how cost outrun distributions for several infrastructure project classes were non-normal and significantly weighted toward overrun, with fat tails containing significant outliers. Similarly, Flyvbjerg and Budzier (2011) found that one-sixth of 1,471 IT projects they studied were black swan fat-tail outliers, with an average 200% cost and 70% schedule overrun. Project classes displaying "regression to the tail" (Flyvbjerg, 2020, p. 2) are susceptible to ever-larger tail risks, signifying outlier salience.

Randomness measures such as the normal distribution cannot effectively describe systems incorporating bias-prone human heuristics (Taleb, 2008). Power laws can describe fat-tailed distributions containing black swans, as demonstrated for IT project cost and schedule overrun distributions by Budzier and Flyvbjerg (2013), evidentiary for similar generating mechanisms as the rest of the distribution. This implies that black swan probability is similar in several more projects than those few in which they manifest, making individual outliers unpredictable (Sornette, 2009). Some extreme outliers lie beyond power laws, which Sornette (2009) calls *dragon-kings*, relating them to high degrees of coupling that amplify emergent outcomes, causing phase transition. The Thunderhorse offshore platform from our dataset exemplifies this. The interaction between a single valve installed backward—a small design error in the hydraulic system—and a hurricane just before

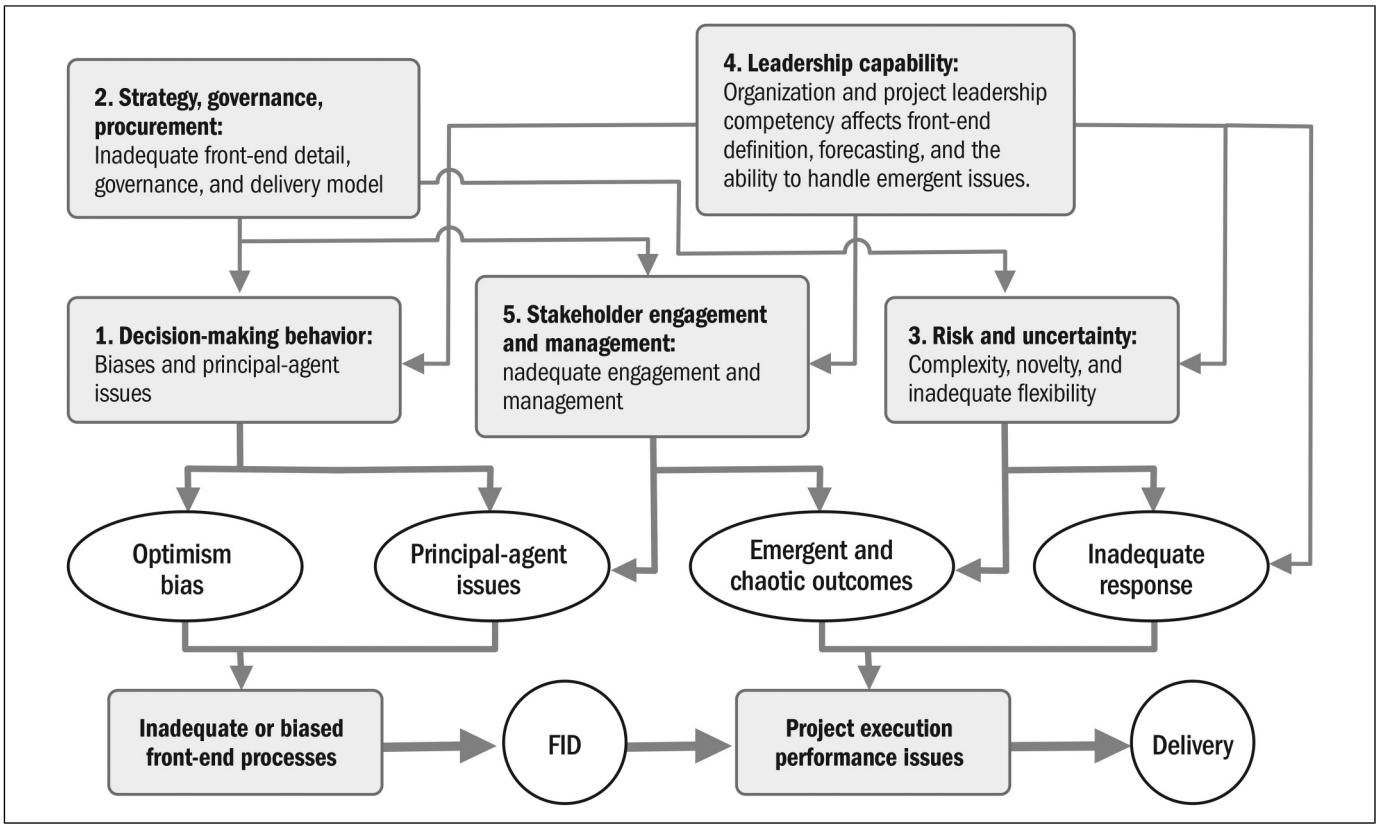

Figure 1. Concept mapping to performance factors.

delivery almost sank it, resulting in catastrophic cost and schedule overruns (Wright, 2009).

Responsiveness

Control-based approaches are inadequate for complex, interconnected projects (Remington & Pollack, 2008). Ackoff (1981, p. 22) saw complex situations as "messes," limiting definite solutions. Emergent changes and black swans require flexibility and responsiveness. Complex projects are often characterized as complex adaptive systems (CAS) (Whyte, 2016).

Heuristics and Biases

Megaprojects are complex systems characterized by uncertain, contested information from several directions (Bruijn & Leijten, 2007). Heuristics can be effective and necessary for forecasting in such environments (Gigerenzer & Brighton, 2011). However, human rationality is limited by cognitive constraints, available information, and time (Simon, 1956), famously labeled *bounded rationality* by Herbert Simon. Furthermore, Kahneman and Tversky (1979) demonstrated how *inside view* forecasting using heuristics can be subjective and biased (Kahneman, 2012). This caused paradigm shifts in several fields, including megaproject management (Flyvbjerg et al., 2018) and was rigorously replicated by an international team of researchers recently (Ruggeri et al., 2020). We interpret bounded rationality in the context of complex megaprojects in Figure 2.

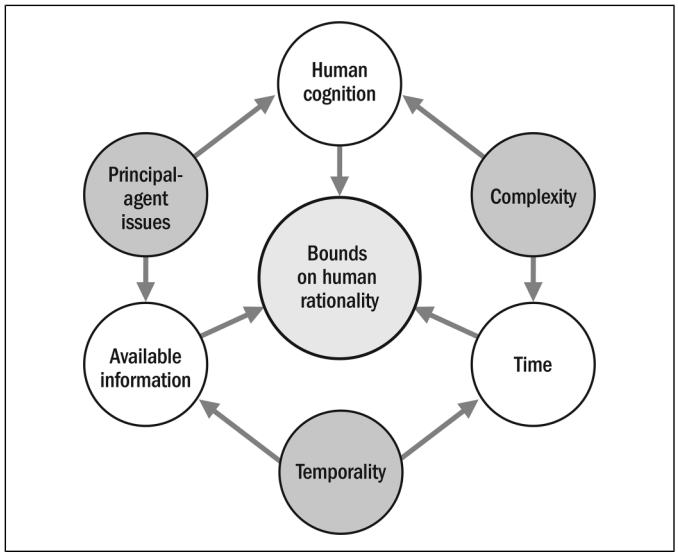

Figure 2. Bounded rationality in complex projects.

Human forecasting exhibits inconsistencies and inadequate consideration of predictability or prior probability (Tversky & Kahneman, 1974). Interpretation of distributional information

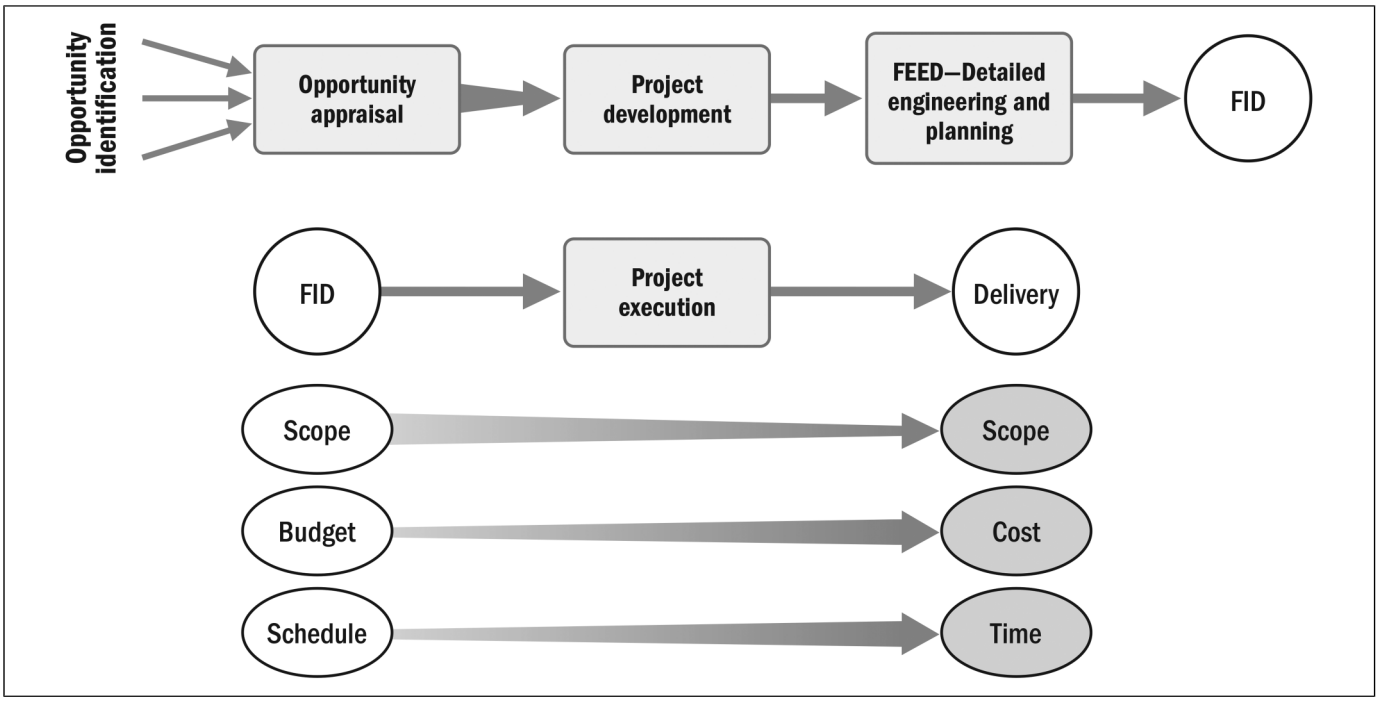

Figure 3. Front-end approach to FID (based on Merrow, 2011, p. 24) and performance deviation measurement.

is prone to biases such as affirmation bias, hindsight bias, and the narrative fallacy of fitting events into simplified, incorrect narratives (Kahneman, 2012). Optimism bias in project forecasting was dubbed the "planning fallacy" and related to the tendency to neglect distributional data by Kahneman and Tversky (1979).

Megaprojects are interorganizational (Sydow & Braun, 2018) temporary meta-organizations of contractually related independent firms with misaligned incentives, asymmetric accountability, and power (Clegg et al., 2017; Lundrigan et al., 2015), challenged by stakeholder conflicts (Locatelli et al., 2014). EPC project complexity is managed by decomposition into subprojects delivered by specialist contractors, making interfaces critical (Davies & Mackenzie, 2014). Typically, O&G companies use EPC contractors to deliver offshore projects using specialist subcontractors (Lee, 2019). Principal-agent conflict across contractual networks is a key project governance problem (Müller, 2009).

Principal-agent issues, optimism, and behavioral biases often result in significant "inside view" underestimations (Flyvbjerg et al., 2018). The causes have been related to "deception" or "strategic misrepresentation" due to misaligned incentives and "delusion" from cognitive biases (Flyvbjerg et al., 2009). Inadequate accountability and risk-sharing mechanisms incentivize cost underestimation and benefit exaggeration (Flyvbjerg, 2014).

Forecasting

Complex projects require decomposition and definition for stability (Remington & Pollack, 2008). O&G projects typically follow a phase-gate approach; a front-end process develops budget and schedule baselines before project approval at final investment decision (FID). A three-stage process (see Figure 3), where estimates are progressively elaborated, is discussed in Merrow (2011) and was present in several projects we investigated. Cost and time overrun measurements require consistent baselines (Flyvbjerg et al., 2018). In our study, the FID budget and schedule define the "budget at the time of decision to build" discussed by Flyvbjerg et al. (2018, p. 175). Overruns are measured between this baseline and actual cost and time at delivery. Scope changes that materially affected projects were baselined to the FID budget and schedule.

Overrun risk is a key FID input. Probability quantifies uncertainty (Goodfellow et al., 2016). Hubbard (2009) defines "strict uncertainty" as possible outcomes with unknown probabilities and risk as the probability of undesirable outcomes. While complexity engendered unknowns limit forecastability, forecasting reduces risk and improves with distributional data. Lovallo and Kahneman's (2003) "outside view" approach to planning was adapted as RCF by using past project distributions to correct biased inside view forecasts for acceptable risk. Batselier and Vanhoucke (2016) demonstrated that RCF outperformed earned value management (EVM) and Monte Carlo simulation for both time and cost forecasting.

Machine Learning

Machine learning (ML) has revolutionized forecasting in several fields. Data-driven project planning using ML is a key recommendation in the Ernst & Young (EY) report on O&G megaprojects (Ernst & Young [EY], 2015). ML models learn by identifying and extracting patterns from data rather than

having knowledge built into them (Goodfellow et al., 2016). A formal definition of ML from Mitchell (1997, p. 2) says: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

In our case, experience E is represented by the reference class distributional data, tasks T are cost and schedule uplift prediction, and the performance P is measured by deviation from actual outcomes. A trained ML model can predict project-specific uplifts from features indicative of biases, emergence, and response by learning correlations using project overrun distributions. This is fundamentally a realist approach to integrate *inside view* project information with the *outside view* using past project data. However, megaprojects are susceptible to high-impact emergence from complexity and length, and the "curse of dimensionality," (Bellman, 2013, p. xxi) a phrase describing the rapid growth in problem difficulty with increasing features. There are a multitude of features with complex interrelationships that affect performance. Their limited number limits datasets. These factors make ML megaproject forecasting extremely challenging.

Research Hypotheses/Questions

In this section we develop the research hypotheses, which are summarized in Figure 4 at the end of the section. RCF improves forecasting accuracy by correcting underestimated forecasts using empirical benchmarking distributions from similar projects that constitute the reference class. However, industrial megaprojects, including O&G projects, are usually privately or market funded (Merrow, 2011) and managed by industry professionals. Merrow (2011) discusses how their tangible, measurable goals and profit motives result in incentivization structures that differ from public infrastructure projects, making them less prone to optimism bias and principal-agent issues. This leads to our first two hypotheses:

Hypothesis 1: Biases, such as optimism and availability biases, are present in heuristics used by offshore industry project managers during forecasting.

Hypothesis 2: Principal-agent misalignment affects offshore project cost and schedule forecasting.

RCF is predicated on the track record of a class of projects showing statistically significant overruns, thus we hypothesize:

Hypothesis 3: Asymptotic distributions of cost and schedule performance outcomes are fat-tailed, non-normal distributions significantly skewed to overruns.

If both cost and schedule distributions show significant overruns, it is desirable to obtain both uplifts. However, if they are not sufficiently covariant, conventional RCF uplifts for both can be overly conservative. A project may have a higher propensity to cost or schedule overrun, affecting the uplift chosen for each.

RCF uses the entire reference class outcome distribution as a probability distribution applicable to any project in that class. However, O&G project performance has shown correlations with factors such as location, front-end detail, and novelty (Merrow, 2012; Rui et al., 2017). Therefore, the

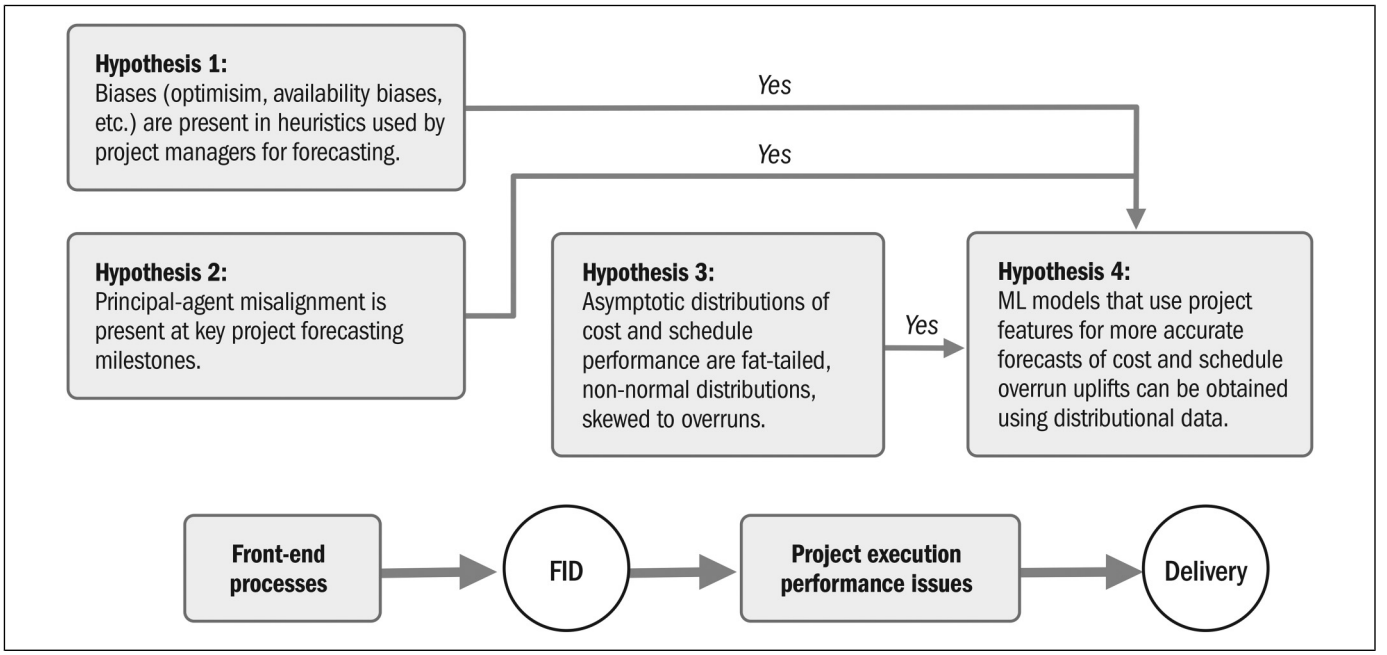

Figure 4. Conceptually linked hypotheses.

ability to identify and use such factors for more accurate project-specific forecasts is of interest. However, factor influences can be modified by other factors in complex ways. For instance, a detailed front-end, which has been correlated with better performance, is characteristic of international O&G companies (IOC), which conversely undertake more challenging projects(Merrow, 2011). Therefore ML, which outputs a model by learning from features and distributions, offers more potential than deterministic forecasting for complex projects, thus we hypothesize:

Hypothesis 4: Generalized ML models can effectively forecast project-specific uplifts for both cost and schedule, using project features by learning from distributional data that are more accurate than conventional RCF.

Research Methods and Data Collection

Research Design

An offshore project is a temporary complex system for creating a complex system whose starting conditions are affected by cognitive biases and principal-agent issues; this limits epistemological questions of finding a definite schedule or budget. Ecological rationality refers to decision-making effectiveness, when bounded rationality is adapted to its environment (Todd & Gigerenzer, 2012). Our goal is to develop effective forecasting methods in the industry context while bound by complexity, available data, and analysis abilities. Distinct research and data collection methods, illustrated in Figure 5, were used for the hypotheses on biases and misalignment on one hand, and hypotheses on project performance on the other. The two hypotheses concerning biases and misalignment were validated using questionnaires, and the two hypotheses on project performance and its prediction were validated from our project dataset and ML model performance.

Research Methods

Decision-Making Biases

A formal judgment elicitation approach using questionnaires answered by experienced offshore project managers validated our hypotheses that heuristic and principal-agent biases challenge offshore project forecasting. The six-step process, based on methodologies for eliciting subjective expert judgment

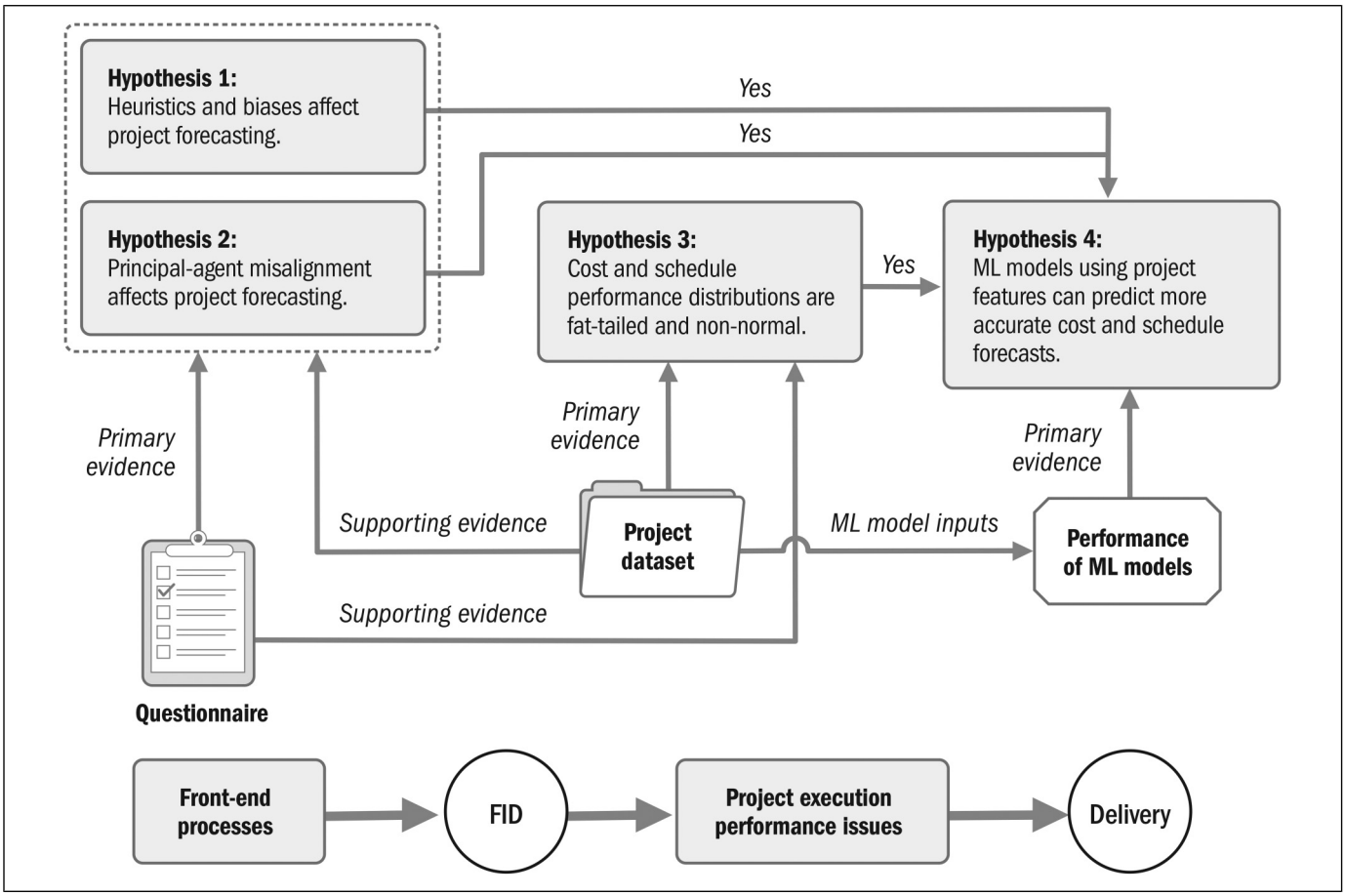

Figure 5. Research methods.

from Walls and Quigley (2001) and Garthwaite et al. (2005) included: preparation, recruitment, briefing, structuring, elicitation, and statistical assessment. The Jisc online survey tool, designed for academic research, was used to prepare the questionnaire, to obtain insights into:

1. Behavioral factors in heuristics-based judgment among project professionals;
2. Differences in expectations and perceptions across contractual boundaries; and
3. Perceived factors that affect project performance; heuristics that have evolved to be ecologically rational in the offshore industry.

Likert scales were used to numerically rank features by their impact on a question. No two features on a scale could have the same rank, and every feature had to be ranked. Thus, the largest rank corresponded to the number of features for that question. Questions to assess subjective probability estimation used percentages instead and were more akin to ratio scales. These are explained further in the analysis sections in context.

Project Performance

Our project performance data collection followed a different approach when compared to the one used for evidence of decision-making biases. It used public and secondary data. Methods for obtaining O&G project data have included pre-collected data from organizations as reported by Merrow (2012) and public and secondary data as reported by Rui et al. (2017) and EY (2014). Our project dataset included project features, FID cost/schedule baselines, and actual cost/schedule performance and was meticulously collected from:

- Annual reports and regulatory filings of O&G companies and contractors;
- Regulatory and government reports, for example, Norwegian Petroleum Directorate, US BOEM;
- Company press releases, Factiva;
- Business information aggregators, such as Capital IQ; and
- Secondary data from reputed industry publications, for example, *Upstream*, *Oil & Gas Journal*.

The FID and installation dates, which involve regulatory approvals and are significant to shareholders, were readily available. Budget and especially cost overrun information required considerable investigation. Discrepancies found in business information sources and secondary data by cross-checking were corrected using corroboratory data or discarded. Information was collated chronologically for each project; the assembled information totaled approximately 400 pages.

Project features affecting performance and forecasting accuracy were identified from:

- Pertinent O&G project performance literature, for example, Merrow (2011 & 2012), EY (2014), Rui et al. (2017), and Steen et al. (2017);
- Cross-case and within-case analysis of our project dataset;
- Theory mapping; and
- Questionnaire feedback.

Features discussed in the literature include front-end development, location, contract management, size, novelty, and company type. Identified features were collected for sampled projects and mapped to our theoretical framework.

RCF and ML application followed the three-step approach described by Flyvbjerg (2006). Offshore projects comprise several categories: fixed or floating production platforms, drilling, or subsea pipelines. Our work emphasized complexity, scope, and measurable FID and delivery milestones. Integrated field development projects with floating platforms fit these criteria well and became our reference class (step 1). Performance measurement was for the aggregate project, as the platform and subsea components are highly interrelated with critical interfaces. O&G companies or their joint ventures (JVs), are referred to as "clients" and primary contractors as "contractors." Performance data and features collected from delivered offshore projects enabled RCF application and training of ML models. Overrun distributions were characterized, probability distributions were established (step 2), and required uplifts were calculated (step 3). Following this, ML models were selected, tested, and deployed for overrun prediction. These three steps are illustrated in Figure 6. The selection of industry-specific features and their weighting to project outcomes make this resemble the robust heuristics with environmental correspondence as discussed by Todd and Gigerenzer (2000).

Data Collection

Questionnaire

Offshore O&G projects rely on expert judgment (Gyasi, 2017). Our questionnaire respondents were 26 individually recruited O&G project professionals, evenly divided between clients and contractors, 25 of whom were currently in offshore projects. This is a good size for expert judgment elicitation in O&G projects from qualitative and quantitative perspectives (Gyasi, 2017).

Respondents were briefed and advised to apply judgment and not refer to data. They represented 624 years of experience, ranging from 10 to 40 years as shown in Table 1. Several had engineering backgrounds and were well acquainted with the quantitative expression of project forecasts. One each from the contractor and client was picked for resembling what Flyvbjerg calls "master builders" (MacNicol, 2016), possessing consistent track records of successful offshore project delivery. The contractor representative had successfully delivered and rescued several challenging projects; the client representative had delivered a complex project in West Africa on time and on budget.

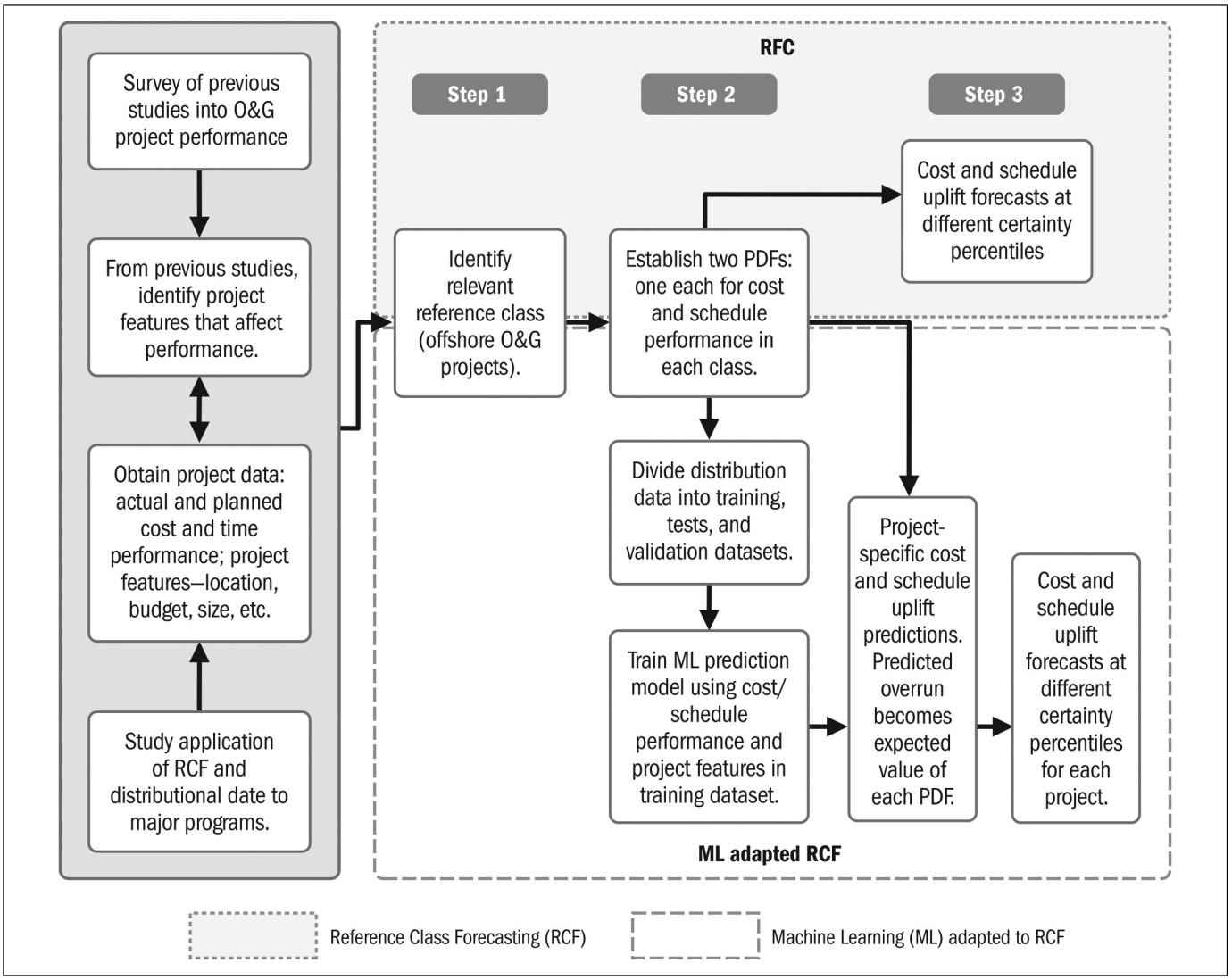

Figure 6. ML and RCF.

Projects

Data collection started with project identification and cross-referencing using industry sources such as the Energy Maritime Associates Floater Systems Report (EMA, 2020) and the 2019 Worldwide FPSO Survey (*Offshore Magazine*, 2019). Approximately 500 projects were found with installation dates from 1990 onward. Subsea tiebacks, where a subsea field is connected to existing platforms, and major upgrades or redeployment of existing platforms were filtered out, leaving 358 projects representing the global population as shown in Table 2. Offshore platforms were often on the critical path and constituted significant portions of aggregate project costs. Schedule overrun information was obtained for 130 projects (~36% of population), of which cost overrun information was obtained for 106 (~30% of population). Our samples represented the population's geographical spread ($p = .71$ and .68). Assessments and predictions involving cost or both cost and schedule used the sample of 106; those involving only schedule used 130 projects.

Table 1. Questionnaire Respondents

		Total	Client	Contractor
Sample Size		26	13	13
Experience	Sum (years)	624	317	307
	Average (years)	24.00	24.38	23.62
	p value, two-sample *t* test			.8
Number of projects	Average	11.27	10.46	12.08
	p value, two-sample *t* test			.6

Access to data on O&G megaproject performance used in EY (2014) was obtained toward the end of our research;

Table 2 Project Dataset

	Population		Cost Sample		Schedule Sample	
Region	Number	Percentage	Number	Percentage	Number	Percentage
Africa	66	18%	24	23%	27	21%
Southeast Asia	68	19%	10	9%	10	8%
Gulf of Mexico	57	16%	26	25%	28	22%
Australia	12	3%	6	6%	6	5%
Canada	2	1%	2	2%	2	2%
Brazil	66	18%	14	13%	32	25%
North Sea	45	13%	21	20%	21	16%
China	18	5%	3	3%	4	3%
Mediterranean	8	2%				
South America	4	1%				
Middle East	12	3%				
Total	358		106		130	

discussions on data collection with the EY research team were helpful. Offshore projects were only part of their sample. The few projects that fit our filtering criteria were cross-checked against our data on those projects and helped validate the robustness of our data collection.

Projects were coded as a design matrix conceptually described in Goodfellow et al. (2016). Each row corresponded to a different project and each column to a project feature. Eighteen features related to project performance and known at FID were collected for each project (see Table 3). Features such as size, novelty, lessons learned, and front-end detail were ranked along a five-point scale using qualitative information and quantitative inputs from within-case analysis. Features such as company type (International [IOC] or National Oil Company [NOC]), contract types, region, and so forth, were categorical data converted to dummy variables for ML algorithms. The average planned FID budget was US$2.3 billion, the median was US$1.3 billion, and the total was US$244 billion for 106 projects (not adjusted for inflation). The average actual cost was US$3.1 billion. Some projects were below the threshold of US$1 billion, which is often associated with megaprojects (Merrow, 2011). The average planned duration was 3.2 years, median was three years, and average actual duration was about four years for 130 projects.

Table 3. Project Features for ML

Features	Measure
Region	Categorical
Company	Categorical
Lease/Own	Categorical
Contract_type	Categorical
Contract_risk_allocation	Categorical
Unit_Type	Categorical
Conv/New	Categorical
Contracting_Date	Date
Planned_Duration	days
Planned_Cost	billion USD
BOE/day	BOE/day
Local_Content_Requirement	5-point scale
Topsides_size	5-point scale
Technology_Novelty	5-point scale
Lessons_Learned	5-point scale
FEED-Detail	5-point scale
Water-Depth	meters
Oil/Gas_Prod	Ratio

Research Findings and Discussion

Data Analysis and Discussion—Questionnaire on Decision-Making Biases

Industry project managers were asked to estimate the proportion of projects facing cost and schedule overruns and also to estimate those overruns. These expert estimates were for understanding heuristics in forecasting helped by comparison with actual performance from our dataset, not for project data. Respondents were asked to rank project features to understand underperformance and outlier causation attribution. Rankings also provided insights into differing perceptions between clients and contractors indicative of principal-agent issues. Two-tailed, two sample *t*-tests were used to compare and discover differences between rankings by client and contractor groups. Paired *t*-tests were used to compare feature rankings within a scale to discover their perceived significance. Statistical significance levels were .05 or less. The results clearly showed errors from optimism, availability and representativeness biases, conjunction fallacies, and principal-agent issues in heuristics used by industry experts, validating Hypothesis 1 and Hypothesis 2. Project feature rankings displayed high standard deviations, correlated mostly to principal-agent boundaries, but also to experience. These results also provided feedback on project feature selection.

Perceived Budget and Schedule Overruns

Respondents were asked to recollect the proportion of their projects that were on budget and on schedule. The low estimate of ~50% (see Table 4) across contractual boundaries can be

Table 4. Recollected Proportions of Projects with Budget and Schedule Overruns

Percentage of Your Projects that you Believe to Have Been		Total	Client	Contractor
On budget	*Mean*	48%	49%	48%
	SD	24%	23%	26%
	Median	45%	45%	35%
	SE	.93%	1.79%	2.00%
	p value_two-sample			.94
On schedule	*Mean*	51%	48%	53%
	SD	27%	28%	27%
	Median	50%	45%	55%
	SE	1.03%	2.14%	2.06%
	p value_two-sample			.62

related to low predictability assessments by experts in their field. This test is recommended by Kahneman and Tversky (1979) for the applicability of reference class corrections. The high underperformance estimate, coming from project professionals, is indicative of principal-agent issues. Furthermore, this high estimate is still less than that from our data (18% on schedule, 20% on budget, 31% within 5% of schedule, 28% within 5% of budget), indicating optimism bias along with principal-agent issues. There was no significant difference between client and contractor perceptions; however, the contractor estimated budget overrun was bimodal, possibly from availability bias reflecting the most recent project.

Respondents were also asked to estimate the actual overruns (see Table 5). Their estimates were similar but significantly less than the average cost overrun of + 33% (n = 106) and average schedule overrun of + 26% (n = 130) from our dataset.

Optimism Bias and Heuristics

Participants were asked to estimate typical industry project performance in two different ways. The first used a ratio scale, Scale 1, with equidistant percentage measures. Given the limitations to objective probability estimation in humans, a more relatable nonlinear canon of probabilities for judgment elicitation, found in a work on the application of Bayes theorem to historical questions (Carrier, 2012), was used as Scale 2 for the second means of elicitation. The two scales are illustrated in Figure 7.

Scale 1	Scale 2
0%–10%	Virtually certain (99.9999%)
11%–20%	Extremely probable (99%)
21%–30%	Very probable (95%)
31%–40%	Probable (80%)
41%–50%	Slightly probable (60%)
51%–60%	Even odds (50%)
61%–70%	Slightly improbable (40%)
71%–80%	Improbable (20%)
81%–90%	Very improbable (5%)
91%–100%	Extremely improbable (1%)
	Virtually impossible (0.0001%)

Figure 7. Same questions, different scales.

Estimates of industry project performance using Scale 1, as shown in Table 6, were only marginally different from participant projects, but were significantly different using Scale 2: (p = .0033 [cost]; .0043 [schedule]). Performance estimates using Scale 2 were much closer to the data: (27% vs. 20% [cost]; 29% vs. 18% [schedule]).

The significant number of actual underperforming projects versus estimations is evidentiary for forecasting bias. The much better performance of Scale 2 indicates the need for more relatable measures for judgment elicitation to avoid bias, even from experts.

Table 5. Recollected Estimates of Budget and Schedule Overruns

Estimated Average Overrun by Respondents in their Projects		Total	Client	Contractor
Cost overrun	*Mean*	17%	19%	15%
	SD	.13	.11	.15
	Median	15%	25%	15%
	SE	.50%	.86%	1.14%
	p value_two-sample			.46
Schedule overrun	*Mean*	20%	20%	20%
	SD	.16	.17	.16
	Median	15%	15%	15%
	SE	.63%	1.31%	1.24%
	p value_two-sample			.95

Table 6. Likelihood Estimations for Project Performing to FID Baselines

	Respondent's Own Projects		Typical Industry Project			
			Scale 1		Scale 2	
	On budget	On schedule	On budget	On schedule	On budget	On schedule
Mean	48%	51%	50%	55%	27%	29%
SD	24%	27%	.24	.25	.26	.25
p value: Own versus industry project			.82	.52	**.0033**	**.0043**

Conjunction Fallacy

Tversky and Kahneman's (1983, p. 293) famous paper on the conjunction fallacy in probability judgment observes that "the probability of a conjunction, P(A&B), cannot exceed the probabilities of its constituents, P(A) and P(B)," is perhaps the "simplest and most basic quantitative law of probability."

To test for this fallacy, respondents were asked to rank the likelihood of an offshore project suffering from cost OR schedule overruns using Scale 2. The mean result was 78%, corresponding to a likelihood of meeting both targets of 22% for P(A&B), only marginally less than their estimates for cost, P(A) = 27%), and schedule, P(B) = 29%. The percentage of projects from our dataset that actually met both cost and time targets is only 4%, which is considerably less than the percentage meeting each performance target. This is indicative of the conjunction fallacy.

Availability Bias

Availability bias—where the probability assigned to an event is biased by the ease of recall of similar instances—was tested by asking participants to choose a 0%–100% probability for an unforeseen event, such as a global pandemic or a catastrophic economic downturn to affect an offshore project materially. This evoked the COVID-19 pandemic, which had caused a severe crisis in the O&G industry at the time of the research.

The mean probability of 42% was much higher than expected and double the mean estimates for outliers caused by external events in a separate question to the same respondents. Projects from our sample showing evidence of being materially affected by such events were less than 10%. This is strong evidence for availability bias. Interestingly, the client and contractor "master builders" assigned 20% and 10% probabilities, respectively, considerably less than the mean assignments of 40% and 50%, respectively, in their groups.

Representativeness Bias

We replicated Tversky and Kahneman's (1983) famous test of the proposition that coupling an outcome with a cause would make it appear more probable than the outcome on its own, whereas the opposite is true. Respondents were asked to assess the probabilities of two events:

1. Cost overrun of more than 40% in a project offshore Southeast Asia.
2. Regulatory and local content issues causing cost overrun of more than 40% in a project offshore West Africa.

The mean estimate for the conjunction overrun due to regulatory and local content issues in West Africa, was higher at 47% versus 35% for the Southeast Asia project. Average cost overruns in Southeast Asia (26%) and Africa (24%) from our sample are very similar. Local content issues are significant in both regions, which form strong evidence for representativeness bias.

Interestingly, the two master builders stood out for correctly ranking the probability of the overrun in Southeast Asia significantly higher than the estimate for the conjunction: 45% versus 15% (client) and 25% versus 5% (contractor).

Table 7. Cost and Schedule Performance Factors

	Average	SD	SE
JV partner management	4.46	2.56	.10
Overoptimistic FID forecasts	6.15	3.18	.62
Market and geopolitical	6.12	3.05	.60
Local content requirements	7.00	2.90	.57
Project cost and size	5.50	3.34	.65
Geographical location	7.00	3.21	.63
Contract issues	5.27	2.85	.56
Technology novelty	5.88	2.49	.49
Insufficient front-end	4.73	3.39	.66
Own over lease	7.04	3.75	.73
Prescriptive over functional requirements	6.92	3.02	.59

Perceived Factors Affecting Cost and Schedule Performance

Respondents were asked to rank factors by importance for meeting cost and schedule FID baselines as shown in Table 7.

No factor stood out, and responses had high variance. JV management issues and insufficient front-end detail, correlated with project underperformance in some studies, were ranked least significant ($p < .05$). These issues made it challenging to filter project features using corroboration from the questionnaire.

Contractors gave significantly more causality to contractual issues ($p = .013$) and high but not statistically significant causality to overoptimistic forecasts at FID ($p = .063$). Both are indicative of principal-agent issues.

Outlier Causation Perceptions

Budzier and Flyvbjerg (2013) studied the impact of outliers in project management and discussed three schools of thought to explain them. The system-centric view focuses on project complexity, the event-centric view focuses on external events coupled with ineffective response, and the process-centric view focuses on the buildup of issues over long periods. Features identified from our theoretical framework and project case analyses were mapped across these views. Respondents were asked to judge which features contributed most to cost overruns greater than 50% (see Table 8).

The normalized attribution was: 53% (system-centric)**:** 20% (event-centric): and 27% (process-centric). Clients ranked local content requirements significantly higher (mean-ranking: 8.15 vs. 5.08, $p = .011$), as well as prescriptive requirements (mean-ranking: 7.08 vs. 4.58, $p = .043$), even though prescriptive requirements typically ensue from their organizations, pointing to principal-agent issues within sponsoring companies. Contractors ranked overoptimistic FID forecasts more

Table 8. Factors Contributing to Severe Cost Overruns

A project shows a cost overrun > 50% from FID. Rank these factors by how they have may have contributed.		Average	SD	SE
System-centric	Own over lease	6.32	3.79	.76
	Prescriptive over functional requirements	5.88	3.23	.65
	Insufficient front-end	5.48	3.42	.68
	JV management	5.12	2.96	.59
	Overoptimistic FID forecasts	6.20	3.50	.70
	Technological novelty	6.16	2.59	.52
Event-centric	Location	6.60	2.90	.58
	Market and geopolitical issues	6.52	3.43	.69
Process-centric	Local content requirements	6.68	3.20	.64
	Project cost and size	5.44	2.99	.60
	Contract issues	5.36	3.16	.63

significantly (average: 7.67 vs. 4.85, $p = .034$), indicating principal-agent issues.

A follow-up question elicited responses coded to the three views to check consistency, with one cause corresponding to each (Table 9). It was also designed to understand internal consistency; a percentage representing the probability of causation was used and, by implication, the sum of probabilities had to be 100%.

The elicited probabilities at 152% totaled more than 100%, pointing to the "Conjunction Fallacy." The normalized attribution was: 47% (process-centric): 21% (event-centric): 32% (system-centric), reversing the previous attributions, pointing to inaccurate quantitative estimations.

Overrun Causation Perceptions Across Contractual Boundaries

Two scales asked respondents to rank overrun causation factors from client and contractor perspectives to highlight the attribution differences between client and contractor participants and corroborate feature selection for our prediction models. While neither scale showed statistically significant deviations between client and contractor responses, there were several pointers to principal-agent issues: Contractors saw the pressure to show reduced cost and schedule as much higher than clients (average: 5 vs. 3.7, $p = .17$); clients ranked improper subcontractor work not discovered in time higher (average: 4.46 vs. 3.5, $p = .137$).

> The extensive biases and principal-agent misalignment in the questionnaire responses by industry project management practitioners validate Hypothesis 1 and Hypothesis 2.

Table 9. Factors Contributing to Severe Cost and Schedule Overruns

An offshore floater project is running into cost overruns and schedule challenges from FID benchmarks. The probability of causation is:			
	Unforeseen issues from project-complexity (system-centric)	Unexpected outside event: economic, geo-political, pandemic, etc. (event-centric)	Inadequate front-end rigor/ time (process-centric)
Average	49%	32%	71%
SD	25%	22%	21%
SE	.94%	.85%	.82%

Data Analysis and Discussion—Project Performance and Forecasting

Analysis of offshore project cost and schedule performance outcomes found that their asymptotic distributions were fat-tailed with black swans and catastrophic dragon-king outliers, validating Hypothesis 3. Both cost and schedule distributions show significant overruns. We demonstrate the ability of clustering to identify black swans and dragon kings and the application of RCF to offshore project uplifts. We then demonstrate ML models for accurate project-specific overrun forecasting, validating Hypothesis 4.

Cost and schedule performance for offshore O&G projects in our sample are shown in Table 10; average cost overrun is +32.8%, schedule overrun is +25.6%, 82% of 130 projects were late, and 80% of 106 projects were over budget. This is comparable with Merrow's (2011) study, which looked at 318 industrial megaprojects, including 130 O&G projects, of which 78% showed cost overruns of +33% and schedule slippage of 30%. EY (2014) reported an average +23% cost overruns for O&G upstream and downstream megaprojects, with 64% of 205 projects facing cost overruns and 73% of 242 projects reporting delays. This performance compares favorably with average cost overruns of +107.2% and schedule overruns of +37.3% reported for large IT projects by Budzier and Flyvbjerg (2013). However, cost overruns in our dataset amounted to about US$83 billion, and schedule slippage had severe cost implications, indicating the issue's seriousness.

We identified outlier projects using the conventional definition of 1.5 inter-quartile ranges from the IQR boxes (Figure 8), similar to Budzier and Flyvbjerg (2013). This resulted in three schedule overrun outliers (2.3% of sample) and six cost overrun outliers (5.7% of sample). Case analysis revealed nothing materially unique about fat-tail outlier projects, keeping with Budzier and Flyvbjerg's (2013) findings for IT projects.

Table 10. Project Outcomes

	Mean	Median	IQR	n
Cost overrun	+ 32.8%	+ 20%	.43	106
Schedule overrun	+ 25.6%	+ 20.2%	.35	130

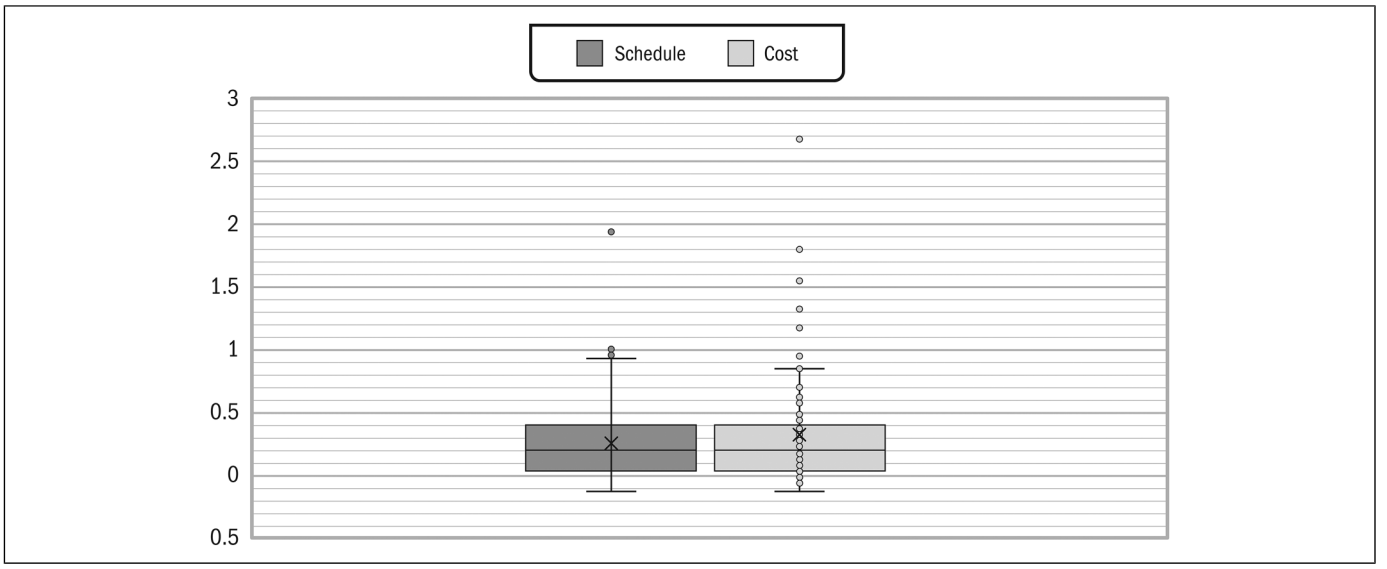

Figure 8. Box plots of overrun data.

Overrun Distribution

Outlier presence and discrepancies between mean and median indicate skewness and fat-tailed distributions (Budzier & Flyvbjerg, 2013). Cost and schedule overrun histograms show fat-tailed non-normal distributions (Figure 9), similar to O&G project cost overruns reported in EY (2014) and Rui et al. (2017). Dragon kings are visible as "obvious bumps in the tail" (Sornette, 2009, p. 5). The schedule, budget, and associated overrun curves in Figure 10 used kernel density estimation (KDE), a non-parametric method to estimate probability density functions (PDF). Only the planned duration is near normal. Cost overrun, which has more outliers and a greater difference between the median and mean, is more fat-tailed. The Kolmogorov-Smirnov (KS) test, a nonparametric test used to test the goodness of fit between probability distributions, was used on the cost and schedule overrun distributions. There was no significant difference ($p = .18$).

Maximum likelihood estimation was used to fit PDFs to overrun distributions. Normal distributions did not fit the schedule ($p = .049$) or cost overrun ($p = .0002$). Exponential, Birnbaum-Saunders (BS), and Pareto distributions were fit to both overruns (Figure 11). The KS test was used to test the null hypothesis that they fit the data ($p > .05$). The Birnbaum-Saunders two-parameter family of extreme-value distributions, used to model structural fatigue and reliability related failures (Birnbaum & Saunders, 1968), provided the

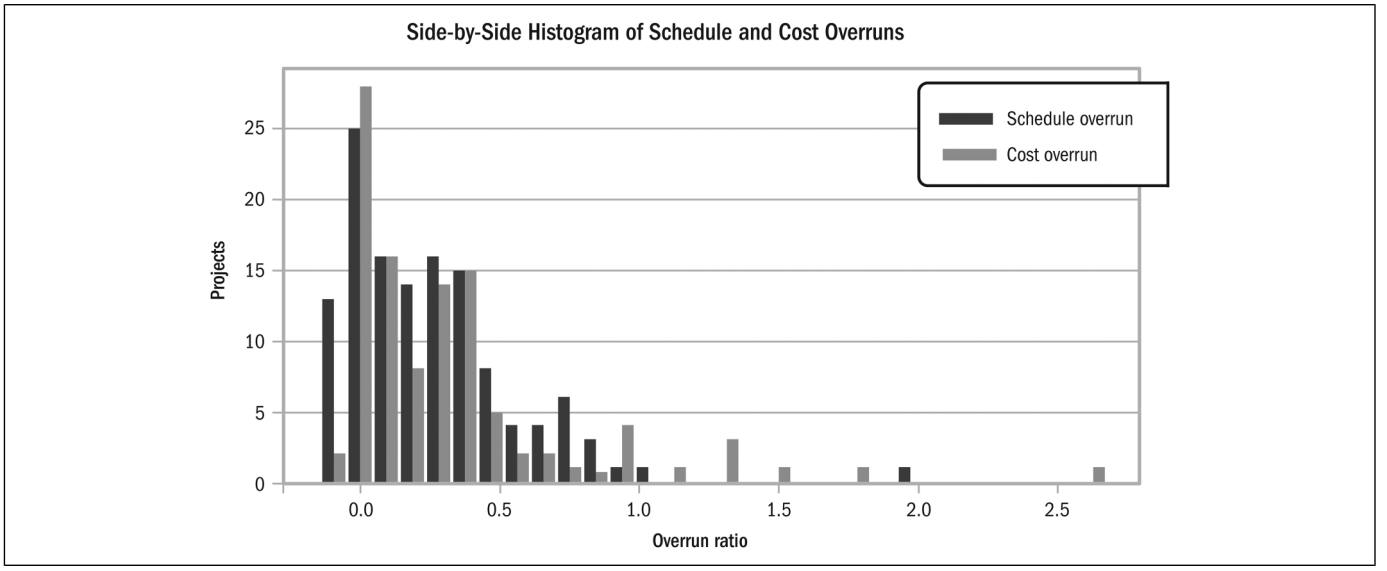

Figure 9. Schedule and cost overrun histograms.

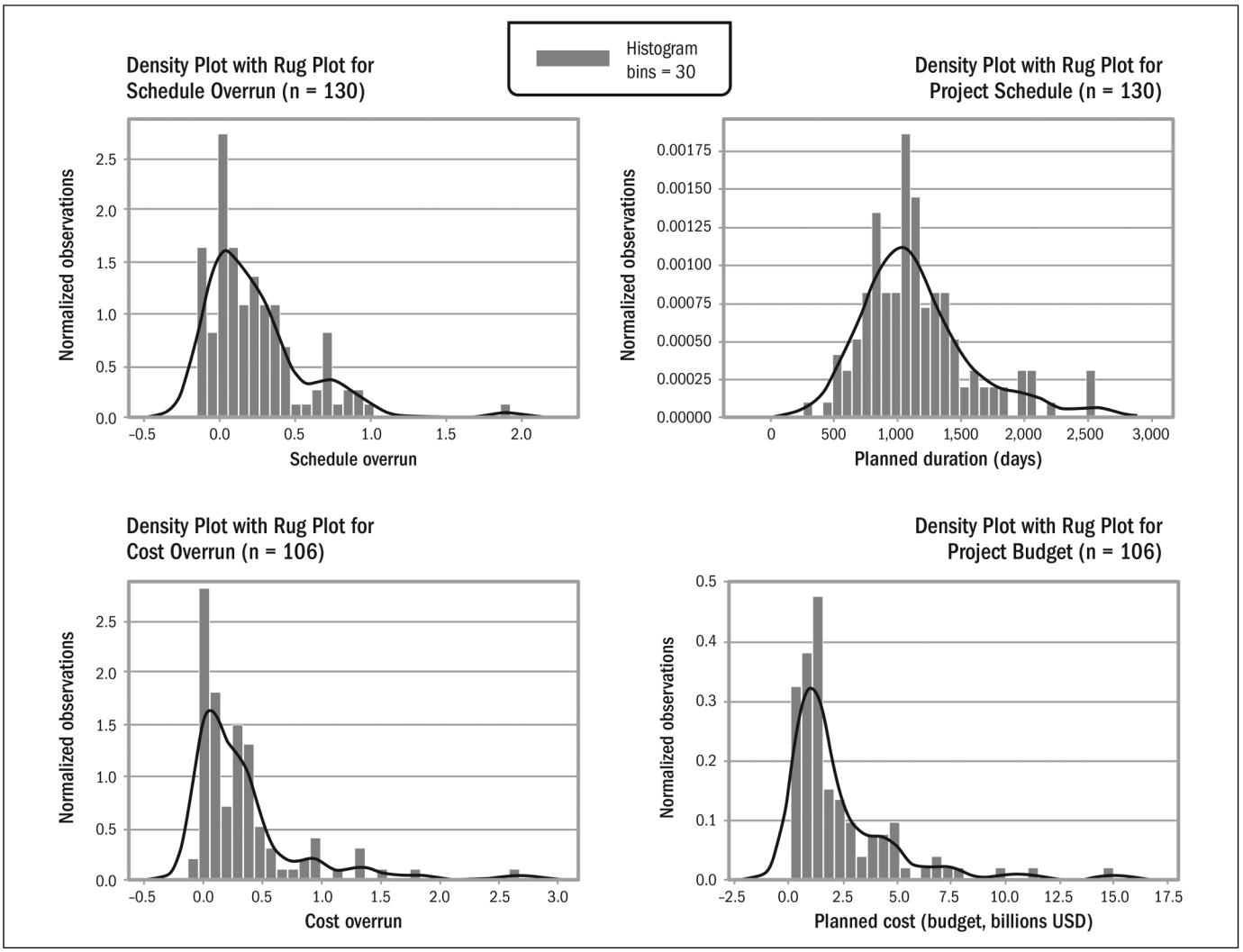

Figure 10. Cost and schedule overruns and corresponding budget and schedule at FID.

best fit to both schedule ($p = .89$), and cost overruns ($p = .14$). Sornette (2009) discusses how the coexistence of power-law distributions with catastrophic dragon-king events can be approximated by calibrating distributions that model material failure, which incorporate positive feedback and phase transition. These results offer possibilities for approximating distributions for forecasting when actual distributions are limited.

Overrun distributions were generated using uniformly distributed pseudorandom numbers and the fitted BS distributions. It was interesting to observe how increasing noise in the BS distribution produced bumps mimicking dragon-king outliers in the tail. As random variables increased from 100 to 10,000, the curve smoothened as it approached the asymptote. The KS two-sample test, showed good fits for schedule ($p = .085$) and cost ($p = .25$), as shown in Table 11 and Figure 12.

Cost and schedule overruns in offshore O&G projects are non-normal, fat-tailed, and skewed toward overruns, validating Hypothesis 3.

RCF Application

After establishing the reference class cost and schedule overrun distributions probability distributions are estimated as their cumulative distributions as shown in Figure 13. Cost overrun is visibly more substantial than schedule overrun.

Functions relating the required uplift on the x-axis to acceptable risk level percentiles on the y-axis for schedule and cost (Figure 14) overruns were obtained from the probability distributions, as demonstrated by Flyvbjerg (2006). Table 12 shows P10, P25, P50, and P75 uplifts from these functions. The uplift percentile choice should follow the desired risk acceptability, as discussed at length by Flyvbjerg (2006).

Joint Distribution of Cost and Schedule Performance

K-means clustering was initially used to identify RCF subclasses in the sample. While the limited sample and population sizes meant that further subdivision was not pursued, clustering analysis yielded valuable information. Project outcomes were located along the two dimensions of cost and schedule

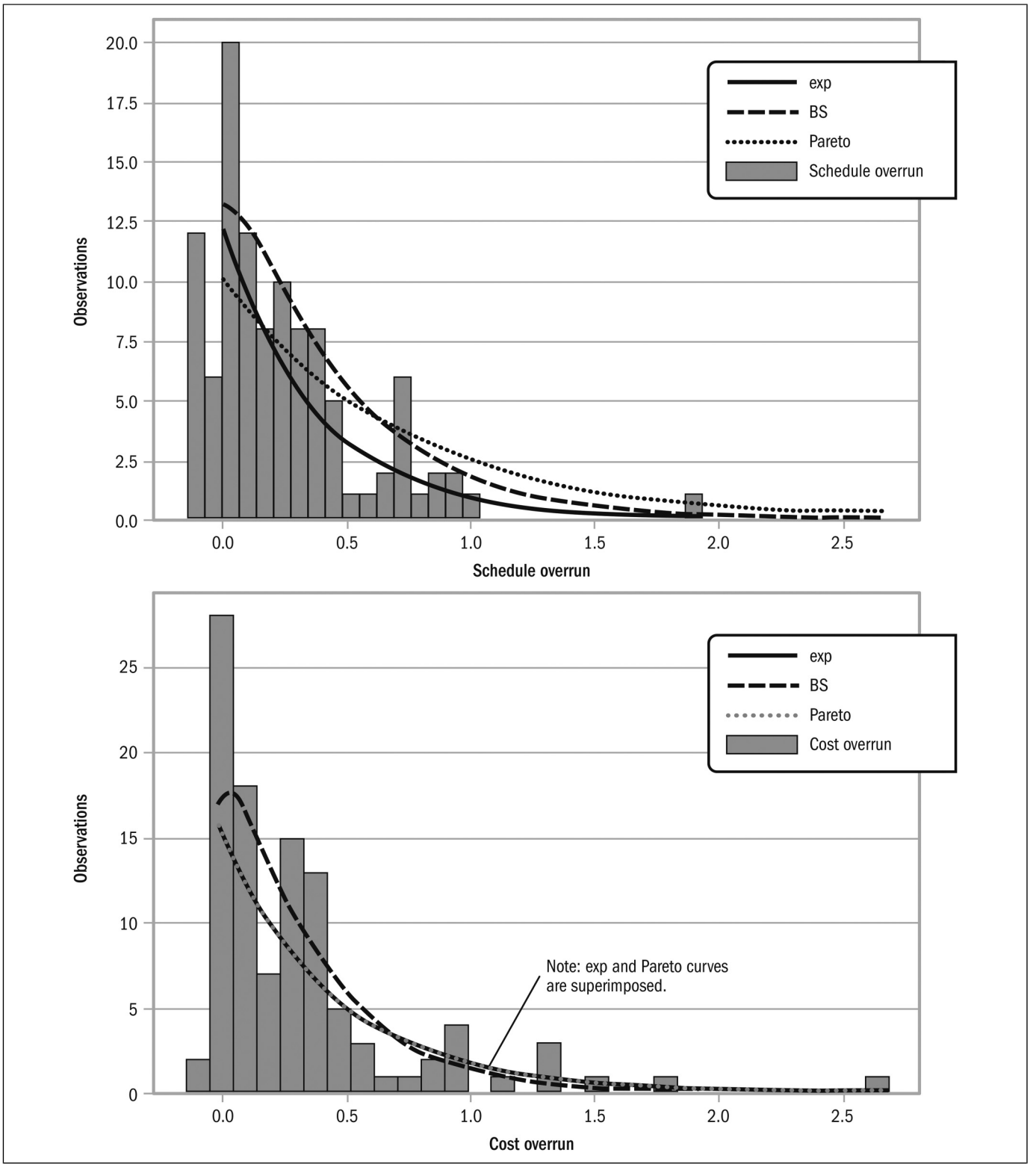

Figure 11. PDF fitting to cost and schedule overrun data.

overruns. Within-cluster sum of squares (WCSS) was used to measure variability within clusters and identify optimal cluster numbers. The WCSS score was plotted against the number of clusters to identify the bend where the variability tapered off (Figure 15). Three clusters offered the greatest separation (Figure 16): the first region with 81 projects corresponds

Table 11. Statistical Parameters, Cost, and Schedule Outcome Generation

	Cost-Overrun-Distribution	Cost-Overrun, Fitted-Curve	Schedule-Overrun- Distribution	Schedule-Overrun, Fitted-Curve
Mean	.326	.258	.24	.29
Variance	.192	.089	.105	.16
Skew	2.53	1.475	1.9	2.56
Kurtosis	8.53	2.685	6.1	11.3

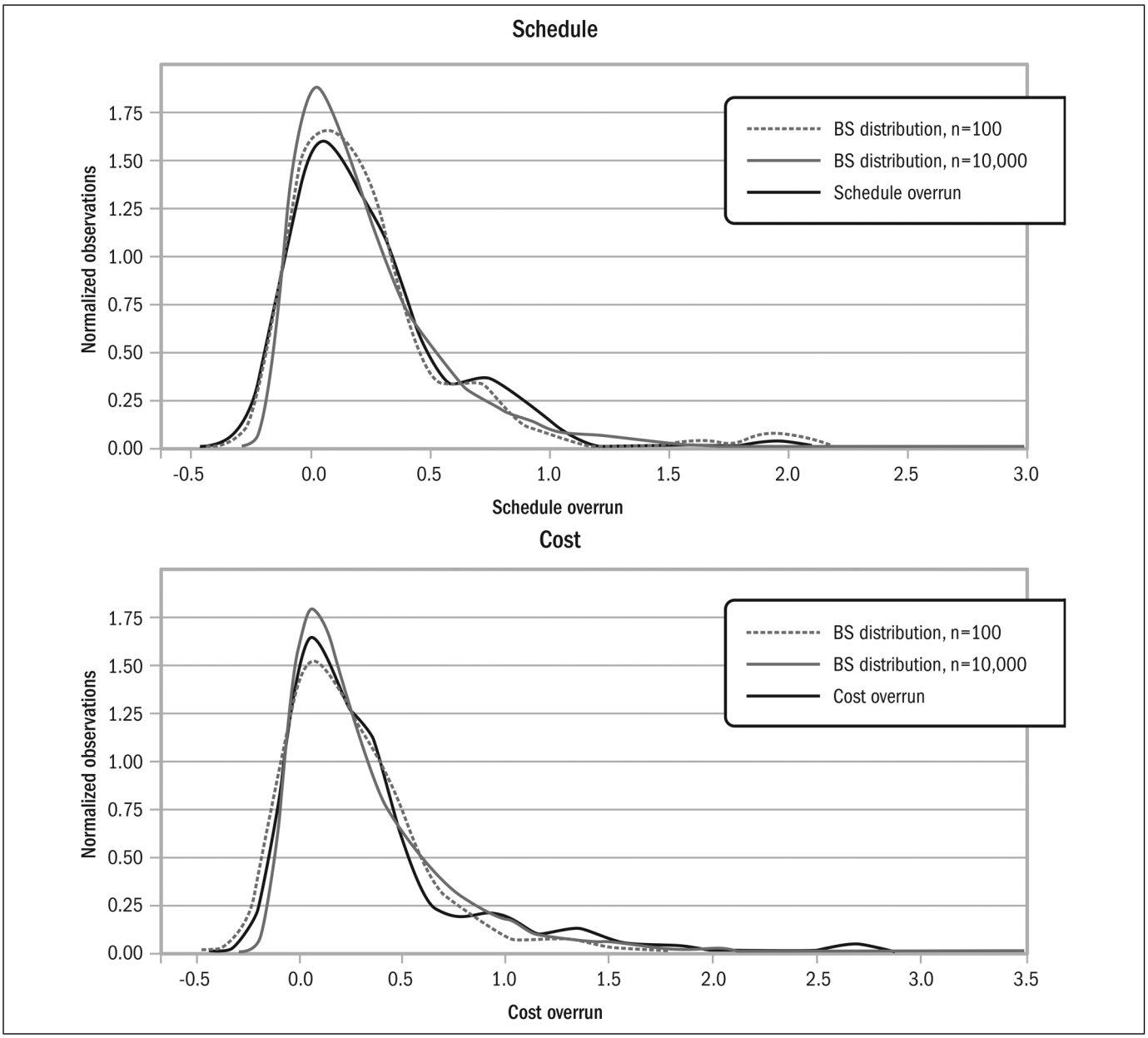

Figure 12. Cost and schedule overrun distribution generation.

to where both overruns are clustered together; the other two clusters divide projects into those showing high cost or schedule overrun. Figure 17 shows how projects separate into distinct cost or schedule overrun clusters as overruns become more extreme. As the number of clusters increases to five and eight, extreme outliers, including black swans and dragon kings, became distinguishable. These findings offer points of departure for the study of outlier occurrence, distribution, causation, and clustering for reference class subdivision beyond this article's scope.

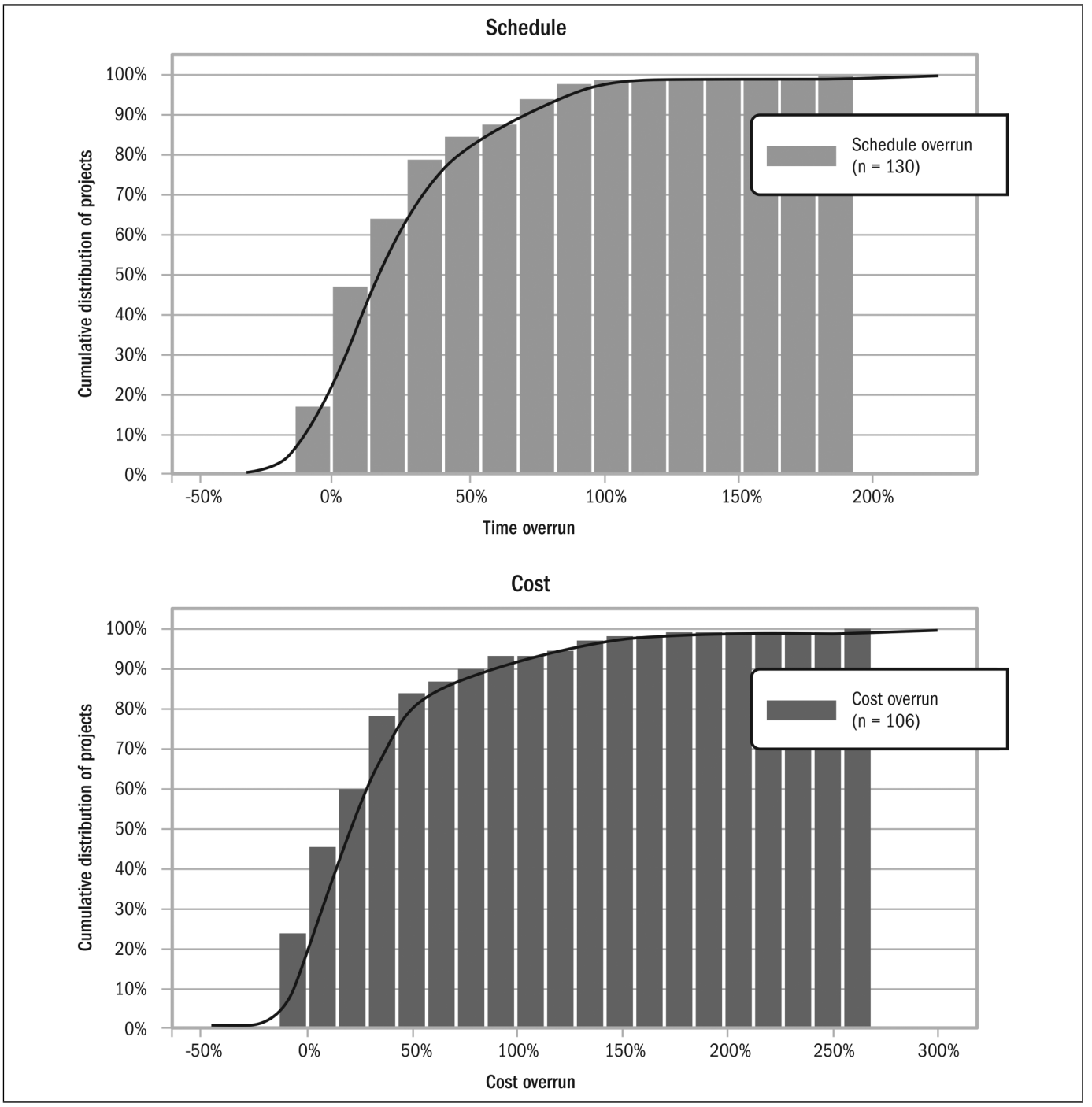

Figure 13. Probability distributions of schedule/cost overruns.

Data-Analytics/ML—Performance Prediction

Cost and Schedule Uplift Forecasting

The distinct populations of cost and outliers reveal limited covariance between cost and schedule overruns. The Wilcoxon signed-rank test, a nonparametric version of the paired *t*-test, only narrowly failed to determine that cost and schedule overruns were significantly different ($p = .08$). A linear regression model fit to a joint distribution of schedule overrun and cost overrun in Figure 18 displays high scatter (PCC = .37, $p < .01$). Even the dense cluster of 81 projects identified by three-means clustering shows substantial scatter (PCC = .25, $p < .05$). Similar uplift percentages for cost and schedule will represent different risk percentiles for different projects,

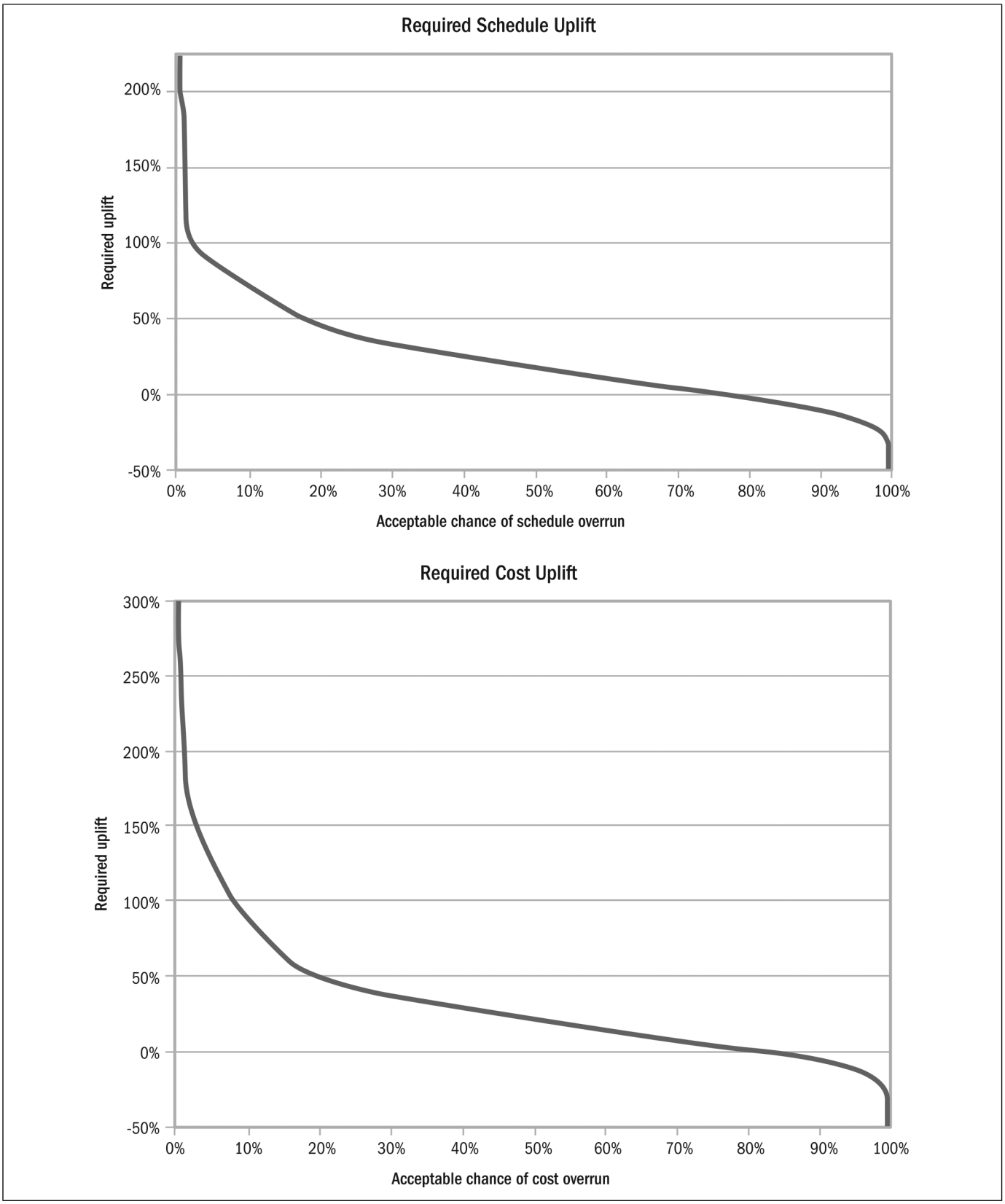

Figure 14. Required uplifts as functions of acceptable risk.

Table 12. RCF Uplifts

Acceptable Risk	Schedule Uplift	Cost Uplift
10%	72%	89%
25%	38%	43%
50%	17%	21%
75%	1%	5%

making RCF application to both challenging. This was resolved by using ML to forecast project-specific cost and schedule uplifts.

ML Cost/Schedule Uplift Forecasting

Deep learning utilizes models that train from a dataset using an optimization procedure and a cost function (Goodfellow

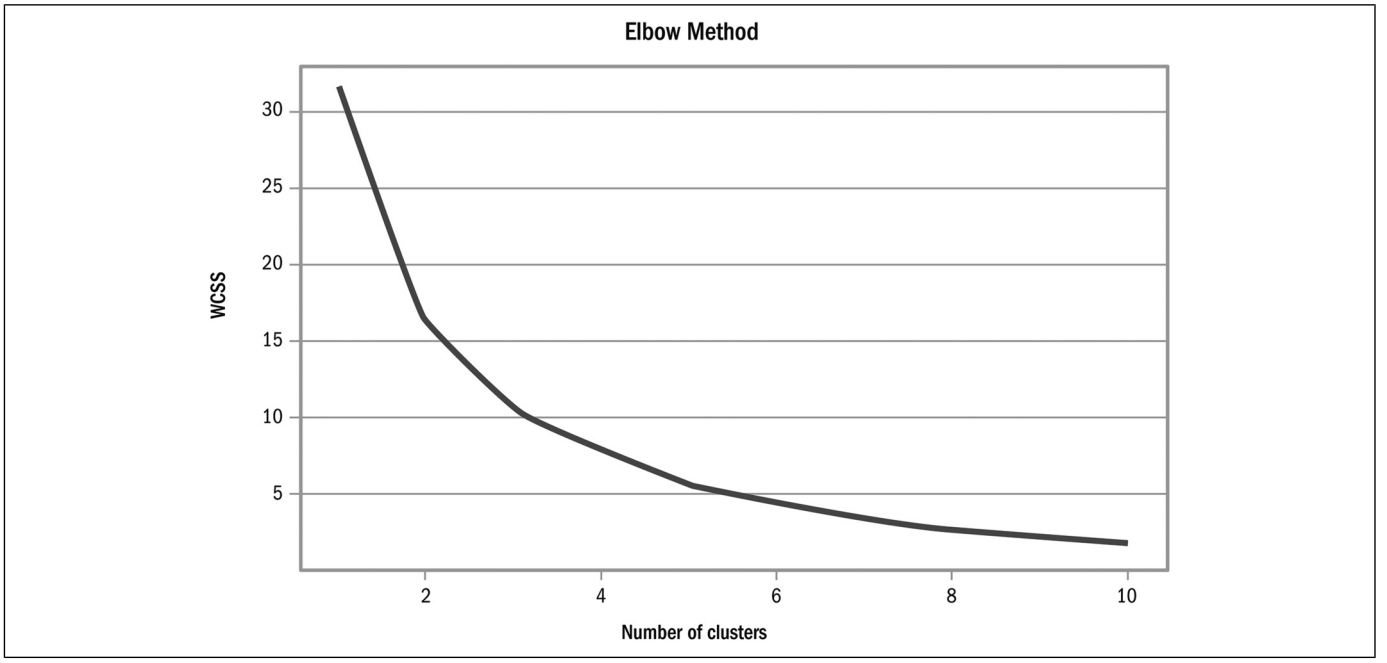

Figure 15. WCSS scores for clusters in cost-schedule distribution.

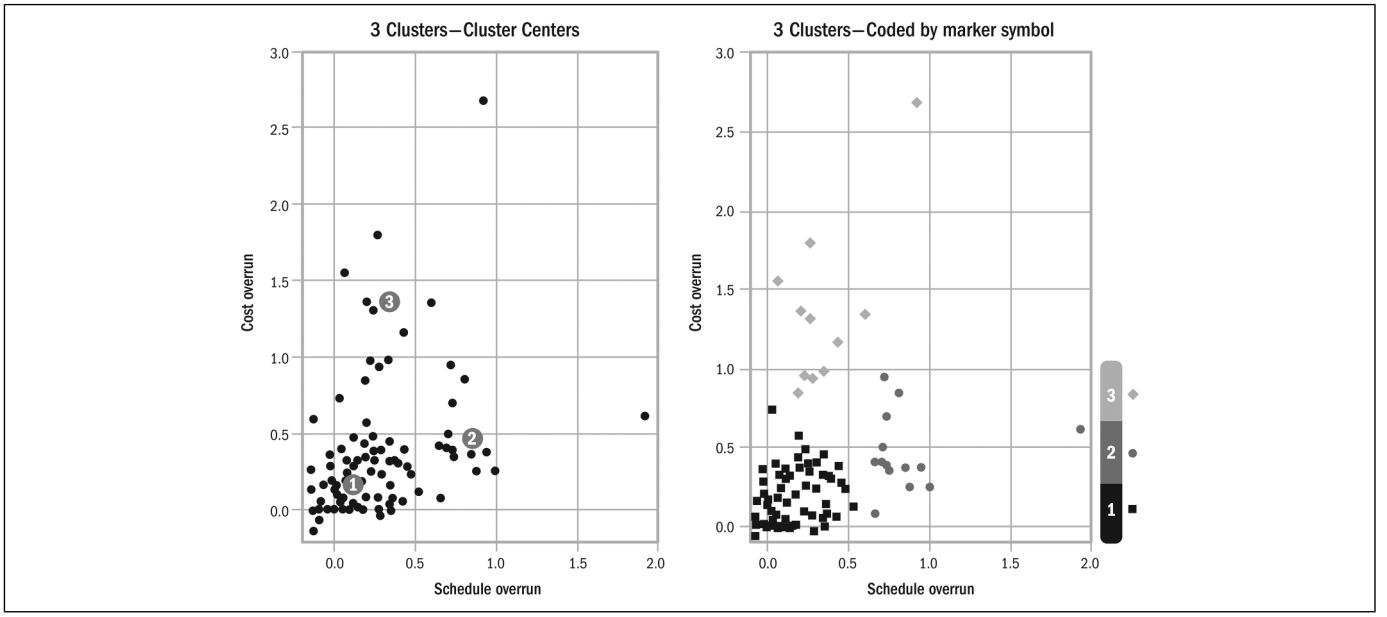

Figure 16. Clustering of project cost and schedule outcomes—three clusters.

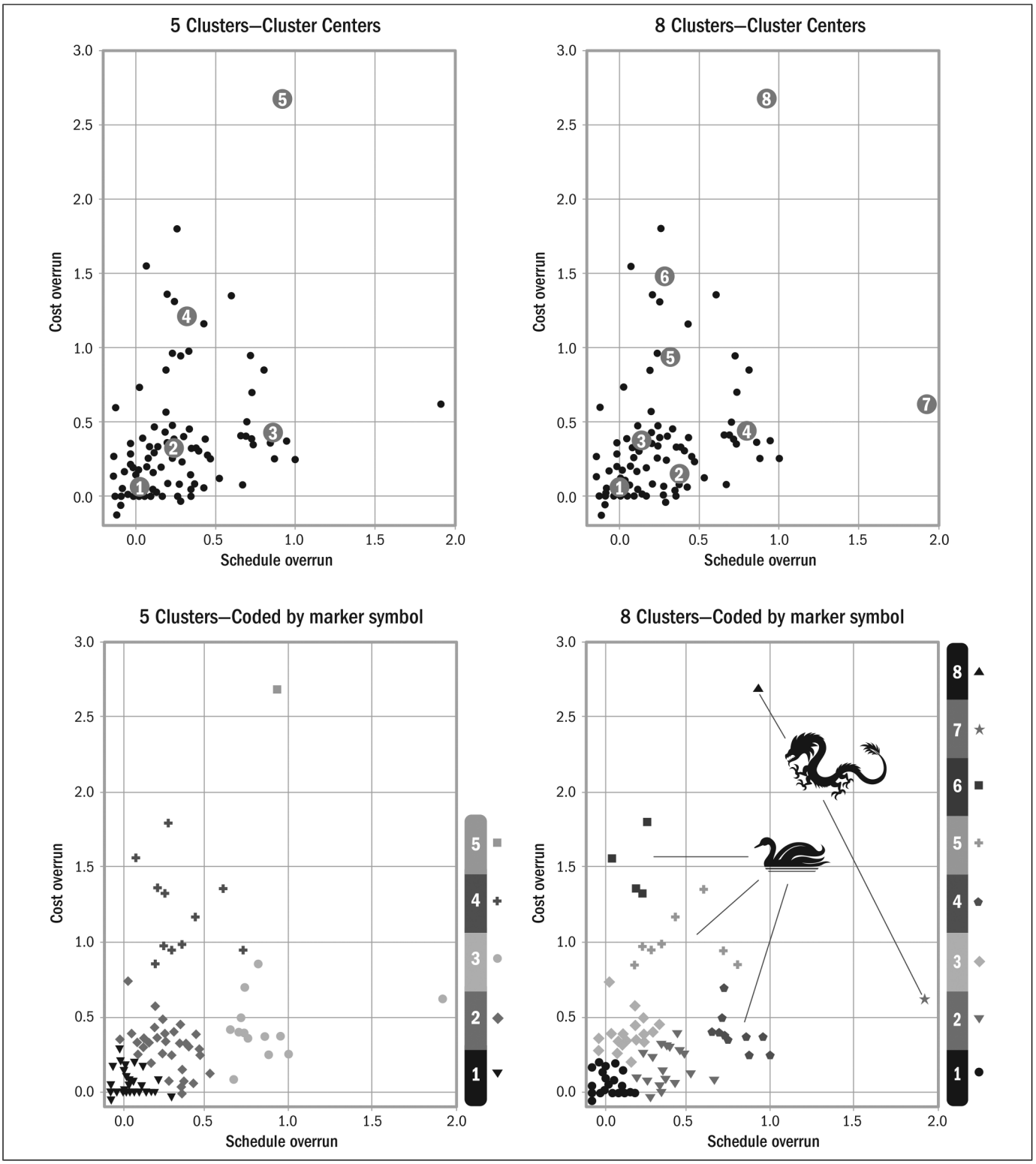

Figure 17. Clustering of project cost and schedule outcomes showing progressive separation of outliers.

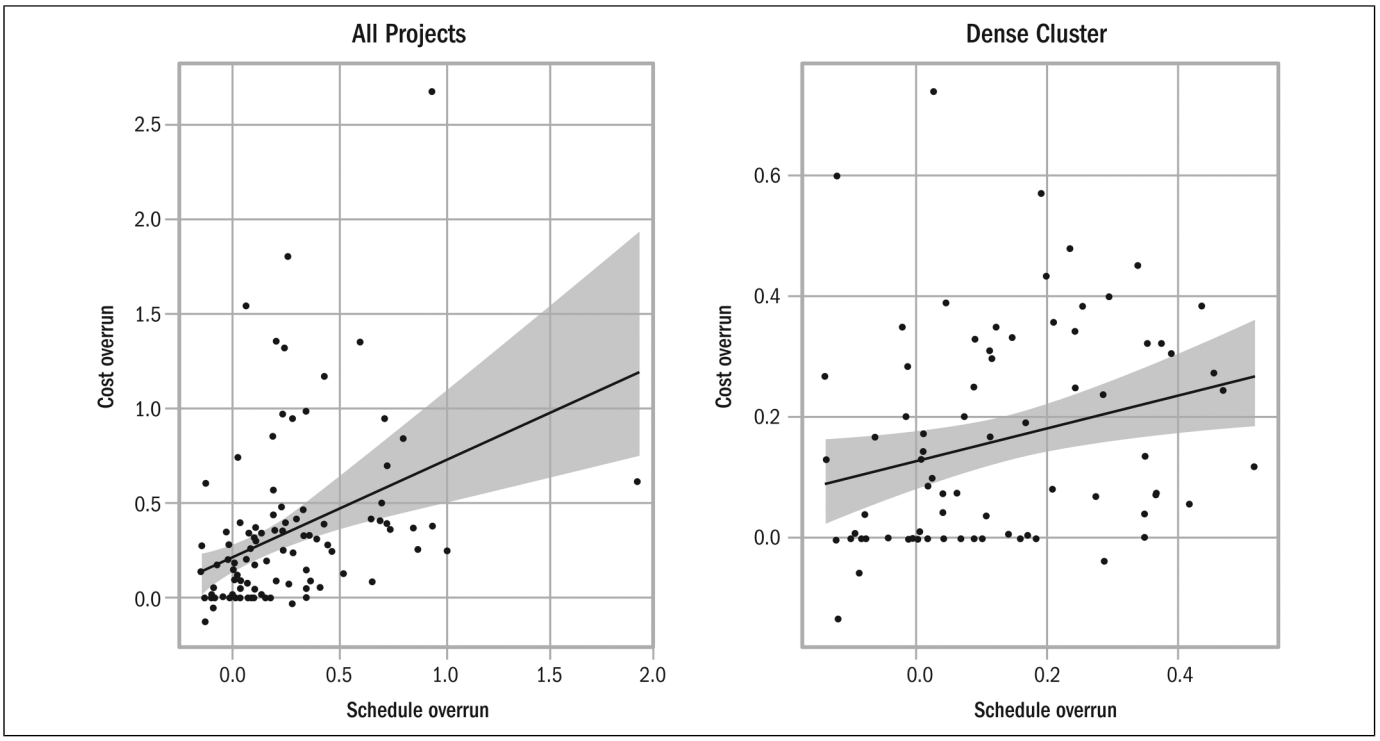

Figure 18. Cost and schedule overrun covariance.

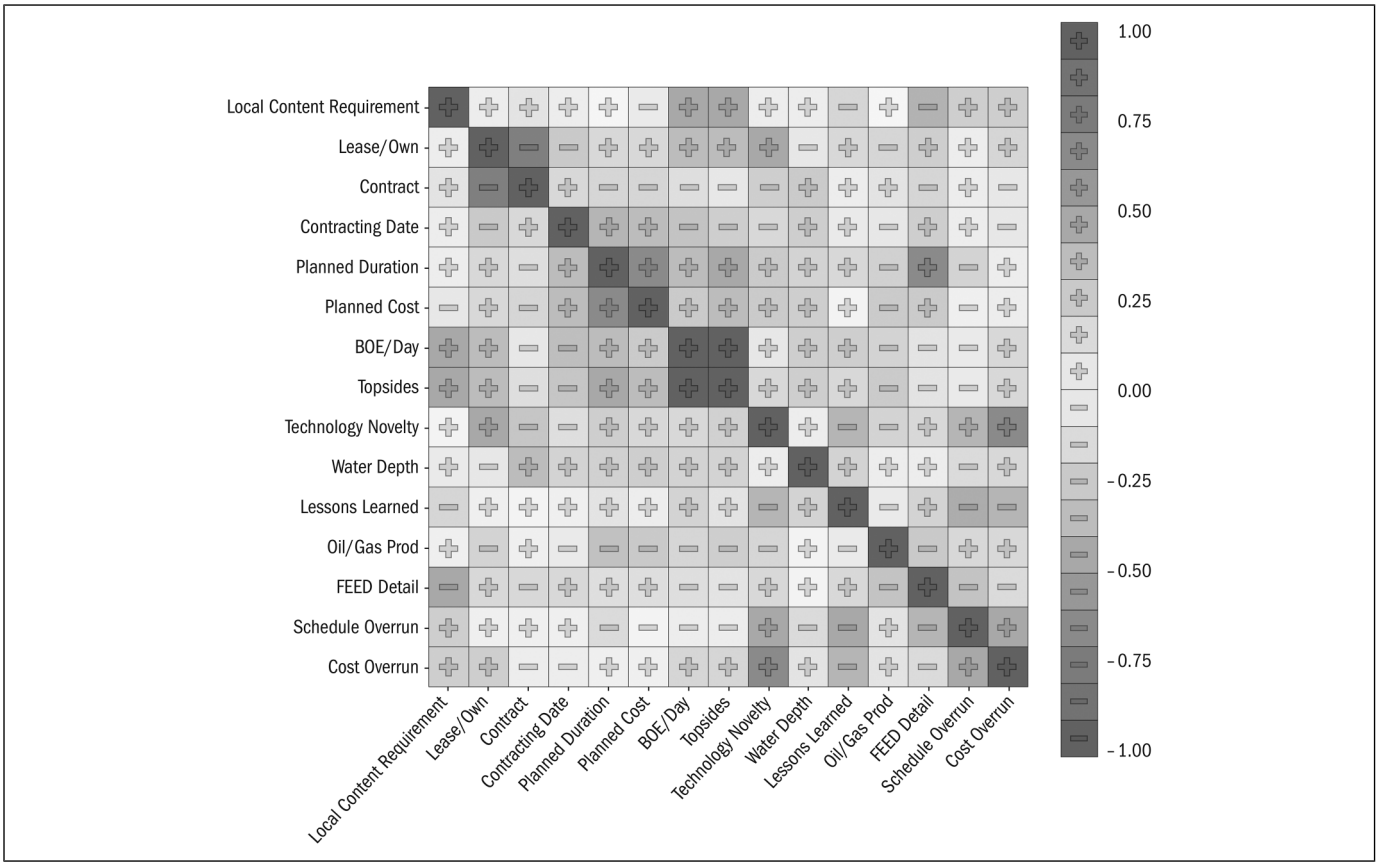

Figure 19. Correlation across project features and outcomes.

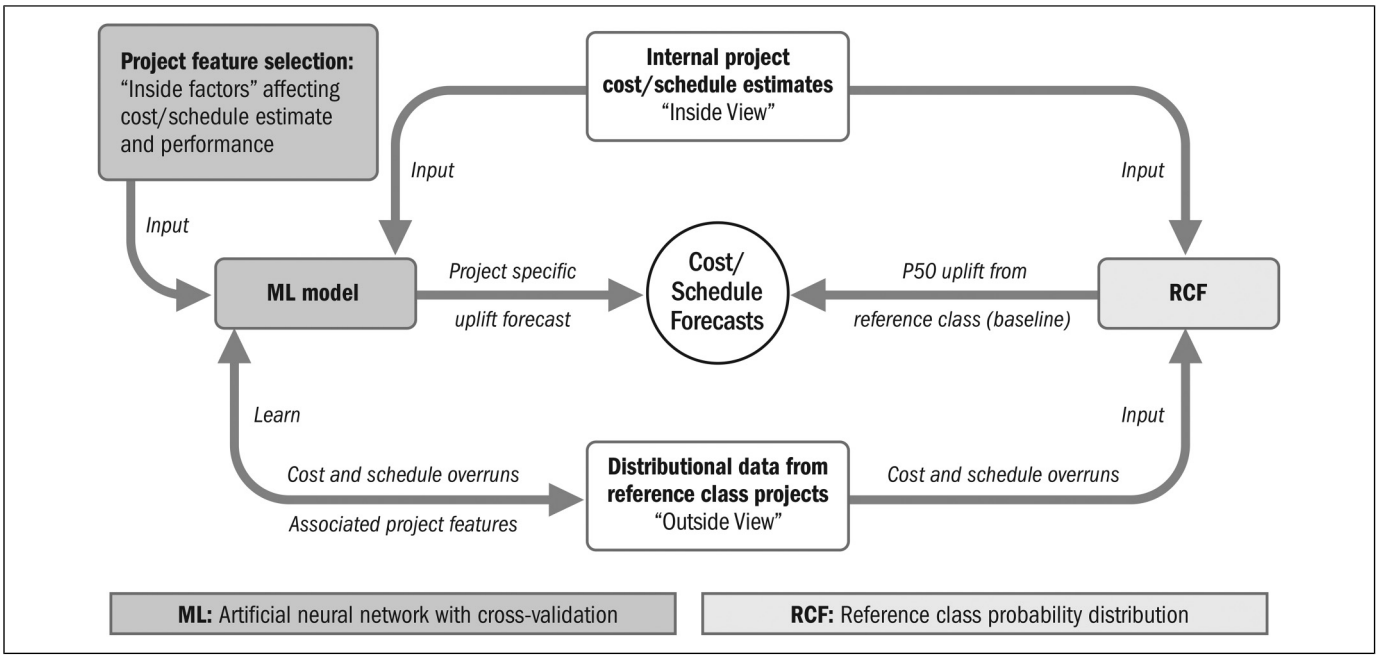

Figure 20. ML model.

Table 13. Models

Four-Layer_model		Two-Layer_model	
neurons/layer	activation_function	neurons/layer	activation_function
26	reLU	26	reLU
117	tanh	52	tanh
39	reLU	output-layer	sigmoid
7	reLU		
output-layer	sigmoid		
Both Models:		Loss_function: MSE; Optimizer: Adam	

et al., 2016). Our challenge was a limited dataset with several features. The model had to relate project features known at FID to overruns at delivery. Pearson correlations between features and overruns are plotted in Figure 19; some correlations, such as lessons learned and technology novelty, are clear. However, the effects of some features could be moderated by other features in complex ways and the model would have to learn these relationships.

Two supervised learning models were trained, one each for schedule and cost overrun. Project data were vertically split into features and outcomes (overruns) and split horizontally, 85:15, into training and validation datasets. Models learned relationships between features and overruns using training data. Validation data represented new data to test models for generalization performance. Given the limited dataset, cross-validation with 10-folds was used to assess generalization. Multilayer neural networks for deep-learning enable the learning of complex concepts from simpler building blocks.

Table 14. MSE Comparison

All Data		New Data		RCF	
Schedule	Cost	Schedule	Cost	Schedule	Cost
.83%	4.2%	6%	12.2%	9.3%	20.2%

Table 15. Forecasting Accuracy Comparison—Percentage of Accurate Forecasts

	ML				RCF	
	Schedule		Cost			
Accuracy	All data	New data	All data	New data	Schedule	Cost
5%	87%	24%	57%	25%	14%	11%
10%	94%	40%	89%	43%	23%	24%
15%	99%	59%	93%	53%	39%	41%

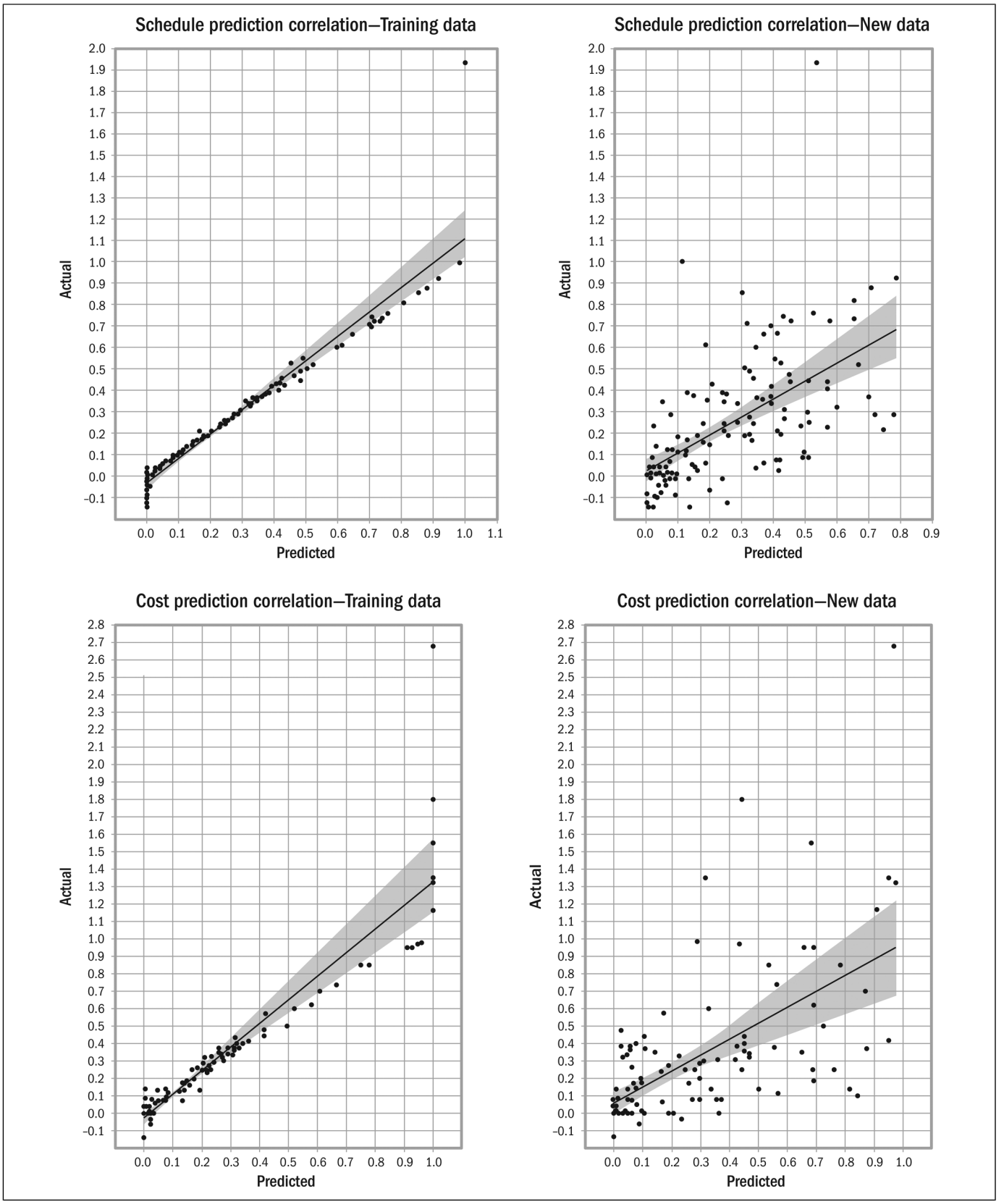

Figure 21. Correlation of predicted and actual overruns.

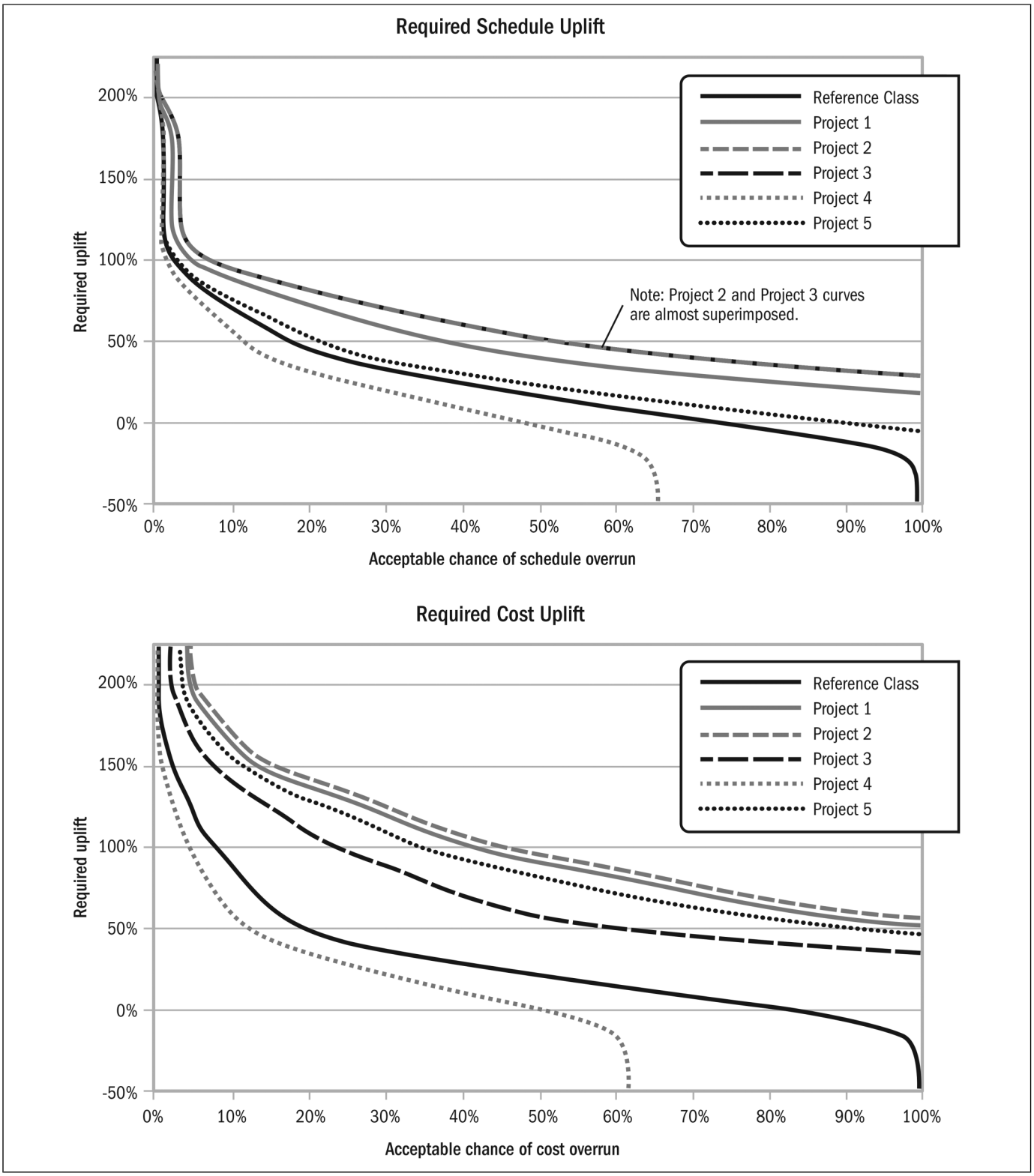

Figure 22. Project-specific uplifts as functions of acceptable risk.

Sequential models implemented in Python using the Keras and TensorFlow 2.0 open-source ML libraries were used. Prediction performance was compared to the P50 RCF schedule and cost overrun uplifts as the baselines as illustrated in Figure 20. The four and two hidden-layer models, which yielded good performance for cost/schedule uplift forecasting are described in

Table 16. Expected Overruns for Dummy Projects

Project	1	2	3	4	5
Water-depth (m)	2,000			100	800
Novelty	Very high			Very low	Medium
Lessons learned	Very low			Very high	
Topsides-size	Very large			Very small	
Planned duration (days)	2,000			500	1,000
Company	IOC	NOC		IOC	
Location	Australia	Brazil	Africa		Brazil
Predicted schedule-uplift	40%	52%	53%	.1%	23%
Predicted cost-uplift	93%	95%	58%	.4%	84%

Table 17. Required Uplifts for Acceptable Risk Percentiles—Dummy Projects

Project		1	2	3	4	5
Schedule	P10	88%	94	94%	56%	75%
	P25	67%	75%	76%	25%	44%
	P50	40%	52%	53%	0%	23%
	P75	29%	38%	39%	0%	9%
Cost	P10	167%	170%	143%	60%	159%
	P25	129%	132%	99%	28%	122%
	P50	93%	95%	58%	1%	84%
	P75	66%	72%	43%	0%	58%

Table 13. The mean squared error (MSE) is compared to conventional RCF in Table 14, and forecasting accuracies are compared in Table 15.

A four hidden-layer model, used to evaluate predictability for these complex projects, performed excellently on trained data: 99% of schedule and 93% of cost overrun forecasts were within 15% accuracy. The MSE was .83% for schedule and 4.2% for cost overrun; five to ten times better than conventional RCF. While extreme outliers were unpredictable, Pearson correlations between predicted and actual overruns were .96 for schedule and .93 for cost (Figure 21).

The four hidden-layer model's performance on validation data not used in training was considerably lower due to overfitting. Generalization performance was improved significantly by reducing layers and employing regularization techniques, limiting the number of epochs and regularization penalties on layers. A two hidden-layer model yielded the best generalization performance, which was significantly better than conventional RCF. Dimensionality reduction using PCA made little difference. Generalization performance was tested on all data by progressively splitting the data ~90:10 into training and validation datasets, such that over 10 iterations, all the projects were present in the validation dataset as new data exactly once. Predictions for all projects were collated and used to recompute MSE (see Table 14), correlation, and accuracy (see Table 15), for new data. The MSE, 6% for schedule and 12.2% for cost, Pearson correlations for new data, .59 and .6, respectively ($p << .01$), and prediction accuracy are very good for this challenging forecasting problem. Generalization performance will improve with better training data on more projects and further model optimization.

The models faltered for extreme outliers as expected from complexity theory, even though high overruns were predicted for those projects. For significant non-outlier deviations, factors absent in the model, such as leadership and unknown emergent issues, were seen to materially affect project outcomes. The excellent performance on training data projects is testament to the prediction model's trainability using distributional data. The comparatively inferior performance on new data reflects ML megaproject forecasting limitations from complexity and emergence; however, a significant portion of it is related to data unavailability and inaccuracy.

Significant overrun predictions not evident in actual data can indicate inaccurate or misleading reporting, or positive black swans and significant influence of factors absent from the model. A West African offshore project exemplified a positive black swan; it performed atypically better at close to 10% below the budget, saving more than US$1 billion from unexpectedly excellent drilling performance. Subsurface geological conditions are significant risks in offshore projects and, in this case, emergence helped. Investigation of another West African project, which reported atypically better performance, revealed evidence of significant cost overruns not reflected in the project but the contractor's books. While this can show misleadingly better results on some projects, having contractors take disproportionate risk is untenable over the long term, given the criticality of their expertise (Merrow, 2011).

A more practical generalization test was performed by predicting outcomes for dummy FPSO projects weighted toward high, medium, and low overruns, using features such as novelty, lessons learned, and front-end detail. The model correctly predicted higher uplifts for riskier projects, as seen in Table 16.

Project-specific uplift distributions for acceptable risk percentiles were generated from the uplift predictions and reference class using a Bayesian approach. Posterior probabilities were computed by assuming a 50% probability for the project-specific predictions and taking the reference class distribution as the prior. This yielded project-specific uplift curves (Figure 22), from which percentile uplifts for acceptable risk are shown in Table 17 for the dummy projects. Outlier risks are incorporated into the curves. While it is impractical to choose risk percentiles incorporating outliers for individual projects, it can be done for long-term portfolios. Novelty, scale, and front-end details can be revisited to reduce risk.

Hypothesis 4 is validated by realizing generalized models for separate cost and schedule forecasts that use individual project features for effective project-specific uplifts.

Conclusion

In this article, we showed that biases, such as optimism, representativeness and availability biases, and principal-agent issues affect O&G offshore project forecasting. Experts displayed significant underestimation of overruns but also significant awareness of them. Projects showed budget and schedule growth after every front-end approval stage, indicating principal-agent issues between and within companies. Responses from the master builders indicate better forecasting performance from some experts. However, these experts still showed bias, and the co-occurrence of the requisite extensive experience and predisposition to obtain such skill cannot be a reliable basis for megaproject planning.

The limitations of heuristics versus the effectiveness of expert intuition are related to the environment's predictability, as discussed by Kahneman and Klein (2009). From in-case and cross-case reviews of our projects, we postulate that expert intuition plays a greater role in response to post-FID execution-phase emergent issues and crises, something to be substantiated in future research. Heuristics can be beneficial in time-critical environments characterized by sparse data and computational capability limitations (Todd & Gigerenzer, 2012). The naturalistic decision-making (NDM) approach focuses on expert intuition's effectiveness in real-world situations bounded by limited unreliable data, computational intractability, and time limits. NDM models such as recognition-primed decision-making (RPD) explain this effectiveness by cue recognition for assessing emerging situations from prior experience and simulation of courses of action based on prior feedback on decision validity (Klein, 1993). Leadership is critical for crisis management (Bundy et al., 2017) during project execution, and effective response under time pressure emphasizes expert intuition. AI/ML may aid project management information systems (PMIS) to recognize emergent issues and support collaborative responses from experts during execution.

However, FID requires forecasting outcomes in planning settings rather than recognizing emerging outcomes in time-critical contexts with sparse information. For megaproject planning, high complexity and lengthy time-horizons can lead to high-impact practically unforeseeable outcomes. The feedback experts "receive from their failures in long-term judgments is delayed, sparse, and ambiguous" (Kahneman & Klein, 2009, p. 523), and principal-agent issues are significant, adversely affecting expert forecasting. In the context of forecasting for FID, the limitations of heuristics are significant.

We established and demonstrated methods to correct forecasts and reduce uncertainty, building on inside view baselines by correcting them. The theoretical foundation and its validation allowed us to find the place for ML in conjunction with expert judgment, balancing benefits from both. The benefits of inside view heuristics, discussed by Gigerenzer and Brighton (2011) and Klein (1993), were present in data interpretation and feature selection; ML models trained on selected features and outside view distributional data discussed by Kahneman and Tversky (1979) and Flyvbjerg (2006) determined project-specific uplift curves. Ecologically valid project features, trained and weighted on outcomes, can attain the benefits of robust heuristics that correspond to their environment discussed by Todd and Gigerenzer (2000). ML application was based on evidence that the same uplift does not apply to every reference class project for a given risk percentile. Our ML model corrected RCF uplifts by learning the relationship between project features and performance outcomes, helped by the chosen features' environmental validity. Our methods can ameliorate principal-agent issues, correct biases, and reduce overruns in project outcome data over time. This work can be substantiated and expanded as better data become available, and several approaches are outlined here.

The significance of our results can be seen from the substantial costs and delays represented by predicted overruns. Furthermore, the performance forecasting models can be used to manage project features to minimize the potential for overruns. Project features can be optimized using structured methods such as the complexity assessment tool described by Maylor et al. (2013) and by improving front-end development; the ML methods outlined here will provide quantitative feedback on overrun risk mitigation.

Potential Limitations of This Work

Models approximate real-world systems and possess inherent limitations. The future is mutable, and correlations between factors could change, affecting weights and relationships learned by models. Post-FID changes to contractors or scope will affect project-feature coding. Cost and schedule can be affected by macroeconomic and commodity cycles. Predictions are constrained by available distributional data and collection accuracy. The limited number of offshore projects is an upper limit, and many project's actual costs can be unreliable or unavailable. These limitations resulted in several approximations to dates, budgets, and costs during data collection. Furthermore, better projects with more readily available data due to outcome reporting bias may be overrepresented compared to problematic ones. Projects professing successful goal attainment can mask overrun costs borne by contractors, as on the Asgard B project in the North Sea (Upstream, 2000). However, rapid growth and advances in digitally enabled project delivery (Whyte, 2019) could lead to exponential growth in performance data reliability and quantity.

General Limitations and Recommendations

Human behavior can lead to nonoptimal adjustments to uplifted forecasts, such as work expansion to allotted capacity, Parkinson's Law (Parkinson, 1955) and procrastination, or Students' Syndrome (Goldratt, 1997). Principal-agent issues

can also result in nonoptimal adjustments, a criticism that has been directed at RCF (Themsen, 2019).

However, such criticism is essentially directed at *how* rather than *whether* RCF should be applied, and there is little reason not to welcome better means of quantifying underperformance risk. It would benefit companies, shareholders, and stakeholders, including societies, given the cost and impact of these projects. The use of uplifts should be in the context of principal-agent issues in an industry, as discussed by Flyvbjerg and Cowi (2004); they can be hidden or used with a fever chart controlling their expenditure.

As Taleb (2008) avers, while true planning is impossible, planning should be done while accounting for limitations. Detailed planning or ML cannot quantify all project risks; strict uncertainty, outcomes with unknown probabilities, and unknown unknowns, are absent from distributional data. Complexity can cause unknown emergent outcomes and chaos. Outliers are very significant to megaproject forecasting but individually unpredictable due to unforeseeable interaction among emergent issues. Outliers in distributional data or "gray swans" (Taleb, 2008, p. 213) have extremely low probabilities of reoccurring the same way. However, they can stand in for future unknown unknowns. Nonoptimal decisions can be satisfactory given temporal, informational, and computational limits when adapted to their environment; this is termed "satisficing" (Simon, 1956, p. 129). Our goals were to extend the reach of satisficing in planning complex projects. Decisions are made in the face of uncertainty, which we have attempted to reduce.

Future Work

This work provides a framework for integrating RCF with ML and can lead to further research on methods and applications. Application to other megaproject classes looks promising: infrastructure and software development projects show similar biases and fat-tailed outcome distributions; the rapidly growing offshore floating wind sector with limited prior projects has similarities to offshore O&G. This, and the limitations imposed on ML models by unavailable or unreliable project data, strongly substantiate the requirement for rigorous cross-industry project accounting standards and practices, such as the IASB and FASB rules for reporting in firms; project governance would benefit immensely. More accurate data collection and de-biasing of reporting can lead to significant improvements in forecasting accuracy.

There is further potential for optimizing the performance of our models. Furthermore, project features not in our models can be incorporated, and more accurate data can be utilized. The random number generated PDF showed great potential for approximating overrun distributions, especially for new sectors with few previous projects, such as offshore wind. Much data were obtained on the performance weightage of theory-coded features from our trained models, offering several promising avenues of inquiry.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ackoff, R. L. (1981). The art and science of mess management. *Interfaces*, *11*(1), 20–26.

Ansar, A., Flyvbjerg, B., & Budzier, A. (2014). Should we build more large dams? The actual costs of hydropower megaproject development. *Energy Policy*, *69*, 43–56.

Baccarini, D. (1996). The concept of project complexity—A review. *International Journal of Project Management*, *14*(4), 201–204. https://doi.org/10.1016/0263-7863(95)00093-3

Batselier, J., & Vanhoucke, M. (2016). Practical application and empirical evaluation of reference class forecasting for project management. *Project Management Journal*, *47*(5), 36–51.

Bellman, R. (2013). *Dynamic programming*. Princeton University Press.

Birnbaum, Z. W., & Saunders, S. C. (1968). *A new family of life distributions*. Boeing Scientific Research Laboratories.

Bruijn, H. D., & Leijten, M. (2007). Megaprojects and contested information. *Transportation Planning and Technology*, *30*(1), 49–69.

Budzier, A., & Flyvbjerg, B. (2013). *Making-sense of the impact and importance of outliers in project management through the use of power aws*. Oslo, International Research Network on Organizing by Projects (IRNOP).

Bundy, J., Pfarrer, M. D., Short, C. E., & Coombs, W. T. (2017). Crises and crisis management: Integration, interpretation, and research development. *Journal of Management*, *43*(6), 1661–1692.

Caron, F., & Ruggeri, F. (2016). Project management in the oil & gas industry—A Bayesian approach. In Wiley (Ed.), *Wiley StatsRef: Statistics reference online*. (pp. 1–14) John Wiley & Sons.

Carrier, R. C. (2012). *Proving history: Bayes's theorem and the quest for the historical jesus* (1st ed.). Prometheus.

Clegg, S. R., Biesenthal, C., Sankaran, S., & Pollack, J. (2017). Power and sensemaking in megaprojects. In B. Flyvbjerg (Ed.), *The Oxford handbook of megaproject management* (pp. 238–258). Oxford University Press.

Davies, A., & Mackenzie, I. (2014). Project complexity and systems integration: Constructing the London 2012 Olympics and paralympics games. *International Journal of Project Management*, *32*(1), 773–790. https://doi.org/10.1016/j.ijproman.2013.10.004

Denicol, J., Davies, A., & Krystallis, I. (2020). What are the causes and cures of poor megaproject performance? A systematic literature review and research agenda. *Project Management Journal*, *51*(3), 328–345.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. Buchanan & A. Bryman (Eds.), *The SAGE handbook of organizational research methods* (pp. 671–689). SAGE.

EMA. (2020). *Floating production systems quarterly report*. EMA.

Ernst & Young (EY). (2014). *Spotlight on oil and gas megaprojects*. https://www.ey.com/Publication/vwLUAssets/EY-spotlight-on-oil-and-gas-megaprojects/$FILE/EY-spotlight-on-oil-and-gas-megaprojects.pdf

Ernst & Young (EY). (2015). *Oil and gas megaproject development*. Ernst & Young.

Flyvbjerg, B. (2006). From nobel prize to project management: Getting risks right. *Project Management Journal*, *37*(3), 5–15.

Flyvbjerg, B. (2014). What you should know about megaprojects and why: An overview. *Project Management Journal*, *45*(2), 6–19. https://doi.org/10.1002/pmj.21409

Flyvbjerg, B. (2020). The law of regression to the tail: How to survive Covid-19, the climate crisis, and other disasters. *Environmental Science & Policy*, *114*, 614–618.

Flyvbjerg, B., Bruzelius, N., & Rothengatter, W. (2003). *Megaprojects and risk: An anatomy of ambition*. Cambridge University Press.

Flyvbjerg, B., & Budzier, A. (2011). Why your IT project might be riskier than you think. *Harvard Business Review*, *89*(9), 23–25.

Flyvbjerg, B., & COWI. (2004). *Procedures for dealing with optimism bias in transport planning: Guidance document*. UK Department for Transport.

Flyvbjerg, B., Garbuio, M., & Lovallo, D. (2009). Delusion and deception in large infrastructure projects: Two models for explaining and preventing executive disaster. *California Management Review*, *51*(2), 170–193.

Flyvbjerg, B., Ansar, A., Budzier, A., Buhl, S., Cantarelli, C., Garbuio, M., Glenting, C., Holm, M.S., Lovallo, D., Lunn, D., Molin, E., Rønnest, A., Stewart, A., & van Wee, B., (2018). Five things you should know about cost overrun. *Transportation Research Part A: Policy and Practice*, *118*, 174–190.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701.

Gigerenzer, G., & Brighton, H. (2011). *Homo heuristicus: Why biased minds make better inferences*. Oxford University Press.

Goldratt, E. (1997). *Critical chain* (1st ed.). North River Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (1st ed.). MIT Press.

Gyasi, E. A. (2017). *A Bayesian approach to cost estimation for offshore deepwater drilling projects*. The University of Warwick.

Hitchins, D. K. (2007). *Systems engineering: A 21st century systems methodology* (1st ed.). Wiley.

Hubbard, D. (2009). *The failure of risk management: Why it's broken and how to fix it* (1st ed.). Wiley.

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526.

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *TIMS Studies in the Management Sciences*, *12*, 313–327.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 138–147). Ablex Publishing.

Lee, C.-Y. (2019). *Integrative trust-based functional contracting: A complementary contractual approach to BIM-enabled oil and gas EPC project delivery*. Curtin University.

Locatelli, G., Mancini, M., & Romano, E. (2014). Systems engineering to improve the governance in complex project environments. *International Journal of Project Management*, *32*(8), 1395–1410.

Lovallo, D., & Kahneman, D. (2003). Delusions of success: How optimism undermines executives' decisions. *Harvard Business Review*, *81*, 56–63.

Lundin, R., & Soderholm, A. (1995). A theory of the temporary organization. *Scandinavian Journal of Management*, *11*(4), 437–455.

Lundrigan, C. P., Nuno, A. G., & Puranam, P. (2015). The (under) performance of mega-projects: A meta-organizational perspective. *Academy of Management Proceedings*, *1*(2015), 11299.

MacNicol, D. (2016). Taking the lead. *Construction Journal*, *1*(1), 16–18.

Maylor, H. R., Turner, N. W., & Murray-Webster, R. (2013). How hard can it be? Actively managing complexity in technology projects. *Research Technology Management*, *56*(4), 45–51.

Merrow, E. W. (2011). *Industrial megaprojects: Concepts, strategies, and practices for success* (1st ed.). John Wiley & Sons.

Merrow, E. W. (2012). *Oil and gas industry megaprojects: Our recent track record*. Society of Petroleum Engineers.

Mitchell, T. M. (1997). *Machine learning* (1st ed.). McGraw-Hill.

Müller, R. (2009). *Project governance* (1st ed.). Routledge.

Offshore Magazine. (2019). *2019 Worldwide survey of floating production, storage and offloading (FPSO) units*.

Olaniran, O. J., Love, P. E. D., Edwards, D. J., Olatunji, O., & Matthews, J., (2015). Chaotic dynamics of cost overruns in oil and gas megaprojects: A review. 17th International Conference on Oil, Gas and Coal Technology, 17(7).

Parkinson, C. N. (1955). Parkinson's law. *The Economist*, 19 November.

Raval, A. (2020). The last frontier: Oil industry scales back exploration. Financial Times, 21 July.

Remington, K., & Pollack, J. (2008). What is a complex project? In *Tools for complex projects* (pp. 1–12). Taylor & Francis Ltd.

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M., Gibson, S. P., Jarke, H., Karakasheva, R., Khorrami, P. R., Kveder, J., Andersen, T. L., Lofthus, I. S., McGill, L., Nieto, A. E., … Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, *4*(6), 622–633.

Rui, Z., Peng, F., Ling, K., Chang, H., Chen, G., & Zhou, X., (2017). Investigation into the performance of oil and gas projects. *Journal of Natural Gas Science and Engineering*, *1*(38), 12–20.

Rystad Energy. (2020). *A new offshore investment cycle is in the making*. Rystad Energy.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138.

Singh, A. (2010). *Strategies for oil and gas companies to remain competitive in the coming decades of energy challenges*. MIT Sloan School of Management.

Sornette, D. (2009). Dragon-kings, black swans and the prediction of crises. *International Journal of Terraspace Science and Engineering*, *2*, 1–18.

Steen, J., Ford, J. A., & Verreynne, M.-L. (2017). Symbols, sublimes, solutions, and problems: A garbage can model of megaprojects. *Project Management Journal*, *48*(6), 117–131.

Sydow, J., & Braun, T. (2018). Projects as temporary organizations: An agenda for further theorizing the interorganizational dimension. *International Journal of Project Management*, *36*(1), 4–11.

Taleb, N. (2008). *The black swan: The impact of the highly improbable*. Penguin.

Themsen, T. N. (2019). The processes of public megaproject cost estimation: The inaccuracy of reference class forecasting. *Financial Accountability & Management*, *35*(4), 337–352.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*(1), 727–780.

Todd, P. M., & Gigerenzer, G. (2012). What is ecological rationality? In P. M. Todd & G. Gigerenzer (Eds.), *Ecological rationality* (pp. 3–30). Oxford University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (New York, N.Y.)*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Upstream. (2000). *Kvaerner licks its Asgard wounds*. Upstream, 22 September.

Upstream. (2002). *Spar losses spur McDermott exit*. Upstream, 11 October.

Walls, L., & Quigley, J. (2001). Building prior distributions to support Bayesian reliability growth modelling using expert judgement. *Reliability Engineering and System Safety*, *74*(2), 117–128.

Warren, K. (2008). Chaos theory and complexity theory. In T. Mizrahi & L. E. Davis (Eds.), *Encyclopedia of social work* (1st ed., pp. 227–233). NASW Press.

Werndl, C. (2009). What are the new implications of chaos for unpredictability? *The British Journal for the Philosophy of Science*, *60*, 195–220.

Whyte, J. (2016). *The future of systems integration within civil infrastructure: A review and directions for research*. Edinburgh, 26th Annual INCOSE International Symposium (IS 2016).

Whyte, J. (2019). How digital information transforms project delivery models. *Project Management Journal*, *50*(2), 177–194.

Wright, B. (2009). *Design error that led to listing the first of many hurdles for project*. Upstream, 24 September.

Author Biography

Ananth Natarajan, PMP, is an experienced project manager with approximately 18 years of experience in O&G offshore megaprojects. His responsibilities have included the planning of complex projects around the world and the leadership of multidisciplinary, multicultural teams. He has bachelor's and master's degrees in mechanical engineering, an MBA, and a master's degree in major program management. Ananth is a Professional Engineer (PE) in Texas and is a Project Management Professional (PMP)® certification holder. Currently, he is the founder/CEO of a startup building a collaborative project governance platform for complex projects, cybereum, using bespoke Distributed Ledger Technology. He can be contacted at ananth.natarajan@cybereum.io