

## Explainable Decision Support for Financial Resilience Monitoring in State-Owned Energy Enterprises

Zhao Xue, Rustamov Ilham Samil

Azerbaijan State Oil and Industry University Baku, Azerbaijan  
*709169882@qq.com & xue.zhao2024@asoiu.edu.az*

**Abstract.** State-owned energy enterprises are exposed to commodity-price, exchange-rate, and inflation shocks, while external stakeholders often rely on fragmented public information to assess their financial condition. This study develops and evaluates a design-science-oriented analytical prototype for explainable financial-resilience monitoring. The prototype integrates audited financial statements, public macroeconomic data, source provenance, and data-confidence metadata, while specifying a separate schema for future management-disclosure evidence. It separates internal financial capacity, represented by profitability, liquidity and cash generation, operating efficiency, and solvency, from external stress arising from oil-price downside, exchange-rate depreciation, and inflation. Direction-aware normalization, hierarchical weighting, CRITIC, TOPSIS, prior-only thresholds, and explicit state rules generate four interpretable monitoring states. Using SOCAR data for 2009–2024, rolling-origin evaluation identifies 2015 as acute risk and 2022 as resilient absorption. The 2020 observation is also classified as acute risk under the baseline specification, although this result is sensitive to the external-stress threshold. The composite capacity score classifies all three prespecified cases consistently with their expected states, but several standalone financial indicators produce the same result. Its added value therefore lies in transparent synthesis, explanation, and traceability rather than demonstrated predictive superiority. A limited cross-firm test using KazMunayGas reproduces the qualitative 2020 and 2022 state pattern, subject to local calibration and an alternative profitability measure. The study contributes an explainable decision-support architecture rather than a fully implemented software system or real-time forecasting model.

**Keywords:** analytical prototype; decision support; explainability; financial-resilience monitoring; public data; state-owned energy enterprises; temporal validation

## 1. Introduction

### 1.1. Background and Information-Systems Gap

State-owned energy enterprises operate at the intersection of commercial performance, public ownership, and macroeconomic exposure. Their financial condition can be affected simultaneously by commodity-price movements, exchange-rate changes, domestic inflation, financing structure, and the organization's own capacity to generate profits and cash. For governments, lenders, managers, and other stakeholders, the relevant question is consequently broader than whether one accounting ratio improved or deteriorated. They need to know whether external pressure is elevated, whether the enterprise has sufficient internal capacity to absorb it, which factors drive the signal, and how much confidence can be placed in the underlying evidence.

Organizational-resilience research conceptualizes resilience as the capacity to anticipate, cope with, and adapt to adverse conditions, but it also recognizes persistent ambiguity in the construct and its measurement (Duchek, 2020; Hillmann & Guenther, 2021). In a financial-monitoring context, resilience should therefore not be equated with a favorable outcome in one year. Meaningful financial monitoring therefore requires a relational assessment of external pressure and the observed financial capacity available to absorb it.

For many state-owned enterprises, extensive internal operational information is not available to external analysts or public authorities in a consistent form. Public evidence is more accessible but fragmented across audited financial statements, annual reports, sustainability reports, commodity-market series, exchange-rate records, and price statistics. Differences in units, accounting definitions, indicator direction, reporting dates, missing values, and source authority make direct integration difficult. A spreadsheet of ratios may summarize historical performance, but it does not automatically provide semantic reconciliation, warning thresholds, traceable transformation rules, or a defensible interpretation of missing narrative evidence.

The core gap is not an absence of financial indicators. Profitability, liquidity, efficiency, solvency, and cash-flow ratios are established tools of corporate analysis, and composite indices can combine them. The unresolved information-systems problem is how to transform heterogeneous public observations into a reusable monitoring artifact whose data lineage, indicator meanings, weighting logic, thresholds, and decision outputs are explicit. Without such an architecture, an apparently sophisticated score may remain a retrospective calculation rather than a decision-support system.

The system must be reproducible, temporally valid, and explainable. Source values and transformations should be auditable, historical warnings should exclude future information, and consequential outputs should rely on interpretable decision logic (Buneman et al., 2001; Hind et al., 2023; Rudin, 2019). Transferability presents a further challenge because accounting definitions, currencies, and data coverage differ across enterprises. A useful system must therefore distinguish portable architecture from firm-specific calibration.

### 1.2. Research Aim and Questions

This study aims to develop and evaluate an explainable analytical prototype for financial-resilience monitoring in state-owned energy enterprises. SOCAR serves as the development case, while KazMunayGas provides a limited cross-firm transferability test. Following the design-science view that information-systems research can create and evaluate purposeful artifacts (Hevner et al., 2004), the study treats the data-integration pipeline, scoring rules, warning logic, explanation structure, and provenance layer as its principal outputs. The prototype is an implemented analytical workflow rather than a standalone software application, database platform, or interactive dashboard (Peppers et al., 2007).

The study addresses four research questions:

1. RQ1. How can heterogeneous public financial, macroeconomic, and disclosure data be integrated into a reproducible and provenance-aware monitoring architecture for a state-owned energy

enterprise?

2. RQ2. How can internal financial capacity and external macroeconomic stress be converted into transparent scores and interpretable decision states without introducing look-ahead information?

3. RQ3. Does the resulting analytical prototype identify prespecified shock-year conditions beyond what is conveyed by individual financial and macroeconomic indicators, and how sensitive are its outputs to weights, thresholds, normalization choices, and indicator selection?

4. RQ4. To what extent can the architecture and decision logic be instantiated in a comparable state-owned energy enterprise?

### **1.3. Proposed Artifact, Evaluation, and Contributions**

The analytical prototype comprises five linked layers. A data layer integrates audited corporate statements and public macroeconomic series. A semantic layer defines indicator meaning, direction, units, source hierarchy, and missing-data treatment. An analytics layer constructs an internal-capacity score and an external-stress score using direction-aware normalization, hierarchical weighting, and TOPSIS aggregation. A decision layer compares the two scores with prior-only thresholds and assigns one of four states: stable, latent vulnerability, resilient absorption, or acute risk. An explanation and governance layer exposes indicator contributions, source provenance, narrative evidence, and data-confidence metadata.

The artifact is evaluated through complementary tests rather than sample fit alone. SOCAR shock years are assessed with rolling-origin rules that use only preceding observations. The composite outputs are compared with standalone financial and macroeconomic indicators. Robustness tests vary the weighting method and remove indicators one at a time. A second state-owned energy enterprise, KMG, is used for a leave-event-out cross-firm transferability test. These procedures follow the design-science principle that artifact utility and contribution must be demonstrated through appropriate evaluation, not inferred from technical novelty alone (Hevner et al., 2004; Gregor & Hevner, 2013).

The study contributes a reusable decision-support architecture that links semantic reconciliation, capacity-stress separation, prior-only warning logic, explainable outputs, and evidence governance. Its contribution lies in the analytical workflow and associated design knowledge rather than in a standalone software interface. The KMG application examines whether this architecture can be instantiated in a second enterprise while allowing local recalibration of accounting measures, weights, and thresholds. The claims remain bounded because the prototype is not a causal model, an automated policy engine, or a real-time forecasting application.

## **2. Literature Review and Design Requirements**

### **2.1. Organizational and Financial Resilience**

Resilience is broadly concerned with maintaining or recovering functioning under adversity, yet organizational research has used the term to denote capacities, processes, outcomes, and desirable traits. Duchek (2020) organizes organizational resilience around anticipation, coping, and adaptation, emphasizing that resilience depends on capabilities that operate before, during, and after disruption. Hillmann and Guenther's (2021) systematic review similarly identifies conceptual fragmentation and measurement ambiguity. These accounts imply that resilience cannot be inferred from one favorable result or from the absence of failure. It concerns the relationship between adverse conditions and the capacity to respond.

Financial-resilience research inherits the same ambiguity. Recent reviews report growing attention to the construct but no universally accepted definition or consistent measurement scale (Tahir & Richards, 2025). Firm-level studies often infer resilience from post-shock outcomes, survival, liquidity, solvency, or the association between pre-shock financial conditions and subsequent performance. For example, Tascón et al. (2024) examine whether pre-pandemic financial

characteristics shaped the resilience of European hospitality firms during COVID-19. Such work shows why profitability, cash generation, liquidity, and leverage matter, but outcome-centered designs do not automatically yield a reusable monitoring mechanism. Longitudinal evidence further indicates that cash-flow management is associated with subsequent firm financial performance, supporting the treatment of cash generation as a distinct component of internal financial capacity rather than merely a short-term accounting outcome (Kroes & Manikas, 2014).

The present study adopts a deliberately narrower construct than organizational resilience. A financial-resilience monitoring state refers to the observed configuration of an enterprise's internal financial capacity under external macroeconomic pressure. Internal capacity is represented by profitability, liquidity and cash generation, operating efficiency, and solvency. These indicators describe financial condition and the apparent capacity to absorb pressure; they do not directly measure anticipation, adaptation, recovery processes, managerial capability, or organizational learning. The analytical prototype therefore separates internal financial capacity from external stress and interprets their combination as a monitoring state rather than a complete measure of organizational resilience. A high-stress year is not automatically classified as failure, and strong financial ratios in a low-stress year are not treated as demonstrated resilience. Because annual public accounts cannot reveal the managerial or ownership mechanisms behind an observed outcome, the resulting states provide diagnostic evidence rather than causal explanations.

## **2.2. Public Data, State Ownership, and Information Quality**

State ownership increases the public relevance of enterprise monitoring while also creating distinctive information conditions. State-owned enterprises may pursue commercial and public-policy objectives, operate under different capital constraints, and receive explicit or implicit state support. International governance guidance consequently emphasizes material financial and non-financial disclosure by enterprises and informed monitoring by the state as owner (OECD, 2020). Public annual reports, audited statements, and sustainability disclosures are therefore legitimate components of an external monitoring architecture, but their availability does not guarantee analytical comparability. Publicly disclosed sustainability information has also been systematically associated with the financial metrics of state-run firms, indicating that accounting records and non-financial corporate disclosures can provide complementary evidence for external monitoring (Naddyrah & Rusmanto, 2024). Recent JLISS research has demonstrated that publicly disclosed sustainability information can be systematically related to the financial metrics of state-run firms, supporting the integration of accounting records and non-financial corporate disclosures for external monitoring (Naddyrah & Rusmanto, 2024).

Public-data integration is a semantic problem as well as a collection problem. A label can change across reporting years; consolidated entities and currencies may change; accounting restatements can revise history; and similar terms can represent different constructs. Total liabilities, for example, should not be interpreted as trade payables, and narrative silence should not be coded automatically as an adverse management response. These risks are particularly serious in a small annual series because one misclassified value can materially alter a normalized score, an objective weight, or a threshold. Research on public-sector document management indicates that the usefulness of an analytical system depends not only on data processing but also on document classification, information retrieval, interoperability, regulatory compliance, and implementation governance (Yang et al., 2025). Although the present prototype does not employ large language models, these considerations support the use of explicit source records, semantic reconciliation, traceable transformations, and controlled treatment of heterogeneous corporate documents.

Information-quality research has long argued that accuracy alone is insufficient. Data quality also depends on contextual fitness, representation, accessibility, and the needs of data users (Wang & Strong, 1996). Provenance research complements this perspective by asking where a value came from

and how it was transformed (Buneman et al., 2001). More recent work links provenance documentation to transparency, accountability, and explainability in analytic systems (Hind et al., 2023). For the present setting, these literatures imply that every indicator should have a documented source, definition, direction, unit, reporting period, transformation history, and treatment of missingness.

Data confidence should remain separate from enterprise condition so that incomplete evidence neither silently lowers the resilience score nor creates false precision. The second design requirement is therefore a provenance-aware integration layer accompanied by a parallel confidence assessment.

### **2.3. Explainable Multi-Criteria Decision Support**

A composite indicator can reduce a multidimensional phenomenon to a tractable signal, but it also embeds choices about conceptual structure, data selection, normalization, weighting, aggregation, and robustness. The OECD/JRC handbook treats these choices as an iterative construction and validation process and stresses sensitivity analysis and links to underlying measures (OECD & European Union, 2008). A composite score is consequently defensible only when its construction rules are transparent and when users can recover the component information that the aggregation compresses.

Multi-criteria decision-making methods are relevant because internal financial capacity contains benefit indicators—such as profitability, liquidity, cash generation, and efficiency—and cost indicators such as leverage. TOPSIS ranks observations by relative proximity to a positive ideal and distance from a negative ideal. Its wide application reflects an intuitive decision logic and its ability to combine criteria expressed on different scales after normalization (Behzadian et al., 2012). In this study, TOPSIS is not used to proclaim a universally best year; it provides a bounded annual capacity score whose drivers remain visible.

Weighting is a central source of model risk. Equal weights are simple and transparent but assume identical importance. Purely judgmental weights can encode domain knowledge but may be difficult to reproduce or defend. CRITIC derives objective weights from contrast intensity and inter-criterion conflict (Diakoulaki et al., 1995), allowing indicators with greater variation and less redundancy to receive more weight. However, applying CRITIC across all ratios in one flat layer can allow a dimension with more indicators to dominate by construction. The present artifact therefore uses a hierarchy: conceptually defined dimensions receive equal shares, while CRITIC allocates weight among indicators within a dimension.

External stress is constructed separately from oil-price downside, exchange-rate depreciation, and inflation. This treatment is consistent with evidence that oil-price movements have heterogeneous firm-level financial consequences, indicating that oil-price pressure should be modeled as an external exposure rather than interpreted as a direct measure of internal organizational resilience (Khan et al., 2023; Narayan & Sharma, 2011). Equal weights are used because the stress layer contains only three conceptually distinct pressures and the short series is insufficient for strong data-driven discrimination among them. In both layers, direction-aware min-max normalization retains interpretability but is sensitive to extreme observations.

Decision support must connect numerical outputs to action-relevant interpretations while allowing users to inspect how each conclusion was reached. In consequential settings, intrinsically interpretable models are preferable when they can perform the required analytical task without relying on an opaque prediction system (Rudin, 2019). Given the small sample and monitoring purpose, explicit scoring, threshold, and state-assignment rules are appropriate for the present artifact.

Explainability operates at the data, indicator, state, and evidence levels. The system records data origin and transformations, displays indicator contributions, makes threshold comparisons explicit, and presents narrative evidence and confidence metadata separately from the quantitative score. Median-based thresholds provide a reproducible rule for the short series without claiming economic

optimality. Threshold crossings are therefore interpreted as prompts for review rather than automated prescriptions.

## 2.4. Temporal Validity, Transferability, and Design Requirements

Historical monitoring systems are vulnerable to look-ahead bias. Full-period normalization, weights, or thresholds use later observations to reinterpret earlier years, which can make a retrospective classification appear more credible than it would have been contemporaneously. Forecast-evaluation research distinguishes in-sample fit from performance on observations not used for estimation and documents the importance of origin choice, window updating, and test-period construction (Tashman, 2000). Time-series validation also requires attention to dependence and chronological structure rather than uncritical use of random folds (Bergmeir et al., 2018).

The appropriate principle for this study is prior-only estimation. For a target year, normalization ranges, eligible history, weights, and thresholds are estimated from preceding years. The target observation is then transformed and classified without allowing future years to affect the decision. This rolling-origin approach approximates the information set available at the time, even though annual public accounts may be released after year-end. It evaluates warning-rule integrity rather than claiming a real-time market forecast.

Small-N evaluation requires benchmarks against individual indicators and sensitivity tests of weights, thresholds, and indicator selection. These tests cannot create statistical power, but they can reveal whether conclusions depend on arbitrary modeling choices.

A single-company artifact remains narrow unless local calibration is distinguished from portable design. External validation can examine whether the architecture transfers to another relevant setting, although accounting definitions, reporting coverage, currencies, and exposure structures may still require local recalibration (Debray et al., 2015).

KMG therefore provides a limited external test of the architecture, direction rules, dimensions, and state logic. Because the event years are excluded but later non-event observations remain available, this is leave-event-out validation rather than real-time prediction.

The review identifies a coherent gap. Financial ratios and resilience concepts are available, but the literature does not by itself specify how a public-data user should construct an auditable, temporally valid, explainable, and transferable monitoring artifact for a state-owned energy enterprise. Table 1 translates that gap into six requirements used to design and evaluate the system.

Table 1. Literature-derived design requirements

| Literature gap                                                | Design requirement                                                  | Artifact implementation                                                                  | Evaluation evidence                                   |
|---------------------------------------------------------------|---------------------------------------------------------------------|------------------------------------------------------------------------------------------|-------------------------------------------------------|
| Resilience conflates pressure, capacity, process, and outcome | DR1: Separate internal capacity from external stress                | Two scores and four explicit decision states                                             | Shock-year state interpretation                       |
| Public data are heterogeneous and uneven                      | DR2: Preserve semantics, provenance, and evidence confidence        | Data dictionary; source hierarchy; narrative and confidence layers                       | Coverage and provenance audit                         |
| Composite indicators conceal modeling choices                 | DR3: Use transparent hierarchical aggregation and expose components | Direction-aware normalization; dimension hierarchy; CRITIC/TOPSIS; separate stress index | Weight, threshold, and indicator sensitivity          |
| Opaque scores are unsuitable for consequential review         | DR4: Make classifications intrinsically interpretable               | Driver decomposition and explicit threshold rules                                        | Traceable state explanations and benchmark comparison |
| Full-period estimation contaminates historical                | DR5: Enforce prior-only temporal validation                         | Rolling-origin normalization, weighting, and thresholds                                  | Held-out 2015, 2020, and 2022 tests                   |

| Literature gap                                   | Design requirement                                                 | Artifact implementation                                             | Evaluation evidence                        |
|--------------------------------------------------|--------------------------------------------------------------------|---------------------------------------------------------------------|--------------------------------------------|
| warnings                                         |                                                                    |                                                                     |                                            |
| Single-firm calibration limits<br>generalization | DR6: Distinguish portable<br>architecture from local<br>parameters | Second-firm instantiation with disclosed<br>proxy and recalibration | KMG leave-event-out<br>external validation |

*Note.* DR = design requirement. The requirements govern both construction and evaluation of the artifact.

Together, these requirements position the SOCAR implementation as an artifact instance and the integration logic, state semantics, and validation principles as the more portable design knowledge (Hevner et al., 2004; Peffers et al., 2007; Gregor & Hevner, 2013).

### 3. Research Design and Method

#### 3.1. Research Design and Case Selection

This study adopts a design-science-oriented longitudinal case approach to develop and evaluate an explainable analytical prototype for financial-resilience monitoring. The prototype is operationalized as a reproducible data-processing, scoring, classification, and explanation workflow. It does not include a standalone graphical interface or database application. The objective is not to estimate a causal relationship between individual ratios and firm performance, but to convert heterogeneous public information into interpretable monitoring states under changing oil-price, currency, and inflation conditions. In this study, “out-of-sample” refers to chronological separation between the parameter-estimation period and the held-out accounting year. It does not mean that the classification was available during that year, because audited annual statements are generally published after the reporting period.

SOCAR serves as the primary development case because its public reports provide a relatively long sequence of audited financial information and supplementary operational and sustainability disclosures. The observation period covers 2009–2024. KazMunayGas (KMG), another state-owned energy enterprise operating in a resource-dependent emerging economy, is used for a limited cross-firm instantiation test of whether the architecture and decision logic can be applied beyond the development case. This longitudinal case structure is intended to generate analytical insight into system design and temporal process patterns rather than population-level statistical generalization (Eisenhardt, 1989; Langley, 1999).

The research follows four stages. First, financial, macroeconomic and disclosure-related records are extracted and reconciled. Second, the records are transformed into internal financial-capacity and external-stress indicators. Third, the system assigns each firm-year to one of four interpretable decision states. Fourth, the classifications are evaluated through rolling-origin testing, prespecified shock-year comparisons, standalone-indicator benchmarks, sensitivity analyses, and a limited leave-event-out cross-firm instantiation using KMG.

#### 3.2. Research Design and Case Selection

The analytical prototype consists of four functional layers: data acquisition and provenance, indicator construction, analytical processing, and decision output. At the data-acquisition layer, the system imports audited consolidated financial statements, annual reports, sustainability reports and external macroeconomic series. Every record is linked to its reporting year, source document, accounting definition and, where available, page or table reference. This layer also identifies restatements and differences between reported, adjusted and researcher-calculated measures.

At the indicator-construction layer, consistently available quantitative data are organized into internal financial capacity and external stress. Narrative and semi-structured disclosures are processed separately as explanatory evidence. This separation prevents incomplete narrative coverage from mechanically changing the quantitative warning score.

At the analytical layer, the system applies direction-aware normalization, hierarchical weighting, TOPSIS aggregation, historically estimated thresholds and rolling-origin evaluation. All parameters used to classify a held-out year are estimated exclusively from earlier observations. This rule prevents look-ahead contamination in retrospective evaluation but does not convert annual accounting data into a contemporaneous forecast.

At the decision-output layer, the prototype produces the monitoring state, external-stress score, internal-capacity score, classification drivers, source provenance, and data-confidence metadata. The architecture can accommodate management-disclosure evidence, but that component is not empirically evaluated in the present study.

The resulting architecture is more than a retrospective calculation of financial ratios. Data reconciliation, missing-evidence handling, weighting, threshold formation, warning classification and explanation generation are explicit and reproducible components of the analytical workflow.

Table 2 Decision-State Classification Rules and Interpretations

| External stress | Financial capacity | Condition                                   | Decision state       | Interpretation                                                       |
|-----------------|--------------------|---------------------------------------------|----------------------|----------------------------------------------------------------------|
| Low             | High               | $ES_t < \tau ES$ and $FC_t \geq \tau FC$    | Stable               | Adequate capacity under ordinary external conditions                 |
| Low             | Low                | $ES_t < \tau ES$ and $FC_t < \tau FC$       | Latent vulnerability | Weak internal capacity may remain hidden in a favourable environment |
| High            | High               | $ES_t \geq \tau ES$ and $FC_t \geq \tau FC$ | Resilient absorption | Sufficient capacity is retained during a material external shock     |
| High            | Low                | $ES_t \geq \tau ES$ and $FC_t < \tau FC$    | Acute risk           | External pressure and weak internal capacity coincide                |

This two-axis design avoids treating every adverse external event as evidence of organizational failure. High inflation or a decline in oil prices may represent substantial external stress, but an enterprise with adequate liquidity, profitability, cash generation and solvency can still be classified as demonstrating resilient absorption.

Conversely, the system can identify latent vulnerability when financial capacity is weak even though external conditions have not crossed the stress threshold. This state is especially relevant for early intervention because deterioration may otherwise remain concealed until a major shock occurs.

### 3.3. Research Design and Case Selection

The primary SOCAR dataset combines audited consolidated financial statements for 2009–2024 with annual-report and sustainability-report records covering operational, economic, and technological disclosures (State Oil Company of the Republic of Azerbaijan [SOCAR], 2010–2025). External variables include annual Europe Brent spot prices obtained from the U.S. Energy Information Administration, the official annual-average AZN/USD exchange rate, and annual consumer-price inflation obtained from the World Development Indicators (U.S. Energy Information Administration, 2026; World Bank, n.d.-a, n.d.-b). The financial records were reconciled through accounting-identity checks. Total assets were compared with total liabilities and equity, while reported profit components and cash-flow measures were checked against their respective statements. Where a later report restated comparative information, the version and treatment of the restatement were recorded in the data notes.

KMG data for 2018–2024 were collected from the company’s official 2019, 2021, and 2024 annual reports (National Company KazMunayGas JSC, 2020, 2022, 2025). Kazakhstan’s annual consumer-price inflation and official KZT/USD exchange rate were obtained from the World Development Indicators, while the Brent series was obtained from the U.S. Energy Information Administration. To preserve cross-company comparability, only indicators that could be constructed

for both enterprises from consolidated financial statements were retained. Firm-specific adjusted measures and EBITDA-based indicators were excluded from the cross-firm implementation.

For KMG, the restated 2020 comparative figures reported in the 2021 annual report were used because the presentation of QazaqGaz as a discontinued operation affected the originally published revenue series. The 2022 and 2023 observations were taken from the comparative figures reported in the 2024 annual report. IFRS net cash generated from operating activities was retained instead of the company's adjusted cash-flow measure. Current assets and current liabilities were reconstructed from the corresponding statement categories when a directly reported subtotal was unavailable. These decisions were recorded in the provenance file so that changes in accounting presentation would not be interpreted automatically as changes in financial capacity.

The available public-data workbook contains extensive structured financial and operational records but does not contain a complete longitudinal corpus of management-disclosure passages for every reporting year. Narrative information is therefore excluded from the financial-capacity score, external-stress score, thresholds, and event classifications. The absence of a retrievable passage is recorded as unavailable rather than coded as evidence of weak management preparedness.

The proposed architecture nevertheless specifies a provenance-aware schema for future narrative evidence. A complete record would contain the firm, reporting year, source document, page or section, evidence category, supporting passage, and coding justification. The proposed categories include risk governance, scenario or sensitivity analysis, liquidity and capital discipline, digital monitoring, and business-continuity or shock-response measures. Evidence could subsequently be classified as not available, general disclosure, or an implemented and verifiable mechanism. The last category would require both an identifiable action, process, system, or governance arrangement and a verification attribute such as an implementation date, responsible unit, quantitative target, reported outcome, or traceable cross-reference.

This narrative schema is a design specification rather than an empirically validated component of the present prototype. The current study does not report category frequencies, coded longitudinal comparisons, or intercoder agreement because the available corpus is incomplete and was not independently double-coded. Narrative information is therefore not used to support the reported classifications. Empirical implementation of this layer will require a complete document corpus, a published coding dictionary, at least two independent coders or validated natural-language-processing rules, and formal reliability assessment.

### **3.4. Research Design and Case Selection**

Candidate indicators were screened using five criteria. First, each measure had to represent a conceptually distinct aspect of internal financial capacity. Second, it had to be reproducibly constructed from audited consolidated statements rather than management-adjusted figures. Third, its accounting definition and direction had to remain sufficiently consistent across the SOCAR observation period. Fourth, the measure had to be interpretable as either a benefit or cost indicator without estimating an additional model from the short annual series. Finally, preference was given to measures that could also be constructed, or approximated with explicit disclosure, for the KMG transferability test. Parsimony was necessary because a large indicator set would make the short rolling windows unstable and could give dimensions containing more variables disproportionate influence.

The resulting set covers profitability through operating margin and return on assets, liquidity and cash generation through the current ratio and operating cash flow to revenue, operating efficiency through asset turnover, and solvency through debt to assets. The measures do not form an exhaustive inventory of financial risk. Interest coverage, debt-maturity structure, refinancing exposure, free cash flow, capital-expenditure commitments, cash reserves, and short-term debt were considered conceptually relevant but were not retained in the primary specification. Some lacked a sufficiently

consistent public series or comparable definition across reporting years and firms, while others would duplicate an existing dimension or increase the influence of liquidity and solvency in the small indicator set. Their exclusion reflects the requirements of longitudinal comparability, reproducibility, and dimensional balance rather than an assumption that they are economically unimportant. Table 3 presents the six indicators retained through this selection protocol, together with their definitions and directional treatment.

Table 3 Internal Financial-Capacity Indicators and Directions

| Dimension                     | Indicator                      | Definition                           | Direction |
|-------------------------------|--------------------------------|--------------------------------------|-----------|
| Profitability                 | Operating margin               | Operating profit / revenue           | Benefit   |
| Profitability                 | Return on assets               | Net profit / total assets            | Benefit   |
| Liquidity and cash generation | Current ratio                  | Current assets / current liabilities | Benefit   |
| Liquidity and cash generation | Operating cash flow to revenue | Net operating cash flow / revenue    | Benefit   |
| Operating efficiency          | Asset turnover                 | Revenue / total assets               | Benefit   |
| Solvency                      | Debt to assets                 | Total liabilities / total assets     | Cost      |

Operating margin is used in the main SOCAR analysis because it measures operating performance before financing and tax effects. For the KMG cross-firm instantiation, net profit margin replaces operating margin because a consistently comparable operating-profit series could not be constructed across all KMG reporting years. The substitution is disclosed as a limitation because net margin incorporates financing and taxation effects that are excluded from operating margin. The remaining five indicators retain the same definitions.

External stress is represented by three variables:

$$OD_t = \max(0, -(\text{Brent}_t - \text{Brent}_{t-1}) / \text{Brent}_{t-1})$$

$$FXD_t = \max(0, (\text{FX}_t - \text{FX}_{t-1}) / \text{FX}_{t-1})$$

$$INF_t = \text{CPI inflation}_t / 100$$

Here,  $OD_t$  measures the downside movement in the annual Brent price,  $FXD_t$  measures domestic-currency depreciation against the US dollar, and  $INF_t$  represents annual consumer-price inflation. Oil-price increases and currency appreciation are assigned a value of zero because this module is designed to monitor adverse pressure rather than symmetric macroeconomic variation. For a state-owned oil enterprise, falling oil prices may constrain revenue and cash generation, while domestic-currency depreciation may increase the local-currency burden of foreign-currency obligations and imported inputs. Inflation enters as a non-negative stressor because sustained price growth can increase operating and investment costs and weaken the purchasing power of nominal cash flows. These transformations represent a downside-monitoring convention rather than a universal structural model of the effects of oil prices, exchange rates, or inflation.

Annual averages were selected because they are consistently available from identified public sources throughout the observation period and correspond to the annual reporting frequency of the financial statements. They may nevertheless smooth concentrated intra-year shocks. The 2015 exchange-rate adjustment, for example, occurred through discrete devaluation episodes that are only partially represented by the annual-average depreciation measure. The present variables are therefore suitable for delayed annual monitoring but not for measuring the maximum intensity or exact timing of a shock. An operational higher-frequency system could supplement them with year-end exchange rates, maximum intra-year depreciation, monthly commodity-price drawdowns, or quarterly inflation measures.

### 3.5. Normalization, Weighting, and TOPSIS Aggregation

Because the indicators have different units and directions, they are transformed using min–max normalization. For a benefit indicator:

$$z_{it} = (x_{it} - \min(x_i)) / (\max(x_i) - \min(x_i))$$

For a cost indicator:

$$z_{it} = (\max(x_i) - x_{it}) / (\max(x_i) - \min(x_i))$$

In the rolling-origin analysis, the minimum and maximum values are calculated exclusively from the training period. Values in the held-out year are transformed using those historical parameters and clipped to the interval from zero to one. Consequently, information from the evaluation year cannot influence its own normalization. Clipping prevents a historically unprecedented observation from generating an unbounded normalized value and dominating the distance calculation. Its limitation is that observations extending beyond the same historical boundary become indistinguishable on that indicator. The robustness analysis therefore repeats the rolling-origin evaluation without clipping, allowing held-out observations to fall outside the unit interval while retaining the normalization parameters estimated exclusively from the training period.

A hierarchical weighting procedure is used to prevent a dimension containing several correlated indicators from receiving a disproportionately large total weight. Profitability, liquidity and cash generation, operating efficiency, and solvency each receive 25% of the internal-capacity weight. Equal dimension weighting is a transparent but normative design choice. The sensitivity analysis therefore assigns 40% to each dimension in turn and 20% to each of the other three dimensions to determine whether the shock-year classifications depend on the equal 25% allocation.

Within the two multi-indicator dimensions, CRITIC weights are used to allocate the corresponding 25%:

$$C_j = \sigma_j \sum_{k=1}^m (1 - r_{jk})$$

$$w_j = C_j / \sum_{j=1}^m C_j$$

Here,  $\sigma_j$  denotes the standard deviation of normalized indicator  $j$ , and  $r_{jk}$  is its correlation with indicator  $k$ . CRITIC assigns greater weight to an indicator when it provides both variation and information not duplicated by other indicators.

$$W_{\text{profitability}} = W_{\text{liquidity/cash}} = W_{\text{efficiency}} = W_{\text{solvency}} = 0.25$$

CRITIC is not used to redistribute weight across the four principal dimensions. This restriction improves interpretability and reduces the sensitivity of the small longitudinal sample to unstable correlations. Nevertheless, the within-dimension CRITIC weights may still vary across short expanding windows. The year-specific weights used for the 2015, 2020, and 2022 evaluations are therefore reported in the Results. The 2015 weights are estimated from only six prior annual observations and are interpreted as provisional calibration values rather than stable population parameters. The external-stress indicators receive equal weights of one third because oil-price decline, currency depreciation and inflation represent different shock channels whose relative importance may change across events.

The normalized decision matrix is multiplied by the indicator weights:

$$v_{it} = w_i z_{it}$$

The positive and negative ideal solutions are identified from the training observations:

$$A^+ = \{\max(v_i)\}, \quad A^- = \{\min(v_i)\}$$

The distances between observation  $t$  and the two ideal solutions are:

$$D_t^+ = \sqrt{[\sum_i (v_{it} - A_i^+)^2]}$$

$$D_t^- = \sqrt{[\sum_i (v_{it} - A_i^-)^2]}$$

The internal financial-capacity score is calculated as:

$$FC_t = D_t^- / (D_t^+ + D_t^-)$$

A higher value indicates stronger internal financial capacity. The same procedure is applied to the three adverse macroeconomic indicators to obtain the external-stress score  $ES_t$ , where a higher value indicates more severe external pressure. TOPSIS preserves the multi-dimensional decision structure and produces an interpretable relative-closeness score without estimating a large-parameter regression model from the small annual sample.

### 3.6. Warning Rules, Validation, and Data Confidence

For descriptive full-period classification, the median internal-capacity score and the 75th percentile of the external-stress score are used as the corresponding decision thresholds. The capacity median provides a transparent non-parametric division between relatively strong and weak internal financial conditions. The higher external-stress threshold is deliberately more selective because the stress module is intended to identify unusually adverse macroeconomic pressure rather than classify every above-median observation as a shock. The same asymmetric threshold rule is retained in the rolling-origin evaluation: the capacity threshold is the median of the prior-only capacity scores, whereas the external-stress threshold is the 75th percentile of the prior-only stress scores.

For out-of-sample evaluation, year  $t$  is classified using only information from years preceding  $t$ . The procedure is:

- Form the training sample from all observations before year  $t$ .
- Estimate normalization parameters from the training sample.
- Calculate hierarchical CRITIC weights from the training sample.
- Determine the training-sample capacity median.
- Determine the 75th percentile of the training-sample external-stress scores.
- Transform and score the held-out year.
- Assign its decision state without re-estimating the model using the held-out observation.

The resulting state represents a chronologically uncontaminated retrospective assessment of the held-out firm-year. Its practical availability remains subject to the publication date of the corresponding annual financial statements. Formally, the four-state rule is:

$$\text{State}_t = g(\text{ES}_t, \text{FC}_t; \tau\text{ES}, \tau\text{FC})$$

The event years were selected from externally observable macroeconomic and institutional conditions rather than from peaks or troughs in the composite scores. The year 2015 represents the joint oil-price decline and two-step depreciation of the Azerbaijani manat documented by the World Bank (2018). The year 2020 represents the pandemic-era disruption and sharp oil-price decline, while 2022 represents a different high-pressure environment characterized by elevated inflation and energy-market volatility. These three episodes were selected to examine whether the artifact could distinguish weak and comparatively strong internal financial configurations under different forms of external pressure. The resulting states are treated as documentary concordance checks rather than independently verified ground-truth labels. Accordingly, three concordant cases cannot be interpreted as an estimate of predictive accuracy.

The system distinguishes the estimated monitoring state from uncertainty arising from public-data limitations. Data confidence is summarized through four components: source authority, indicator coverage, accounting consistency, and provenance completeness. Source authority receives 1.00 when the observation is obtained from an audited financial statement or an identified official institutional dataset, 0.50 when it is obtained from an identifiable but non-audited corporate or secondary public source, and 0 when the source cannot be independently traced. Indicator coverage is the proportion of required quantitative inputs available for the relevant firm-year. Accounting consistency receives 1.00 when the reported values satisfy the applicable accounting-identity and comparative-period checks, 0.50 when a documented comparability issue remains only partially resolved, and 0 when an unresolved inconsistency affects the constructed indicator. Provenance completeness is the proportion of required metadata fields available, including the reporting entity, reporting year, source document or dataset, page, table or series locator where applicable, and transformation rule.:

The four components are combined with equal weights:

$$\text{DC}_t = (\text{SA}_t + \text{IC}_t + \text{AC}_t + \text{PC}_t) / 4$$

A value of 0.80 or above is labelled high confidence, a value from 0.50 to below 0.80 is labelled medium confidence, and a value below 0.50 is labelled low confidence. These cut-offs are transparent governance rules rather than empirically calibrated probabilities of correctness. Similarly, equal

component weighting is used only to provide a simple metadata summary and does not imply that the four dimensions have identical consequences for decision reliability. The confidence result accompanies the monitoring state but does not modify the financial-capacity or external-stress score.

## 4. Empirical Results and System Validation

### 4.1. Capacity Weights and Shock-Year Classifications

The hierarchical CRITIC procedure produced a balanced four-dimension structure. Operating efficiency and solvency each retained a dimension weight of 0.25. Within profitability, operating margin received a weight of 0.14 and return on assets received 0.11. Within liquidity and cash generation, the current ratio received 0.11 and operating cash flow to revenue received 0.14. The indicator weights sum to one and remain interpretable at both the indicator and decision-dimension levels.

Using all years only for descriptive visualization, the median capacity threshold was 0.5064 and the 75th-percentile external-stress threshold was 0.3787. Figure 1 maps all SOCAR firm-years against these two thresholds. This representation is useful for understanding the system's state logic, but it is not the principal validation design because the thresholds use information from the entire period.

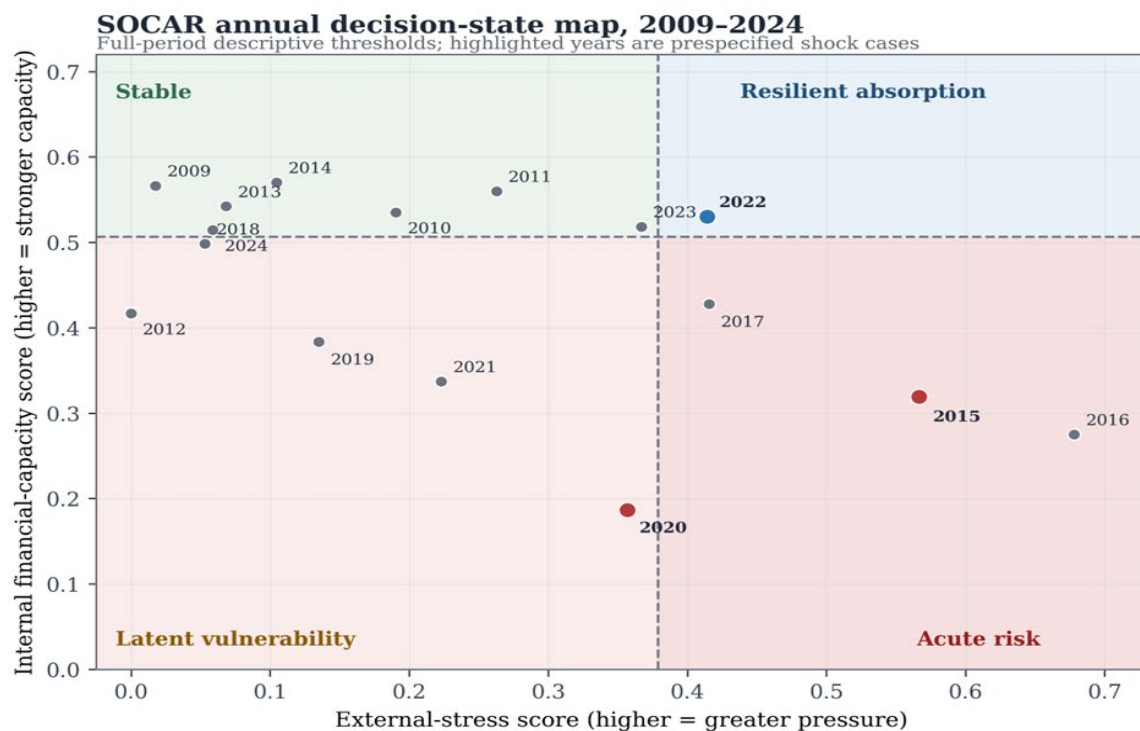


Fig. 1: SOCAR annual decision-state map, 2009–2024.

*Note.* The full-period descriptive threshold is the median for internal capacity and the 75th percentile for external stress. The shock-year classifications below use the same threshold definitions estimated exclusively from prior observations.

The primary validation asked whether the system could classify the prespecified 2015, 2020 and 2022 shock years using only information that would have been available before each target year. For every shock year, the normalization ranges, CRITIC weights, median capacity threshold, and 75th-percentile external-stress threshold were re-estimated from earlier observations. This prevents the target year's outcomes from determining its own warning boundary.

Table 4 SOCAR rolling-origin out-of-sample shock classifications

| Year | Capacity | Capacity Threshold | Stress | Stress Threshold | State                |
|------|----------|--------------------|--------|------------------|----------------------|
| 2015 | 0.3039   | 0.4754             | 0.6384 | 0.4739           | Acute risk           |
| 2020 | 0.1861   | 0.4941             | 0.3586 | 0.3565           | Acute risk           |
| 2022 | 0.5390   | 0.4478             | 0.4142 | 0.3586           | Resilient absorption |

*Note.* Each target year is evaluated using parameters estimated only from prior SOCAR observations. Capacity below its cut indicates weak internal capacity; stress above its cut indicates high external pressure.

The system classified 2015 as acute risk. The external-stress score of 0.6384 substantially exceeded its historically estimated threshold, while the capacity score of 0.3039 fell below the corresponding capacity threshold. The underlying records show that this combination was associated with a 47.1% annual Brent-price downside, 30.6% AZN depreciation, negative return on assets and a debt-to-assets ratio of 0.6334.

The 2020 observation was classified as acute risk under the principal specification, but this classification was marginal on the external-stress axis. Its out-of-sample stress score was 0.3586, only 0.0021 above the prior-only threshold of 0.3565. Internal capacity was unambiguously weak at 0.1861, reflecting a low operating margin, negative return on assets, a current ratio below one, and reduced operating cash flow to revenue. The evidence therefore supports a robust weak-capacity signal, while the distinction between acute risk and latent vulnerability depends on the selected external-stress threshold.

In contrast, 2022 was assigned to the resilient-absorption monitoring state. Inflation increased to 13.85%, pushing the stress score above its prior-only threshold, but the capacity score of 0.5390 also exceeded its capacity threshold. The classification is supported by stronger operating margin, return on assets, operating cash generation, asset turnover and a lower debt-to-assets ratio than in the acute-risk years.

### SOCAR rolling-origin shock-year validation

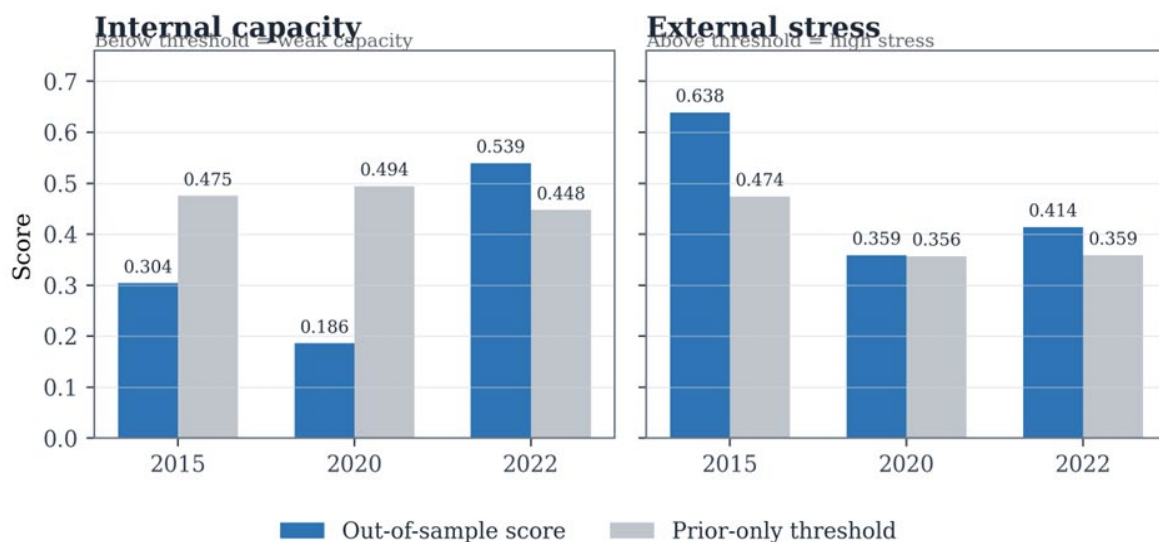


Fig. 2: SOCAR shock-year validation using prior-only thresholds.

These classifications are annual retrospective monitoring outputs rather than real-time warnings. In particular, the financial information used to classify 2020 became fully available only after the corresponding audited statements were released.

#### 4.2. Comparison with Standalone Indicators

The composite capacity score produced classifications concordant with the prespecified interpretation in all three cases: weak capacity in 2015 and 2020 and comparatively strong capacity in 2022. Operating margin, return on assets, operating cash flow to revenue, and debt to assets produced the same three-case pattern, while the current ratio and asset turnover were concordant in two cases. Each standalone capacity indicator was evaluated using its direction-specific prior-only median. The external-stress indicators were evaluated using their prior-only 75th percentiles. The benchmarks therefore followed the same temporal information restriction as the composite scores.

Table 5 Concordance of Composite and Standalone Benchmarks

| Benchmark Type    | Model or indicator                   | Correct cases |
|-------------------|--------------------------------------|---------------|
| Internal capacity | Hierarchical Decision-Support System | 3 / 3         |
| Internal capacity | Operating margin                     | 3 / 3         |
| Internal capacity | Return on assets                     | 3 / 3         |
| Internal capacity | Operating cash flow / revenue        | 3 / 3         |
| Internal capacity | Debt / assets                        | 3 / 3         |
| Internal capacity | Current ratio                        | 2 / 3         |
| Internal capacity | Asset turnover                       | 2 / 3         |
| External stress   | Composite External-Stress Score      | 3 / 3         |
| External stress   | Oil-price downside                   | 2 / 3         |
| External stress   | Exchange-Rate Depreciation           | 2 / 3         |
| External stress   | Inflation                            | 1 / 3         |

*Note.* Internal-capacity benchmarks use direction-specific medians estimated exclusively from observations preceding each event year. External-stress benchmarks use corresponding prior-only 75th percentiles. The results describe concordance across three prespecified documentary cases and are not estimates of population-level classification accuracy.

The comparison does not demonstrate superior event discrimination by the composite capacity score, because several standalone financial indicators produced the same three-case pattern. Its incremental value lies instead in the reproducible integration of multiple dimensions, the transparent reconciliation of conflicting signals, the identification of the dimensions driving each state, and the simultaneous representation of external pressure and internal financial capacity.

#### 4.3. Sensitivity Analysis

The full-period capacity ranking remained positively associated with the principal hierarchical specification under three alternative indicator-weighting methods. Spearman's rank correlation was 0.9176 for equal indicator weighting, 0.7912 for entropy weighting, and 0.7559 for flat CRITIC weighting. The lower correlation under flat CRITIC indicates that removing the conceptual dimension hierarchy materially affects the ordering of firm-years, although it does not by itself establish that either specification is superior. The rolling-origin CRITIC weights showed limited variation within the profitability and liquidity dimensions. For the 2015 evaluation, which used only six prior observations, the weights for operating margin, return on assets, current ratio, operating cash flow to revenue, asset turnover, and debt to assets were 0.1275, 0.1225, 0.1189, 0.1311, 0.2500, and 0.2500, respectively. The corresponding weights were 0.1367, 0.1133, 0.1107, 0.1393, 0.2500, and 0.2500 for 2020, and 0.1267, 0.1233, 0.1060, 0.1440, 0.2500, and 0.2500 for 2022. The estimates are reasonably similar across the three expanding windows, but the 2015 values remain provisional because correlations and standard deviations were estimated from only six annual observations.

Table 6 Sensitivity of Rankings and Shock-Year Classifications

| Sensitivity test    | Alternative specification | Spearman correlation | Shock-year classification result        |
|---------------------|---------------------------|----------------------|-----------------------------------------|
| Indicator weighting | Equal indicator weighting | 0.9176               | Not evaluated as a separate state model |

| <b>Sensitivity test</b> | <b>Alternative specification</b>                                      | <b>Spearman correlation</b> | <b>Shock-year classification result</b>                 |
|-------------------------|-----------------------------------------------------------------------|-----------------------------|---------------------------------------------------------|
| Indicator weighting     | Entropy weighting                                                     | 0.7912                      | Not evaluated as a separate state model                 |
| Indicator weighting     | Flat CRITIC weighting                                                 | 0.7559                      | Not evaluated as a separate state model                 |
| Dimension weighting     | Profitability assigned 40%; other dimensions 20% each                 | 0.9500                      | 3/3 states unchanged                                    |
| Dimension weighting     | Liquidity and cash generation assigned 40%; other dimensions 20% each | 0.9147                      | 3/3 states unchanged                                    |
| Dimension weighting     | Operating efficiency assigned 40%; other dimensions 20% each          | 0.5471                      | 3/3 states unchanged                                    |
| Dimension weighting     | Solvency assigned 40%; other dimensions 20% each                      | 0.8765                      | 3/3 states unchanged                                    |
| Indicator exclusion     | Return on assets removed                                              | 0.9882                      | Not separately examined                                 |
| Indicator exclusion     | Current ratio removed                                                 | 0.9882                      | Not separately examined                                 |
| Indicator exclusion     | Operating margin removed                                              | 0.8912                      | Not separately examined                                 |
| Indicator exclusion     | Operating cash flow/revenue removed                                   | 0.8824                      | Not separately examined                                 |
| Indicator exclusion     | Asset turnover removed                                                | 0.7735                      | Not separately examined                                 |
| Indicator exclusion     | Debt/assets removed                                                   | 0.5647                      | 3/3 states unchanged                                    |
| Profitability measure   | Net margin replaces operating margin                                  | 0.9794                      | 3/3 states unchanged                                    |
| Normalization           | Held-out values not clipped to [0,1]                                  | Not applicable              | 3/3 states unchanged                                    |
| Threshold variation     | Capacity threshold varied by $\pm 20\%$                               | Not applicable              | 3/3 states unchanged                                    |
| Threshold variation     | Stress threshold reduced by 20%                                       | Not applicable              | 3/3 states unchanged                                    |
| Threshold variation     | Stress threshold increased by 10%                                     | Not applicable              | 2/3 states unchanged; 2020 becomes latent vulnerability |

*Note.* Spearman correlations refer to full-period capacity-score rankings relative to the principal hierarchical specification. “3/3 states unchanged” means that 2015 remains acute risk, 2020 remains acute risk, and 2022 remains resilient absorption. Rank correlation is not applicable to threshold perturbations because these tests alter the classification boundary rather than the underlying continuous score.

Alternative dimension allocations did not change the three shock-year states. Assigning 40% to profitability, liquidity and cash generation, operating efficiency, or solvency in turn preserved the acute-risk classifications for 2015 and 2020 and the resilient-absorption classification for 2022. The full-period rank correlation was lowest when operating efficiency received 40%, indicating that dimension weights affect the ordering of ordinary firm-years more strongly than the three focal event classifications. Debt to assets was the most influential indicator in the exclusion analysis. Removing it reduced the full-period rank correlation to 0.5647. Nevertheless, after the remaining three dimensions were reweighted equally, 2015 and 2020 remained below their prior-only capacity thresholds and 2022 remained above its threshold. The influence of leverage therefore concerns the ordering and relative distance of firm-years rather than the qualitative classification of the three prespecified events.

Replacing operating margin with net margin produced a rank correlation of 0.9794 and also left all three event classifications unchanged. This result reduces, but does not eliminate, the measurement-equivalence concern created by the profitability substitution in the KMG application.

The event classifications also remained unchanged when held-out observations were transformed without clipping and when the capacity threshold was varied by up to 20% in either direction. The external-stress boundary was less stable. Raising it by 10% changed 2020 from acute risk to latent vulnerability, while raising it by 20% additionally changed 2022 from resilient absorption to stable. Reducing the stress threshold by 20% did not change the three classifications. These results confirm that the weak internal-capacity signal in 2020 is robust, but its designation as a high-stress rather than low-stress observation depends on a narrow threshold margin..

#### 4.4. Cross-Firm Instantiation with KazMunayGas

The cross-firm instantiation used KMG public data for 2018–2024 and retained the same four capacity dimensions, indicator directions, stress channels, hierarchical aggregation procedure, and four-state decision logic. Net margin replaced operating margin because a consistently comparable operating-profit series was unavailable across all KMG years. The locally estimated weights were 0.1188 for net margin, 0.1312 for return on assets, 0.1227 for the current ratio, 0.1273 for operating cash flow to revenue, and 0.2500 each for asset turnover and inverse debt to assets. These are locally calibrated KMG parameters rather than weights transferred directly from SOCAR.

The 2020 and 2022 observations were jointly excluded from KMG parameter estimation and were subsequently assessed against normalization ranges, weights, and thresholds derived from the five remaining years. Under this leave-event-out specification, 2020 was classified as acute risk and 2022 as resilient absorption. Because the calibration sample contains later non-event observations and only five training years, the result demonstrates that the architecture can be instantiated in a second enterprise. It does not constitute prospective external validation or evidence of stable predictive performance.

Table 7. KMG Leave-Event-Out Cross-Firm Instantiation

| Year | Capacity | Capacity Threshold | Stress | Stress Threshold | State                |
|------|----------|--------------------|--------|------------------|----------------------|
| 2020 | 0.3420   | 0.5604             | 0.5838 | 0.5099           | Acute risk           |
| 2022 | 0.6950   | 0.5604             | 0.5444 | 0.5099           | Resilient absorption |

*Note.* KMG financial records were derived from its official 2019, 2021, and 2024 annual reports. Both event years were excluded jointly, leaving five non-event observations for local calibration. Because the calibration sample includes observations occurring after the evaluated years, this is a cross-firm instantiation test rather than prospective external validation or a real-time forecast.

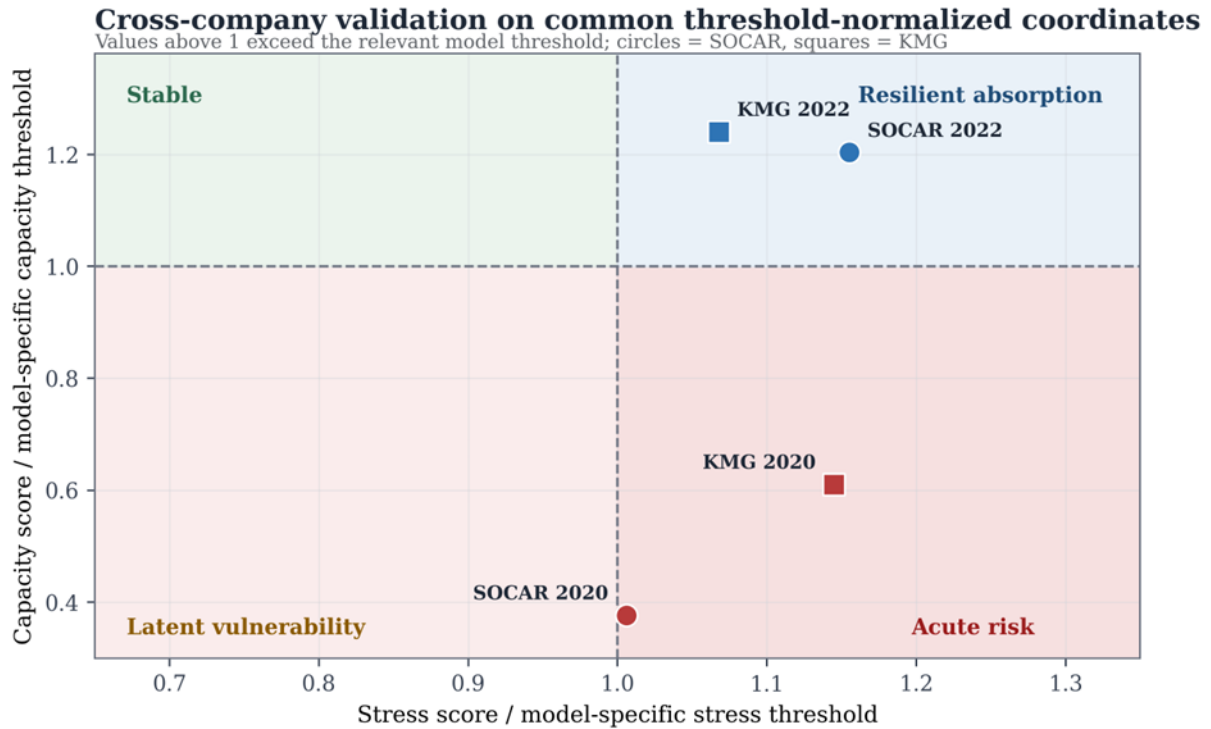


Fig. 3: Threshold-normalized cross-firm instantiation for SOCAR and KMG.

Figure 3 places the two firms on threshold-normalized coordinates while retaining firm-specific calibration. Both 2020 observations fall in the acute-risk quadrant, while both 2022 observations fall in the resilient-absorption quadrant. The qualitative replication supports the portability of the architecture, indicator directions, dimension structure, and decision-state logic. It does not demonstrate invariance of raw scores, fitted weights, thresholds, or accounting measures across firms. Each new implementation therefore requires firm-specific provenance checks, accounting reconciliation, and transparent documentation of any indicator substitution.

#### 4.5. Data Confidence and Explainable Quantitative Outputs

The three SOCAR event observations each received a data-confidence value of 1.00 and were therefore classified as high confidence under the prespecified rule. Source authority equalled 1.00 because the financial inputs came from audited consolidated statements and the macroeconomic inputs came from identified official datasets. Indicator coverage equalled 1.00 because all six capacity indicators and three external-stress indicators were available. Accounting consistency equalled 1.00 because the applicable accounting identities and comparative-period checks were satisfied. Provenance completeness also equalled 1.00 because each input was linked to an identified reporting period, source, locator, and transformation rule. These confidence values describe the traceability and completeness of the quantitative inputs; they are not probabilities that the monitoring states are correct.

Narrative evidence did not contribute to the event scores or classifications because a complete longitudinal corpus and independent reliability assessment were unavailable. Table 8 therefore reports quantitative drivers, provenance status, and data-confidence values only. The proposed narrative schema remains part of the future architecture, but no empirical claim about management preparedness is made from missing or selectively available disclosures.

Table 8. Explainable Quantitative Outputs for the SOCAR Event Years

| Year And State | Primary External Drivers | Primary Internal Interpretation | Provenance And Data Confidence            |
|----------------|--------------------------|---------------------------------|-------------------------------------------|
| 2015 — Acute   | Oil downside 47.1%; FX   | Negative Return on Assets       | Audited financial statements and official |

| Year And State              | Primary External Drivers                                 | Primary Internal Interpretation                                 | Provenance And Data Confidence                                                                                     |
|-----------------------------|----------------------------------------------------------|-----------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|
| risk                        | depreciation 30.6%                                       | and elevated leverage outweigh liquidity                        | macroeconomic series; complete quantitative provenance; DC = 1.00 (high)                                           |
| 2020 — Acute risk           | Oil downside 34.7%; stress narrowly above prior-only cut | Very low margin, negative ROA, Current Ratio Below 1.00         | Audited financial statements and official macroeconomic series; complete quantitative provenance; DC = 1.00 (high) |
| 2022 — Resilient absorption | Inflation 13.85%                                         | Profitability, turnover and solvency support capacity above cut | Audited financial statements and official macroeconomic series; complete quantitative provenance; DC = 1.00 (high) |

The analytical prototype was evaluated against the six design requirements derived in Section 2. Table 9 links each requirement to its implemented component, evaluation procedure, available evidence, and remaining limitation. The purpose is not to claim that every requirement has been fully validated, but to distinguish demonstrated properties from components that remain provisional.

Table 9. Evaluation of the Design Requirements

| Design Requirement                                                  | Implemented Component                                                                                                                            | Evaluation Method                                                              | Evidence                                                                                                                        | Remaining Limitation                                                                                                  |
|---------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| DR1: Separate internal capacity from external stress                | Independent capacity and stress scores combined through four decision states                                                                     | Rolling-origin classification of the 2015, 2020, and 2022 episodes             | The prototype distinguishes acute risk from resilient absorption and does not treat every high-stress year as financial failure | The evidence is limited to three prespecified SOCAR episodes                                                          |
| DR2: Preserve semantics, provenance, and evidence confidence        | Data dictionary, source records, accounting-reconciliation rules, narrative-evidence fields, and confidence metadata                             | Source and coverage audit for the evaluated firm-years                         | Financial and macroeconomic inputs are linked to identified sources and transformation rules                                    | Narrative coverage is incomplete, and the confidence rules require further operational validation                     |
| DR3: Use transparent hierarchical aggregation and expose components | Direction-aware normalization, equal dimension weights, within-dimension CRITIC weights, TOPSIS aggregation, and visible component contributions | Alternative weighting, threshold, normalization, and indicator-selection tests | The broad ordering is stable under several alternatives, but leverage and the 2020 stress threshold are influential             | Short rolling windows may produce unstable weights; additional robustness results are reported in Section 4.3         |
| DR4: Make classifications intrinsically interpretable               | Explicit state rules, threshold comparisons, driver decomposition, and evidence display                                                          | Inspection of the explainable outputs for the three SOCAR episodes             | Each state can be traced to its capacity, stress, and principal indicator contributions                                         | User comprehension and decision usefulness have not yet been evaluated with practitioners                             |
| DR5: Enforce prior-only temporal evaluation                         | Expanding historical windows for normalization, weighting, thresholds, and standalone benchmarks                                                 | Held-out evaluation of 2015, 2020, and 2022                                    | Each target year is classified without using later SOCAR observations                                                           | Annual audited data support delayed monitoring rather than real-time warning                                          |
| DR6: Distinguish portable architecture from local calibration       | Common dimensions, direction rules, state semantics, and evaluation logic with firm-specific parameters                                          | KMG leave-event-out cross-firm instantiation                                   | The qualitative 2020 and 2022 state pattern is reproduced in KMG                                                                | The KMG sample is short, later non-event years remain in calibration, and net margin substitutes for operating margin |

Table 9 shows that the strongest evidence concerns the separation of capacity and stress, prior-only evaluation, transparent state logic, and cross-firm instantiation. The weaker components are the stability of parameters in short rolling windows, empirical demonstration of narrative coding, confidence-score calibration, and user-oriented evaluation. These qualifications define the present contribution as an evaluated analytical prototype rather than a completed operational decision-support system.

## **5. Discussion**

### **5.1. Principal Findings and Their Interpretation**

This study developed and evaluated an analytical prototype that converts heterogeneous public information into two separate signals: internal financial capacity and external macroeconomic stress. Their combination produces four financial-resilience monitoring states. These states describe whether an observed financial configuration appears weak or adequate under the external conditions prevailing in a given year. They do not measure the broader organizational capabilities involved in anticipating, adapting to, or recovering from disruption. The distinction is important because a high-stress observation does not necessarily imply financial failure, while a favorable financial outcome does not by itself demonstrate organizational resilience.

The rolling-origin results for SOCAR show that this distinction is empirically consequential. Using only information available before each evaluated year, the system classified 2015 and 2020 as acute-risk states: internal capacity was below its prior-only threshold while external stress was above its prior-only threshold. In 2022, external stress again exceeded the applicable threshold, but internal capacity remained above its threshold, producing a resilient-absorption state. The substantive interpretation is not that 2022 was benign. Rather, the firm entered a high-pressure environment with a stronger observed capacity configuration than in the two earlier shock episodes. The different descriptive and rolling-origin classifications of 2020 reinforce the importance of temporal validity: only the latter approximates the historically available information set and avoids look-ahead contamination.

### **5.2. Informatics and Decision-Support Contribution**

The contribution is an analytical decision-support prototype composed of a versionable data pipeline, explicit semantic rules, a transparent scoring procedure, a state-classification layer, and provenance-aware outputs. It should not be interpreted as a completed software platform. No graphical interface, operational database, or user-interaction module was developed or evaluated. The contribution instead lies in specifying a reproducible workflow through which financial statements and macroeconomic series are reconciled, transformed, classified, and traced back to their underlying sources and assumptions.

The design separates quantitative measurement from incomplete narrative information. The reported scores are calculated only from longitudinally comparable financial and macroeconomic variables. The architecture specifies how management disclosures could later be stored with source and coding metadata, but the present study does not empirically evaluate that layer. Excluding incomplete narrative records prevents disclosure availability from being misinterpreted as evidence of organizational capability. A future implementation will require a complete longitudinal corpus and a reproducible coding process before narrative evidence can be incorporated into evaluated decision outputs.

Explainability is embedded at three levels. First, the score is decomposable into indicators and dimensions, allowing users to identify which capacity or stress components drive a state. Second, thresholds and state rules are explicit rather than hidden in an opaque prediction model. Third, data-

confidence metadata records source authority, temporal coverage, accounting consistency, and provenance completeness. The confidence assessment accompanies the decision state; it does not silently alter it. Users can therefore distinguish a concerning signal supported by strong evidence from a similarly concerning signal based on incomplete public information.

### **5.3. Incremental Value, Robustness, and Transferability**

The benchmark results require a deliberately qualified interpretation. The composite capacity score detected all three prespecified shock episodes, but operating margin, return on assets, and operating cash flow to revenue also achieved three of three under the same directional event test. The system should therefore not be presented as uniformly superior to every standalone ratio. Such a claim would be unsupported by the small event set and would obscure the artifact's actual value.

Its incremental value is instead structural and decision-oriented. A single ratio conveys one financial mechanism and can produce an isolated alert, but it does not jointly represent profitability, liquidity and cash generation, operating efficiency, solvency, and external pressure; specify how conflicting signals should be reconciled; or translate them into a stable set of interpretable decision states. The system preserves the underlying ratios while adding a transparent synthesis layer. It can show whether a high-stress year is accompanied by broad-based financial weakness or by an observed financial configuration that remains above its historically estimated capacity threshold. This makes the output suitable for monitoring, escalation, and diagnostic follow-up rather than merely retrospective performance description. The stress composite similarly integrates heterogeneous macroeconomic pressures without establishing predictive superiority from only three events.

The capacity ordering remained broadly stable across alternative weighting schemes, although the indicator-exclusion analysis identified leverage as the principal sensitivity. A practical implementation should therefore display the solvency contribution explicitly and require review when debt definitions or capital-structure policies change.

Transferability was examined with KazMunayGas (KMG), a second state-owned energy enterprise. Under a leave-event-out design, the portable system logic classified 2020 as acute risk and 2022 as resilient absorption, matching the qualitative pattern observed for SOCAR. This is evidence that the architecture, direction rules, hierarchical dimensions, and state logic can be instantiated beyond the development firm. It is not evidence that raw scores, fitted weights, or thresholds are universally invariant. Firm-specific histories, accounting choices, ownership arrangements, and macroeconomic settings remain relevant to calibration.

Transferability remains qualified because KMG net margin substitutes for SOCAR operating margin and the leave-event-out design uses non-event observations on both sides of the evaluated episodes. The test therefore supports cross-firm instantiation, not real-time predictive performance.

### **5.4. Managerial and Governance Implications**

For enterprise finance, risk, and strategy teams, the prototype provides a repeatable structure for delayed annual financial review using publicly auditable information. Its two-axis output supports a disciplined sequence of questions: whether external pressure is elevated, whether internal capacity is weak or strong relative to prior experience, which dimensions drive the classification, and whether the underlying evidence is complete and traceable. The resulting state can support escalation to more detailed treasury, operating, refinancing, or scenario analysis. It should not be interpreted as an automatic recommendation or as a substitute for information available to internal managers.

For state owners, public authorities, lenders, and external analysts, the artifact offers a common evidence structure for comparing annual signals while retaining source provenance. Because audited statements may be released several months after year-end, the current prototype supports ex-post or

delayed monitoring rather than contemporaneous crisis warning. A threshold crossing should prompt review rather than directly trigger a credit, investment, or policy action. More timely operational use would require quarterly or monthly financial, market, debt-maturity, and production data together with explicit controls over their publication dates. Parameter changes should also be governed through documented approval, version control, periodic recalibration, and expert challenge.

### **5.5. Limitations and Future Research**

The scope of the evidence remains limited. SOCAR provides sixteen annual observations, the focused validation covers three prespecified shock episodes, and KMG is the only external comparator. The findings support analytical claims about the usefulness and potential transferability of the artifact, but they do not permit population-level generalization, causal inference, or precise estimates of predictive sensitivity. Further studies could apply the architecture to a larger group of state-owned energy enterprises operating under different national and ownership arrangements. The study also does not include a user-oriented evaluation. No financial managers, state-owner representatives, lenders, regulators, or external analysts were asked to assess the comprehensibility, usefulness, or decision relevance of the four states, driver explanations, provenance records, or confidence metadata. The present evaluation therefore establishes computational reproducibility and limited analytical utility, but not usability or organizational decision impact. Future research should use structured expert review, task-based evaluation, or controlled user studies to examine comprehension, consistency of interpretation, decision time, perceived usefulness, and the handling of conflicting indicators.

The annual and consolidated nature of the public data also restricts the timeliness and depth of the analysis. A state assigned to year  $t$  should therefore be understood as an assessment of the financial condition reported for year  $t$  not necessarily as a warning delivered during year  $t$ . Warning signals may become available only after the reporting year, while consolidated statements provide little information about intra-year liquidity, differences among business units, contingent state support, contractual exposure, production disruption, or management responses during a shock. Accounting restatements, currency translation, and changes in disclosure practices create additional uncertainty. Provenance records and confidence metadata make these issues visible, but cannot remove them. Quarterly or monthly financial, market, and operational data could improve warning timeliness, provided that reporting dates and measurement definitions are documented consistently and future information is excluded from historical classifications.

The current indicator set does not capture every source of financial pressure. Internal capacity is represented by profitability, liquidity and cash generation, efficiency, and solvency. External stress is measured using oil-price downside, exchange-rate depreciation, and inflation. Interest rates, sanctions, fiscal transfers, regulatory changes, output disruptions, refinancing maturities, and political risk may also affect financial resilience. The indicator-selection protocol improves consistency and parsimony but cannot ensure that the six retained measures capture every financially relevant form of capacity. Management disclosures are retained as explanatory evidence rather than incorporated into the score because a complete and consistently coded longitudinal corpus is not yet available. Future extensions could test whether additional indicators improve out-of-sample performance and develop the narrative layer using a published coding protocol, independent coders or validated natural-language-processing procedures, and formal reliability assessment. A human-in-the-loop interface could also allow users to inspect the evidence, adjust assumptions, compare scenarios, and record the reasons for overriding a system-generated warning.

Model risk remains despite the use of transparent methods. Historical min-max normalization is sensitive to extreme observations, and median-based thresholds do not reflect the specific costs of false alarms and missed events. Alternative weights and indicator-exclusion tests show whether the results depend heavily on particular modeling choices, but they are not substitutes for prospective

evaluation. Future evaluations should consider warning lead time, decision quality, user comprehension, rating changes, refinancing stress, covenant pressure, liquidity interventions, and extraordinary state support. Versioned code, machine-readable data dictionaries, parameter records, and source-level provenance would support independent replication and help institutions use the system responsibly.

## 6. Conclusion

This study developed and evaluated an explainable analytical prototype for financial-resilience monitoring in state-owned energy enterprises. The artifact integrates audited financial statements, macroeconomic series, source provenance, and operationalized data-confidence metadata, while specifying but not empirically evaluating a separate management-disclosure evidence schema. Internal financial capacity and external stress are represented as separate axes and converted into four interpretable monitoring states through direction-aware normalization, hierarchical weighting, TOPSIS aggregation, and prior-only thresholds.

Rolling-origin evaluation classified SOCAR's 2015 observation as acute risk and its 2022 observation as resilient absorption. The principal specification also classified 2020 as acute risk, but its external-stress score exceeded the applicable threshold by only 0.0021 and changed to latent vulnerability when the stress threshold was increased by 10%. The weak internal-capacity signal for 2020 remained stable across the tested specifications. Although several standalone financial ratios matched the capacity composite across the three prespecified events, the prototype adds value by reconciling multiple indicators, identifying the drivers of each state, and preserving traceability from the output to source data and parameter choices. The qualitative 2020–2022 pattern was reproduced in a limited KMG cross-firm instantiation, showing that the architecture and decision logic can be applied beyond the development enterprise, although this exercise does not constitute prospective external validation.

The findings support a bounded proof of concept rather than causal, population-level, prospective-validation, or real-time forecasting claims. The resulting states describe reported financial condition under external pressure and should not be interpreted as comprehensive measures of organizational anticipation, adaptation, or recovery capability. Annual reporting frequency, the small SOCAR sample, three documentary event cases, one external comparator, and the absence of user-oriented evaluation constrain the claims. Nevertheless, the study establishes a reproducible and explainable monitoring architecture whose data transformations, parameters, thresholds, classifications, and provenance records can be inspected independently. Future research should evaluate the artifact prospectively, across larger multi-firm samples, with higher-frequency data, and through structured user studies.

## References

- Behzadian, M., Otaghsara, S. K., Yazdani, M., & Ignatius, J. (2012). A state-of-the-art survey of TOPSIS applications. *Expert Systems with Applications*, 39(17), 13051–13069. <https://doi.org/10.1016/j.eswa.2012.05.056>
- Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83. <https://doi.org/10.1016/j.csda.2017.11.003>
- Buneman, P., Khanna, S., & Tan, W.-C. (2001). Why and where: A characterization of data provenance. In *Database Theory—ICDT 2001* (pp. 316–330). Springer. [https://doi.org/10.1007/3-540-44503-X\\_20](https://doi.org/10.1007/3-540-44503-X_20)

- Debray, T. P. A., Vergouwe, Y., Koffijberg, H., Nieboer, D., & Steyerberg, E. W. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, 68(3), 279–289. <https://doi.org/10.1016/j.jclinepi.2014.06.018>
- Diakoulaki, D., Mavrotas, G., & Papayannakis, L. (1995). Determining objective weights in multiple criteria problems: The CRITIC method. *Computers & Operations Research*, 22(7), 763–770. [https://doi.org/10.1016/0305-0548\(94\)00059-H](https://doi.org/10.1016/0305-0548(94)00059-H)
- Duchek, S. (2020). Organizational resilience: A capability-based conceptualization. *Business Research*, 13(1), 215–246. <https://doi.org/10.1007/s40685-019-0085-7>
- Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, 14(4), 532–550. <https://doi.org/10.5465/amr.1989.4308385>
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Hillmann, J., & Guenther, E. (2021). Organizational resilience: A valuable construct for management research? *International Journal of Management Reviews*, 23(1), 7–44. <https://doi.org/10.1111/ijmr.12239>
- Hind, M., Wei, D., Campbell, M., Codella, N. C. F., Dhurandhar, A., Mojsilović, A., Natesan Ramamurthy, K., & Varshney, K. R. (2023). Provenance documentation to enable explainable and trustworthy AI. *Data Intelligence*, 5(1), 139–162. [https://doi.org/10.1162/dint\\_a\\_00119](https://doi.org/10.1162/dint_a_00119)
- Khan, M. H., Ahmed, J., Mughal, M., & Khan, I. H. (2023). Oil price volatility and stock returns: Evidence from three oil-price wars. *International Journal of Finance & Economics*, 28(3), 3162–3182. <https://doi.org/10.1002/ijfe.2588>
- Kroes, J. R., & Manikas, A. S. (2014). Cash flow management and manufacturing firm financial performance: A longitudinal perspective. *International Journal of Production Economics*, 148, 37–50. <https://doi.org/10.1016/j.ijpe.2013.11.008>
- Langley, A. (1999). Strategies for theorizing from process data. *Academy of Management Review*, 24(4), 691–710. <https://doi.org/10.5465/amr.1999.2553248>
- Naddyrah, N., & Rusmanto, T. (2024). Analyzing effect of sustainability disclosure on financial metric of state-run firms: Evidence from Indonesia. *Journal of Logistics, Informatics and Service Science*, 11(8), 179–192. <https://doi.org/10.33168/JLISS.2024.0811>
- Narayan, P. K., & Sharma, S. S. (2011). New evidence on oil price and firm returns. *Journal of Banking & Finance*, 35(12), 3253–3262. <https://doi.org/10.1016/j.jbankfin.2011.05.010>
- National Company KazMunayGas JSC. (2020). Annual report 2019. <https://ar2019.kmg.kz/en/>
- National Company KazMunayGas JSC. (2022). Annual report 2021. <https://ar2021.kmg.kz/en/>
- National Company KazMunayGas JSC. (2025). Annual report 2024. <https://ar2024.kmg.kz/en/>
- OECD & European Union. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing. <https://doi.org/10.1787/9789264043466-en>
- OECD. (2020). *Transparency and disclosure practices of state-owned enterprises and their owners*. OECD Publishing. <https://doi.org/10.1787/1ea99dbf-en>

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

State Oil Company of the Republic of Azerbaijan. (2010–2025). *Audited consolidated financial statements and corporate reports covering financial years 2009–2024*. <https://socar.az/en/page/financial-reports>

Tahir, M. S., & Richards, D. W. (2025). A systematic literature review of financial resilience: Antecedents, consequences and future research. *Journal of Financial Reporting and Accounting*, 33(4), 592–620. <https://doi.org/10.1108/JFRC-10-2024-0204>

Tascón, M. T., Castro, P., & Valdunciel, L. (2024). Effects of financial restrictions on firms' financial resilience against the COVID-19 pandemic: Evidence from the European hospitality industry. *Applied Economics*, 56(58), 8226–8241. <https://doi.org/10.1080/00036846.2023.2289951>

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–33. <https://doi.org/10.1080/07421222.1996.11518099>

World Bank. (2018, October 29). *Azerbaijan: Just-in-time support for the economy*. <https://www.worldbank.org/en/news/feature/2018/10/29/azerbaijan-just-in-time-support-for-the-economy>

Yang, H., Sutan Ahmad Nawi, H., & Zhang, Y. (2025). Artificial intelligence and large language models in government document management: A systematic review of applications, challenges, and implementation strategies. *Journal of Logistics, Informatics and Service Science*, 12(4), 129–145. <https://doi.org/10.33168/JLISS.2025.0408>