

*Research Article*

Segmentation of Manufacturing Enterprises by Operational–Financial Profiles Based on Digital Analysis of Corporate Data

Nikita Blagoy^{1*}, Nikolay Dmitriev¹, Olga Lavrova², Andrey Zaytsev¹, Dmitry Rodionov¹

¹Graduate School of Industrial Economics, Peter the Great St. Petersburg Polytechnic University, 29 Polytechnicheskaya Street, St. Petersburg 195251, Russian Federation

²Department of Economics, Faculty of Engineering and Economics, Belarusian State University of Informatics and Radioelectronics, 6 P. Brovki Street, Minsk 220013, Republic of Belarus

*Corresponding author: nikita.1915@bk.ru; Tel.: +79992262808

Abstract: Persistent operational-financial heterogeneity across manufacturing firms complicates peer benchmarking and digital investment program targeting. We segment Russian manufacturing enterprises using a reproducible pipeline that combines feature standardization, principal component analysis, k-means clustering, and panel-level label consolidation. The methodological framework integrates principal component analysis on standardized features and the k-means algorithm as tools of digital data analytics. The optimal number of clusters was determined by combining the elbow method, average silhouette, and Calinski–Harabasz and Davies–Bouldin indices. The cluster labels were consolidated across the observation horizon through modal membership, with tie-breaking based on the minimum average distance to the centroids. Five interpretable profiles were identified that differed in business scale, capital intensity, revenue and margin dynamics, leverage resilience, and liquidity. The final map contains five profiles ($n = 6,479$ firms): 45, 1,746, 993, 1,286, and 2,409 firms per cluster, differing in scale, capital intensity, profitability and turnover patterns, leverage tolerance, and liquidity. Across random restarts, agreement with the final partition is $ARI \approx 0.49$ and $NMI \approx 0.43$; the low mean silhouette is consistent with soft boundaries between adjacent profiles. To monitor managerial practices, we build a composite index from six indicators using intracluster min–max normalization and multi-year averaging, positioning each firm relative to its closest peers and highlighting reserves for productivity growth and faster capital turnover. Novelty lies in a fully reproducible segmentation workflow that integrates panel-level label consolidation and intracluster normalization for managerial benchmarking.

Keywords: Clustering method; Digital transformation; Industrial data analytics; Manufacturing industry; Principal component analysis

1. Introduction

Manufacturing firms differ sharply in terms of scale, capital intensity, and financial behavior, which distorts industry average benchmarking. A unified panel of corporate indicators enables reproducible segmentation into comparable peer groups and supports stability checks under real sector heterogeneity (Arruda et al., 2023; Glukhov et al., 2023). To avoid unstable partitions, unsupervised clustering provides an operational taxonomy of firms but requires a transparent data pipeline and explicit quality and robustness diagnostics. Standardized pre-processing, dimensionality reduction, and multi-metric validation make the resulting profiles suitable for managerial benchmarking and for designing targeted industrial support measures (Pittino et al., 2020). Digital technologies—including elements of gamification as managerial mechanisms—act as a bridge between technological change and operational practice: improved

algorithms for analyzing production flows increase accuracy in detecting local defects and robustness to noise, directly affecting quality, process reliability, and real-time managerial decisions (Wang et al., 2025). Taken together, these considerations motivate cluster-based segmentation on a unified digital panel of indicators and subsequent profile monitoring to target digital initiatives.

Previous studies often segment firms using cross-sectional snapshots, report limited diagnostics beyond a single internal index, and rarely address profile instability when a company's position changes over time. In addition, benchmarking scores are frequently computed on pooled samples, which mixes heterogeneous peer sets and weakens managerial interpretability. The present study addresses these limitations using panel-level label consolidation, reproducible clustering with robustness reporting, and peer-normalized monitoring.

Manufacturing enterprises exhibit persistent heterogeneity in scale, capital intensity, and operational–financial behaviour, which complicates benchmarking and weakens the effectiveness of uniform digital and managerial interventions. Addressing this heterogeneity requires a segmentation framework that is both empirically grounded and reproducible on panel data, allowing firms to be compared within relatively homogeneous peer groups rather than against aggregated industry averages.

In this context, the study integrates an industry-level cluster map with a quantitative assessment of managerial practices. Such integration transforms segmentation from a descriptive classification into an operational tool that supports targeted training, motivation schemes, and resource allocation across digital initiatives while accounting for sectoral diversity.

The empirical focus is on Russian Federation manufacturing enterprises, represented by a longitudinal panel of firms described by a unified set of operational and financial indicators with de-identified company codes. This data structure ensures comparability across firms and years and provides a stable basis for clustering and productivity analysis.

Methodologically, a reproducible framework for data preparation and segmentation that combines preprocessing and standardization, multi-criteria selection of the number of clusters, consolidation of firm-level labels over time, and the construction of a feature-based aggregate index with within-cluster normalization is developed. Such a design reduces the sensitivity to short-term fluctuations and enhances the interpretability of the resulting profiles.

The analytical objective is to empirically validate this methodology on panel data and to demonstrate its applicability for monitoring operational–financial profiles. This study addresses three sequential tasks: preparation of an informative panel dataset, construction and validation of a clustering model with consolidated firm membership, and development of an aggregate index for peer-based comparison with productivity and financial performance indicators.

2. Literature Review

2.1 Unsupervised Clustering and Manufacturing Data Analytics

Manufacturing enterprise segmentation relies on unsupervised procedures that reduce sensitivity to initialization and noise. Enhanced k-means variants with symbiotic search heuristics improve seed selection and ensure stable convergence on large industrial corpora; in parallel, "improved" k-means versions help identify managerial control factors and enable cluster assignment sensitivity analysis (Ikotun & Ezugwu, 2022; Lu et al., 2021). In big-data settings, balancing computational cost and partition quality is essential: scalable initialization and cost-aware stopping criteria sustain enterprise segmentation reproducibility at the industrial scale (Pérez-Ortega et al., 2018).

In service-oriented manufacturing, matching capacities and capabilities are addressed via spectral clustering followed by service ranking (an improved Slope One), which accelerates top-k selection by functional similarity. At the resource level, clustering based on manufacturability features groups equipment into coherent sets aligned with routing rules and capacity planning (Hu et al., 2023; Zhao et al., 2022). The practical value of cluster descriptors also appears in

delay prediction: data mining pipelines achieve high F-scores and convert cluster characteristics into actionable features for scheduling and maintenance (Kannan et al., 2022).

Synthetic datasets compensate for gaps when real data are scarce or sensitive: systematic frameworks for generation and validation preserve decision boundaries and cluster structure when testing algorithms without exposing proprietary records (Buggineni et al., 2024). For dynamic operations, unsupervised learning extends to sequences and events: multidimensional time-series clustering groups traces from production and energy meters, supporting the shift from preventive to condition-based maintenance (Giordano et al., 2021b). In product flows, mode-based binary search isolates item clusters and helps detect regime shifts during changeovers, thereby reducing quality losses. Density-aware methods further increase alarm robustness under noise (Blachowicz et al., 2025; Herrero et al., 2023).

2.2 Network and Cluster Spatial Context

Beyond the shop floor, k-means is used to typologize inspection regimes, reducing subjectivity in compliance assessment. Combining PCA with hierarchical clustering yields more representative cycles for hybrid platforms than templates based on "plain" k-means (Hou et al., 2022; Wang et al., 2023). Selecting the number of clusters is best grounded in a combination of diagnostics (elbow, silhouette, gap) and harmonized K-selection methods that are stable across benchmarks and industrial datasets (Yuan & Yang, 2019). In the Industry 4.0 context, unsupervised identification/characterization of unknown modes supports anomaly detection and automatic labeling of states in complex production systems (Molinié et al., 2022).

Flexible scheduling for cyber-physical systems, leveraging IoT connectivity and analytics, reduces makespan, tardiness, and resource waste under mass customization and frequent disturbances; metaheuristic surveys confirm the advantages of hybrid search with decomposition and learning-guided operators under makespan–energy–cost criteria (Mourtzis, 2022; Zhang et al., 2025). In process industries, hybrid modeling (first principles + data-driven components) is a key mechanism of the digital transition, retaining interpretability while achieving high predictive accuracy. Synthetic-data architectures that preserve cluster structure and classification boundaries support analytics at scale (Chen et al., 2023; Gargalo et al., 2024). Expert studies in Logistics 4.0 emphasize the contribution of sensing and analytics to responsiveness and transparency, while educational simulators emphasize modular scenarios and experiment tracking for prototyping segmentations and "what-if" schedules (Chiang et al., 2021; Kılıç & Nielsen, 2022). Methodological work on hierarchical clustering of process models complements this picture by ensuring comparability and interpretability when reconstructing production schemes (Ahn & Chang, 2019).

Extra-plant risk monitoring and adjacent domains demonstrate the transferability of k-means: from electrical-equipment identification in smart buildings to industrial safety systems based on wireless sensor networks, thereby reducing subjectivity and formalizing control typologies (Seo et al., 2022; Zhang et al., 2020). The formation of driving cycles for hybrid platforms benefits from the combination of principal component analysis with hierarchical clustering, producing more representative trajectories in mixed urban traffic than templates based on "plain" k-means (Wang et al., 2023). In the Industry 4.0 setting, unsupervised identification and characterization of previously unknown data-supply modes support early anomaly detection and auto-labeling of states in complex production systems (Giordano et al., 2021a).

2.3 Operational Monitoring, Institutions, and Digital Transformation Financing

The analytics layer that underpins our clustering must be comprehensive and resource-aware—i.e., it should account for where and how data are captured, how they circulate across shops and firms, and which managerial decisions they are meant to support. Network community-detection methods on multilayer weighted supply-chain graphs help reveal hidden connectivity among sites, product families, and routes, thereby sharpening the boundaries of

enterprise clusters and isolating risk "communities" for coordinated inventory and scheduling policies (Zhu et al., 2024). Spatial industrial clusters have been shown to drive sustainable regional growth at the macro level, implying that a firm-level cluster typology is useful not only for within-firm optimization but also for aligning industrial policy and support measures (Chen et al., 2020).

Even in adjacent domains—smart buildings and industrial safety—k-means and related methods have demonstrated practical value: characteristic load profiles and sensor signals enable automatic identification of electrical equipment and anomalies, which transfers to production lines as fine-grained segmentation of assets, nodes, and monitoring zones (Seo et al., 2022; Zhang et al., 2020). A network view of economies and business ecosystems complements firm-level segmentation: the nested structure of urban business clusters explains why "clusters within clusters"—from product niches to technological specializations—frequently appear in sectoral panels (Cottineau & Arcaute, 2020). Models of land-use efficiency in industrial clusters further show how infrastructural constraints and density shape enterprise profiles and capital intensity—an important context for interpreting our "heavy" versus "light" manufacturing clusters (Cui et al., 2024). Finally, links between Industry 4.0 industrial policy, ESG practices, and total factor productivity suggest that cluster membership can guide the targeting of digitalization and sustainability programs: different clusters exhibit distinct return trajectories to "green" and digital initiatives (Liu & Ren, 2025).

Beyond the firm, spatial industry clusters and their institutional context are pivotal because they host the knowledge transfer, standardization, and joint investment mechanisms that materially influence productivity within our clusters. A recent review of sustainable manufacturing underscores the role of cluster platforms as an Industry 4.0 policy instrument: they synchronize digitalization, environmental objectives, and operational practices, thereby boosting firm-level returns and easing the diffusion of best practices across cluster participants (Scharmer et al., 2024). Simulation "sandboxes" that integrate sustainable/circular manufacturing tools with scenario modeling are useful to test segment-specific managerial interventions before they are rolled out on the shop floor (Khan & Abonyi, 2022).

Interactions between firms and counterparties within an industrial cluster can be formalized as a supplier/partner selection game under uncertainty. Such models align risk, pricing, and lead times, thereby improving supply chain resilience across our cluster profiles (Zaytsev et al., 2024). Financing sources for digital initiatives are institution-dependent: empirical evidence shows that digital technologies and capital markets reinforce each other, easing access to modernization finance in developing economies—an argument for cluster-targeted support instruments (Sandoyan et al., 2025). For monitoring macro- and meso-trends, short-term forecasting methods under digital transformation help steer cluster programs and capture structural breaks in our corporate panels (Eremina & Rodionov, 2023).

Cluster policy inevitably affects labor. Assessments of AI's impact on sectoral labor-market scenarios indicate which clusters require the largest investments in upskilling and reskilling to benefit from automation without employment losses (Faizullin et al., 2025). Finally, for energy planning it is advisable to embed industrial clustering explicitly into multi-node energy-system models—this realistically represents the demand, flexibility, and decarbonization potential of manufacturing nodes and helps align energy and production strategies for different enterprise types within our clusters (Dhondt et al., 2024). In summary, our methodology is strengthened when firm-level clustering by operational-financial profiles is complemented by institutional-spatial analysis, policy simulation, and the incorporation of financial, labor, and energy channels—turning cluster profiles into actionable units for coordinating digital investment and sustainable development.

2.4 Cluster Validation, Reproducibility, and Analytics for Digital Maturity Profiles

The structure and governance of industrial clusters are critical for collaboration and risk management: institutional and geographic proximity facilitate knowledge exchange, coordina-

tion, and shared-value creation, which directly enhances firm-level productivity within clusters and accelerates the diffusion of digital practices (Chen et al., 2022; Shirasawa & Seo, 2025). A complementary network perspective shows that the nested structure of business clusters helps identify cores and peripheries of interactions among firms and related activities, which is useful for targeting support measures and allocating infrastructure investments (Cottineau & Arcaute, 2020). Automated pipelines for spatiotemporal data collection and processing provide the technological backbone for sustained cluster monitoring and for evaluating the effects of digitalization at both enterprise and territorial scales—from sustainability indicators to operational-financial metrics (Zaytsev & Dmitriev, 2025).

At the operational level of manufacturing networks, time-series clustering and anomaly detection offer “early-warning” capabilities: they group similar process trajectories, flag real-time deviations, and thereby reduce operational risks within clusters of firms sharing comparable profiles (Giordano et al., 2021b). Dimension reduction and hierarchical clustering—proven effective in engineering tasks such as constructing representative duty/driving cycles—transfer to the segmentation of assets and production lines, helping align operating modes, maintenance schedules, and constrained resources (Wang et al., 2023). Effective cluster governance depends on resilient financial circuits and investment rules: regional heterogeneity in fixed-capital investment drivers necessitates differentiated modernization and digitalization instruments by cluster type (Dmitriev et al., 2022). In parallel, corporate social responsibility practices act as a driver of capability accumulation and intellectual potential, strengthening the “soft infrastructure” of clusters and reducing cooperation risks (Dmitriev & Satmurzaev, 2025). Taken together, these results motivate our strategy: to improve supply resilience, coordination quality, and returns on digital investment, firm-level clustering by operational-financial profiles should be complemented by interaction-aware network analysis, automated spatiotemporal data processing, and institutional governance settings.

3. Methods

3.1 Data Collection and Preprocessing

The study follows a reproducible workflow: data cleaning and harmonization, missing-value handling and robust tail treatment, feature standardization, principal component analysis (PCA)-based dimension reduction, clustering and model selection, validation and robustness checks, firm-level label consolidation over the panel, and construction of a peer-normalized monitoring index.

We harmonize firm–year reports into a unified panel and exclude indicators with missingness of more than 70%. For skewed non-negative variables, we apply $\log(1 + x)$; extreme values are winsorized at the 0.5th and 99.5th percentiles. Column medians impute missing entries, and all retained features are standardized to zero mean and unit variance. Detailed diagnostics and plots of missingness are presented in Figure 1.

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (1)$$

where x_{ij} is the value of feature j for observation i (company–year); μ_j is the mean of feature j ; σ_j is the standard deviation of feature j ; and z_{ij} is the standardized value.

3.1.1 Rationale for the feature blocks

The baseline indicators are grouped into six blocks reflecting standard corporate finance and operations logic: (1) scale and resources, (2) operating efficiency and profitability, (3) working capital turnover and operating-cycle speed, (4) capital structure and debt burden, (5) liquidity and short-term solvency, and (6) investment and innovation activity. This grouping aligns the segmentation with managerial levers (e.g., cost/margin control, cycle-time reduction, leverage policy, and investment capacity) and improves the interpretability of cluster profiles.

3.2 Dimensionality Reduction

The baseline PCA threshold was set at 85% cumulative explained variance, and the resulting number of components was reported. The sensitivity checks repeat the full clustering workflow for thresholds of 70%, 80%, and 90% to verify that the main profiles and their interpretation remain stable. The stability of component loadings is assessed via bootstrap resampling of firm-year records.

$$\text{EVR}_k = \frac{\lambda_k}{\sum_{j=1}^p \lambda_j} \quad (2)$$

where λ_k is the eigenvalue of the covariance matrix for component k ; p is the number of features; and EVR_k is the explained-variance ratio of component k .

3.3 Clustering

Clustering is performed using k-means (k-means++ initialization) either on standardized features or in the PCA space. The optimal k is selected using a suite of diagnostics: elbow (inertia), mean silhouette, Calinski–Harabasz, and Davies–Bouldin indices. We use multiple random starts (*n_init*) and an iteration cap (*max_iter*) to improve stability and validate solution consistency across distance metrics (Euclidean/cosine). At the company level, we consolidate membership using the modal cluster label across years and resolve ties by assigning the firm to the cluster whose centroid has the smallest mean distance from its annual observations.

$$\min_{\{C_1, \dots, C_K\}} \sum_{k=1}^K \sum_{i \in C_k} \|y_i - \mu_k\|_2^2 \quad (3)$$

where C_k is the set of observations in cluster k ; y_i is the feature vector of observation i (original or PCA space); μ_k is the centroid of cluster k ; $\|\cdot\|_2$ is the Euclidean norm; and K is the number of clusters.

As a robustness comparison, we also cluster in the PCA space and on the original standardized indicators, reporting agreement between solutions via ARI/NMI. We further compare k-means against Gaussian mixture models and Ward hierarchical clustering to assess whether overlapping profiles are better separated under alternative assumptions.

3.4 Assessment of Quality and Robustness

We examine the silhouette distribution by cluster and compute ARI/NMI over repeated initializations, the inter-centroid distance matrix, and outlier diagnostics via the Mahalanobis distance using Ledoit–Wolf covariance shrinkage (thresholded by the χ^2 quantile with degrees of freedom equal to the feature-space dimension). Structural robustness is tested by recomputing the silhouette after excluding outliers and by “rolling” the ARI over time to track the profile drift. As a sensitivity check for boundary overlap, we also fit Gaussian Mixture Models (GMMs) as an alternative.

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (4)$$

where s_i is the silhouette of observation i ; a_i is the mean intra-cluster distance for i ; and b_i is the smallest mean distance from i to another (best neighboring) cluster.

3.4.1 Computational efficiency

All experiments were conducted with fixed random seeds. We report wall-clock time for preprocessing, PCA fitting, and clustering across $k = 2, \dots, 10$, together with the software stack and hardware specification, to support scalability assessment.

3.5 Visualization and Interpretation of Data

We present heatmaps of cluster-wise means of standardized features (with 95% bootstrap CIs) and two-dimensional projections (PCA/t-SNE/UMAP) with trustworthiness to monitor the preservation of local structures. For practical interpretation, "representative" firms (minimum distance to the centroid) and back-transform centroids to original units (inverse standardization) are identified, which facilitates alignment with operational-financial KPIs and the derivation of actionable recommendations.

4. Results and Discussion

4.1 Study Data and Pipeline Processing

The empirical analysis uses corporate reports of Russian manufacturing enterprises from the Spark Interfax database for 2015–2024. After filtering and harmonization, the panel contains 6,479 firms and 29,460 firm–year observations. Indicator semantics are fixed in a codebook to ensure consistent interpretation across years. Preprocessing includes basic cleaning and type casting, conversion to a numeric format, missingness screening, robust tail treatment where needed, and feature standardization. Then, the workflow proceeds to PCA-based dimension reduction, clustering (k-means and auxiliary checks), validation, and firm-level label consolidation, followed by index construction for monitoring.

The processing pipeline includes the following steps:

1. Type casting and cleaning, removal of non-printable characters and non-breaking spaces, and unification of field names (Company, Year, X1...X146).
2. Conversion of values to a numeric format and labeling of incorrect entries as missing. Removal of rows without a year and rows with empty features.
3. Computation of the share of missing values for each indicator and exclusion of features with a missing-value share greater than 70%, with the excluded features retained in an audit list.
4. Calculation of summary statistics, including the mean, standard deviation, minimum, and maximum, and construction of diagnostic plots. These plots include a histogram of the number of non-missing features per company–year pair and a horizontal bar chart of missing-value shares for the 10 most problematic indicators.
5. Feature standardization and, where required, robust tail treatment and power transformations to reduce distributional skewness.

The subsequent stages comprise dimensionality reduction (PCA), clustering (k-means and hierarchical schemes), validation of partition quality by the silhouette, the Calinski–Harabasz index, and the Davies–Bouldin index, agreement of partitions by ARI and NMI, covariance estimation by Ledoit–Wolf, and construction of two-dimensional projections (t-SNE and UMAP) with control of local topology preservation (trustworthiness). The result is a reproducible pipeline—from cleaning and normalization to projections and clusters to quality checks and artifact storage—that secures comparability across companies and years and provides the technical basis for computing aggregate indices, productivity, and the effect of digital practices with cluster adjustment.

4.2 Assessment of Data Completeness and Robust Preprocessing

Missingness is evaluated under two conventions: treating zero values as valid observations and treating zeros as missing for indicators where zeros typically encode the absence of reporting rather than real magnitudes. Figure 1(a-d) summarizes the missingness patterns. Indicators

with missingness above the adopted threshold are excluded, and remaining gaps are imputed using a robust rule consistent with the preprocessing design. The pipeline applies sequential normalization steps to improve cross-feature comparability and reduce sensitivity to extreme values, combining standardization with robust tail trimming and power transformations for strongly skewed variables.

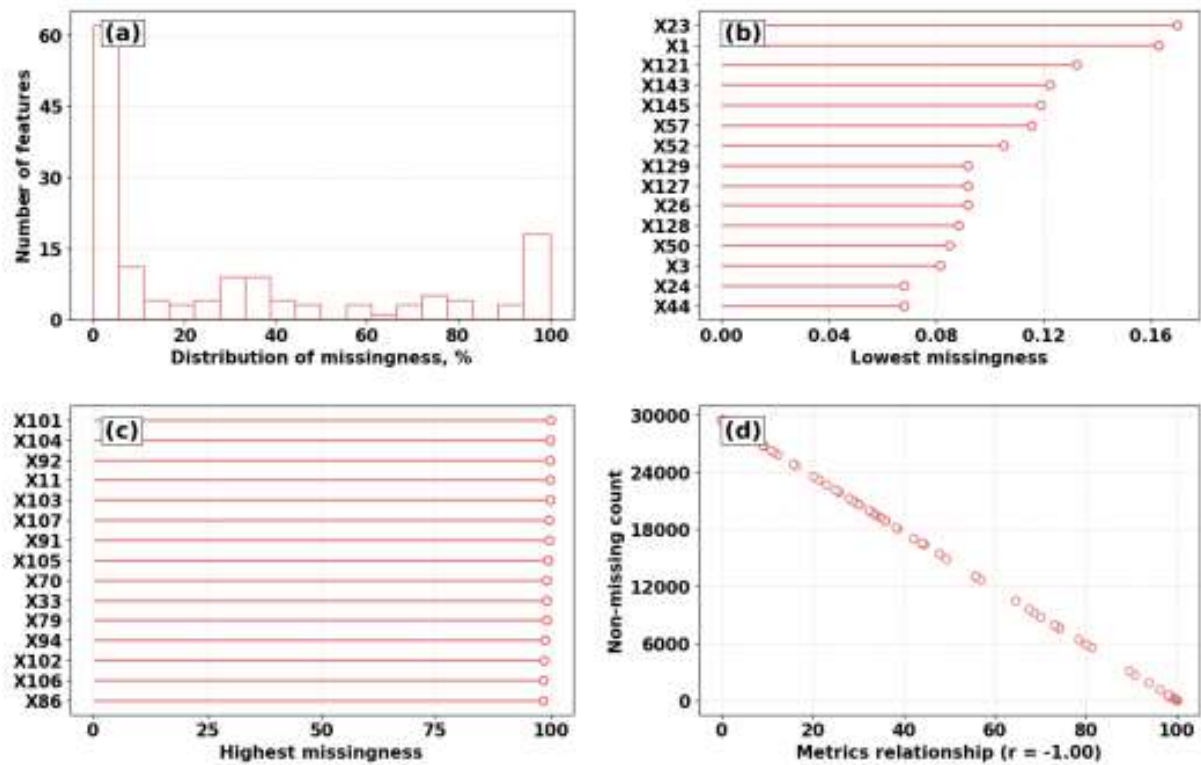


Figure 1 Missing values distribution across corporate indicators in the initial dataset (a) Distribution of missingness (b) Lowest missingness (c) Highest missingness (d) Metrics relationship

A sequential procedure is applied to enhance robustness to outliers and ensure comparability of scales. First, features are identified in which $>95\%$ of the values are non-negative and pronounced right skewness is present. A log transformation of the form $\log(1 + x)$, which shortens the distribution tails, is applied. Robust tail trimming is then performed for extreme observations. Values below the 0.5th percentile and above the 99.5th percentile are replaced with the boundary values, reducing the influence of extremes without dropping rows. We impute missing values with the median of each feature because this statistic is robust to skewed distributions. The final step is standardization by subtracting the mean and dividing it by the standard deviation, after which all features have zero mean and unit standard deviation. The output is a feature matrix that is ready for factor decompositions and clustering without scale distortion and the undue impact of outliers.

4.3 Choice of the Number of Clusters k

We select k by combining inertia, Calinski–Harabasz, Davies–Bouldin, and silhouette trends (Figures 2(a-b) and 3). The criteria favor a parsimonious partition with interpretable separation; $k = 5$ balances stability and economic readability, whereas larger k fragments adjacent profiles without improving robustness.

4.4 Final Labeling and Label Consolidation

Clustering is performed on the standardized feature space. For each firm–year record, the algorithm stores the assigned cluster label along with distances to all centroids. Firm-level

profiles are consolidated over the full horizon because yearly assignments may vary for a subset of firms: the modal label is selected, and ties are resolved using the smallest mean distance to the corresponding centroid. The resulting firm distribution by cluster is presented in Table 1.

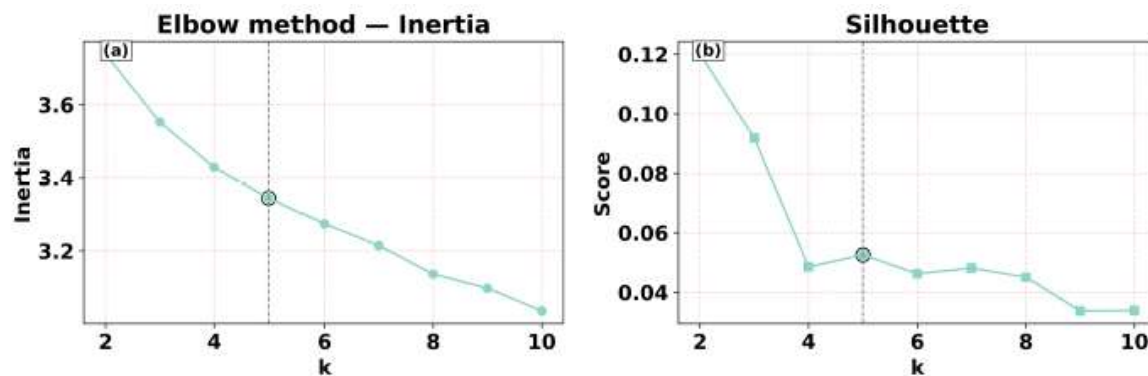


Figure 2 (a) Elbow and (b) mean silhouette criteria used to select the number of clusters

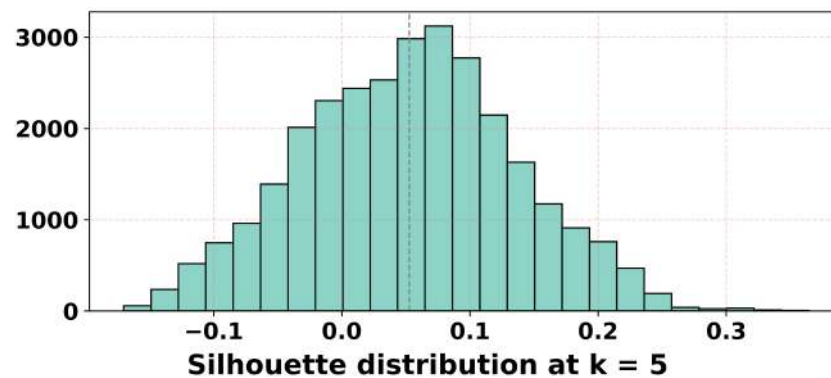


Figure 3 Distribution of individual silhouette values for the five-cluster solution selected

Table 1 Distribution of manufacturing enterprises across clusters

Cluster	Number of Unique Companies
1	45
2	1,746
3	993
4	1,286
5	2,409

4.5 Cluster Profiles and Their Interpretation

The segmentation summarizes manufacturing heterogeneity along six economic dimensions: scale/resources, operational efficiency, profitability, working-capital cycle and turnover, capital structure/debt burden, and liquidity/short-term solvency. The final industry map contains the following five interpretable profiles (Table 2):

- **Cluster 1** (45 firms): a stressed profile with weak profitability and frequent losses, primarily used for monitoring atypical financial patterns and potential restructuring trajectories.

- **Cluster 2** (1,746 firms): stable mid-sized producers with moderate margins and balanced operating–financial behavior, forming a broad middle segment.
- **Cluster 3** (993 firms): growth-oriented firms with accelerating revenue and margins and stronger dynamics in key performance measures, representing a narrower group with higher management intensity requirements.
- **Cluster 4** (1,286 firms): capital-intensive enterprises with large asset bases and long cycles, often linked to long-horizon investment regimes.
- **Cluster 5** (2,409 firms): small and microenterprises with low turnover and short operating cycles, forming the largest group due to the sector structure.

Table 2 Operational–financial profiles characterizing the identified clusters

Cluster	Brief profile description	Reasons for group size	Number of companies
1	Stressed profile with low profitability, frequent net losses, a high share of short-term liabilities, and slow turnover.	Extreme configuration is rare; some firms exit the market or move to other groups after restructuring.	45
2	Stable mid-sized producers with moderate margins, normal inventory and receivables turnover, a balanced equity-to-debt mix, and moderate leverage.	A common business model forms a broad “middle” of the distribution.	1,746
3	Growth profiles with accelerating revenue and margins, active investment in fixed and intangible assets, staff expansion; debt sometimes above average with acceptable interest coverage.	Require access to financing and strong growth management, so the share of such companies is limited.	993
4	Capital-intensive enterprises with large asset bases, long operating cycles, substantial leverage, and subdued current profitability due to depreciation and interest expenses.	Strategy is available to a limited set of firms; many projects have long payback periods.	1,286
5	Small and micro-enterprises with low turnover, short cycles, modest margins, and limited investment.	An industry structure dominated by small business creates the largest group.	2,409

Cluster size differences reflect the underlying population structure, sectoral composition, and firm-level consolidation rule, which reduces fragmentation from short-lived year-to-year fluctuations.

4.6 Feature Selection and Factor Projections

After cleaning and standardization, feature selection was performed to preserve the five-cluster structure and facilitate interpretation. The first step used the full set of indicators covering scale and resources, profitability, capital turnover, capital structure, liquidity, and the digital and innovation block. Such a full matrix provides a detailed picture but contains many weakly related and narrowly specialized dimensions. As a result, the first two principal components jointly explain only 22.99% of the variance. Numerous minor sources of variation, duplicated features, and local noise obscure the underlying cluster structure (Figure 4).

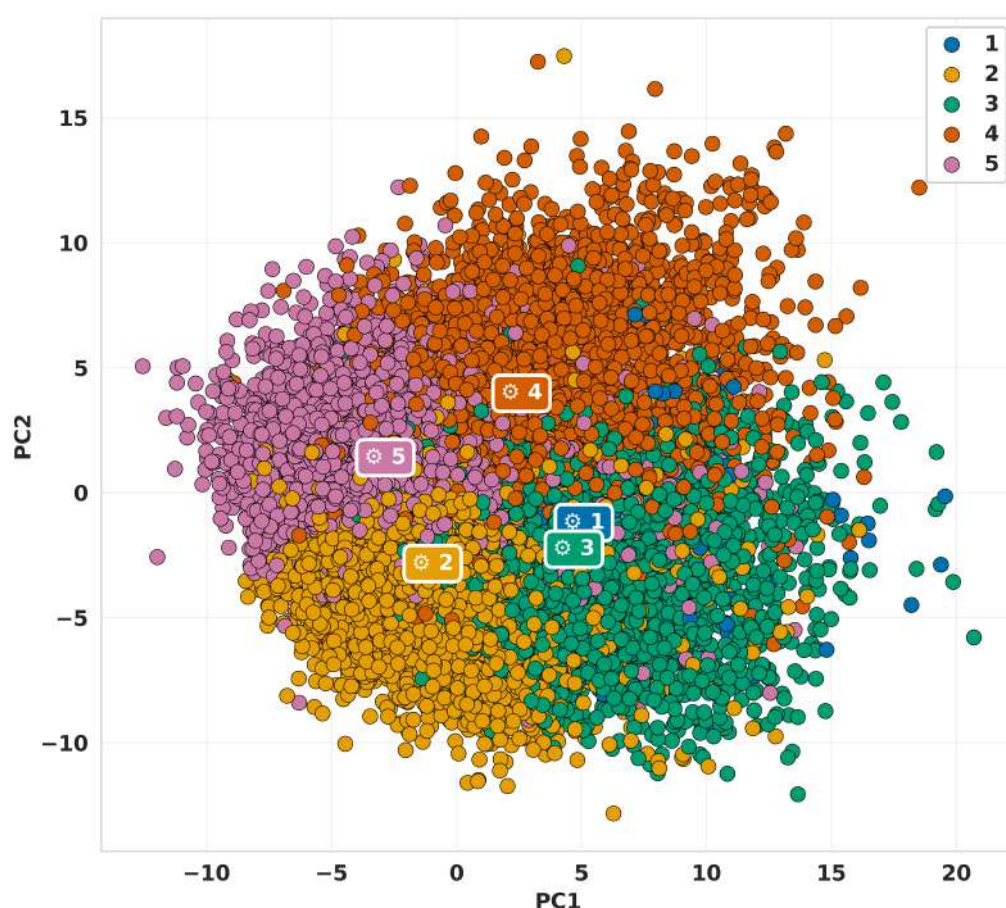


Figure 4 Principal component projection based on the full set of corporate indicators

The set was then reduced to 15 features. The grounds for selection were as follows: First, the cluster differences are ranked by the F statistic and by mutual information to retain the most discriminative dimensions. Second, correlation control with duplicate removal. Third, mandatory coverage of all semantic blocks relevant to the influence of gaming and behavioral practices on production and office processes, including downtime and defect reduction, turnover acceleration, and cash flow and liquidity strengthening. The core list includes X6 as the average number of employees, X52 as revenue, X73 as payroll and related benefits, X57 as profit or loss from sales, X134 as return on sales, and X137 as return on costs. With this composition, the first two components explain 64.45% of the variance because noise and redundancy are removed, and the main axes reflect real channels of operational and financial change (Figure 5).

A compact set of six key indicators was identified for regular monitoring, as shown in Table 3. This set is sufficient for stable profile separation and for building maps in two-dimensional projections, as shown in Figure 6. X6 captures workforce scale, X52 reflects revenue, and X73 measures payroll and related benefits. X57 represents profit or loss from sales, while X134 and X137 measure return on sales and return on costs, respectively.

In practical terms, the selected indicators characterize workforce scale, revenue, payroll expenses, sales profit, return on sales, and return on costs. The three-tier approach (full set to top 15 to key six) ensures detailed expert analysis on the full matrix, robust separation of profiles and year-to-year comparability on the set of 15 features, and readable maps with regular reports based on six indicators.

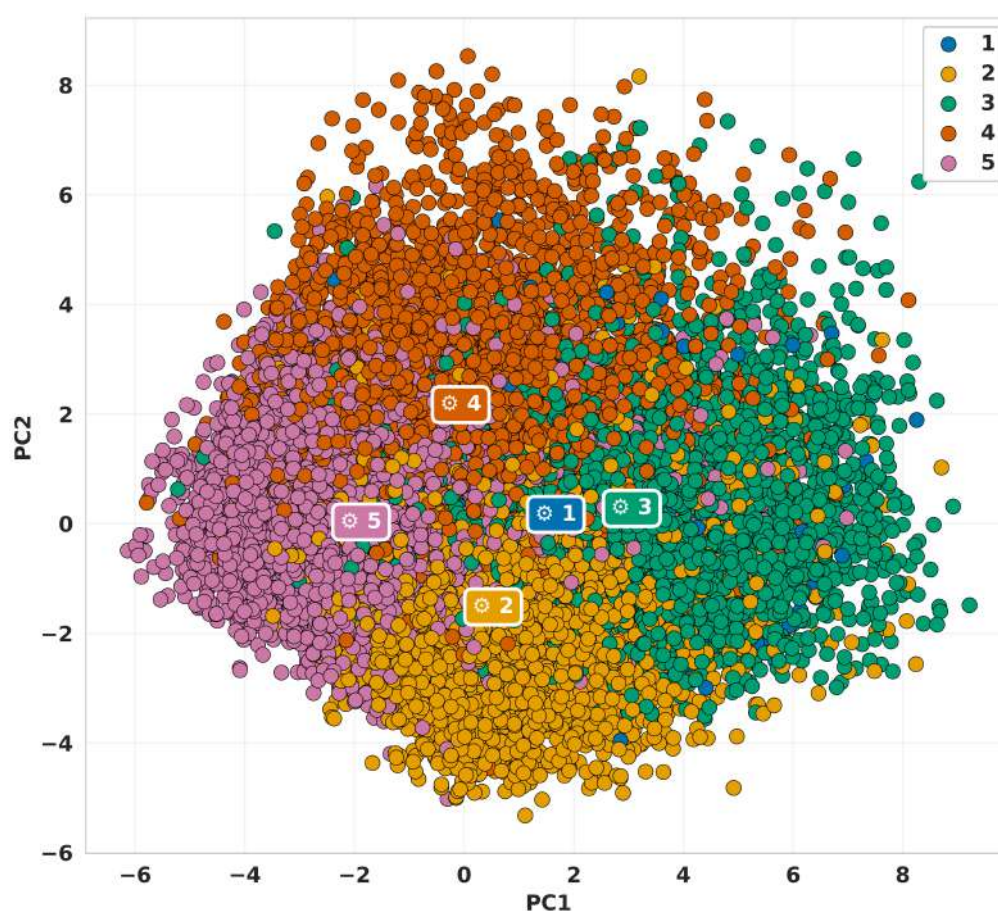


Figure 5 Principal component projection based on the reduced set of 15 indicators

Table 3 Key operational and financial indicators used for clustering and monitoring

Code	Indicator	Definition / formula	Unit
X57	Profit (loss) from sales	Revenue – Cost of goods sold – Selling expenses – Administrative expenses	RUB
X134	Return on Sales (ROS)	(Profit from sales / Revenue) × 100%	%
X137	Return on costs	(Profit from sales / (Cost of goods sold + Selling expenses + Administrative expenses)) × 100%	%
X52	Revenue	Revenue from sale of goods, works, and services	RUB
X73	Payroll (wages & related benefits)	Expenses for wages and related payments	RUB
X6	Average number of employees	Average headcount over the period	persons

4.7 Aggregated Index and Its Interpretation

The composite monitoring index aggregates six indicators reflecting productivity and turnover reserves and is computed using min–max scaling within each profile to preserve peer comparability. The index ranks firms relative to their nearest peers and highlights whether margin control, asset utilization, cycle acceleration, or financing discipline have potential for

improvement. Multi-year averaging reduces one-year volatility and supports program targeting and follow-up evaluation.

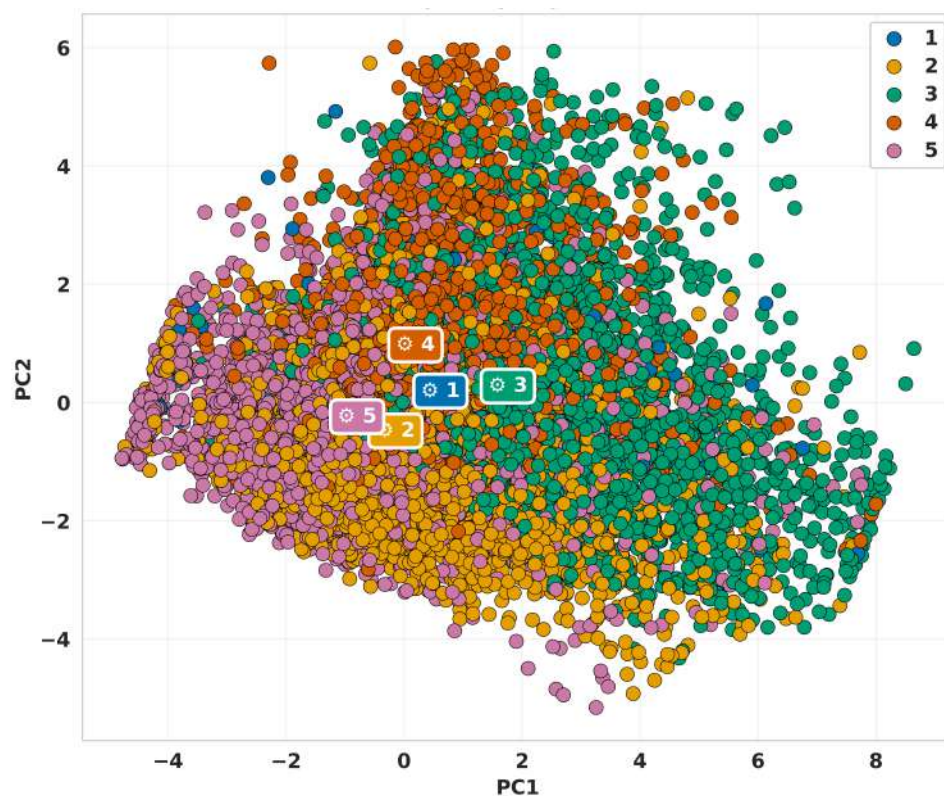


Figure 6 Principal component projection based on six key operational and financial indicators

Table 4 Cluster-level values of the composite peer-based monitoring index

Cluster	Mean composite index
1	0.555397
2	0.470112
3	0.461002
4	0.501893
5	0.445076

Interpretation. The closer a company's value is to 1 within its cluster, the stronger the economic effect of gamification channels among its peers. The closer to 0, the greater the scope for improvement through faster operations, lower defect rates, and stronger discipline of execution. The pipeline records the influence of gaming practices on output, margins, capital turnover, debt resilience, liquidity, and digital maturity. Correct comparability is preserved through standardization, equal weights, averaging over time, and intracluster normalization.

In our case, the indicator reflects the aggregated index's average position of cluster companies between the worst and the best profiles on a scale from 0 to 1. A value of approximately 0.5 corresponds to a near equilibrium distribution. Values above 0.5 indicate a shift toward the upper part of the range, whereas values below 0.5 indicate a shift toward the lower part. Cluster 1 has the highest mean index value, equal to 0.555397. A few strong profiles pull the mean upward, whereas weaker profiles are rare or filtered out by quality controls. The value of 0.501893 for cluster 4 is close to the middle, indicating an almost symmetric internal structure among capital-intensive enterprises. The mean index values for clusters 2 and 3 are 0.470112 and 0.461002, respectively. Most firms lie in the lower half of the range. Cluster 3 has a noticeable share of players at the growth stage without full returns on investment. In cluster 2, resilient

producers predominate with moderate improvements. Cluster 5 has the lowest mean index value, equal to 0.445076. The upper bound is stretched, but the bulk lies below the midpoint. Comparisons across clusters do not form a single ranking because each group is normalized. The values reflect the distributional shape and headroom for growth relative to the intracluster benchmarks.

4.8 Indicator of Informativeness and Heat Profiles

The informativeness of each of the six indicators is evaluated by two independent measures based on the table of standardized values and consolidated cluster labels. The first measure is the F statistic of one-way analysis of variance across clusters, which captures the strength of differences in means under comparable within-group dispersion. The second measure is the mutual information between the feature and the cluster label, which captures arbitrary, including nonlinear dependencies. The features are ordered by F values, and a bar chart is visualized in Figure 7, which enables a quick view of which metrics effectively separate the groups.

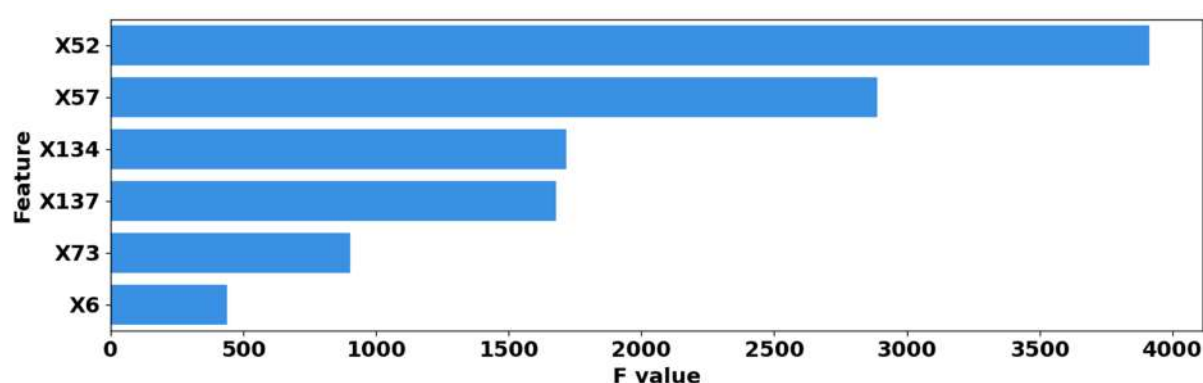


Figure 7 Relative informativeness of key indicators assessed using the F-statistic

Next, the mean standardized values are computed for each cluster, and a heat map is produced for the same features. Warm tones indicate values above the sample average, and cool tones indicate values below the average (Figure 8). The joint reading of Figures 7 and 8 yields a clear logic of the differences. Elevated X52 and X73 values indicate larger revenue and payroll scales, respectively. X57 reflects profit or loss from sales, X134 measures return on sales, X137 measures return on costs, and X6 characterizes workforce scale. Standardization secures the comparability of indicator contributions in averaging and interpreting the color scale.

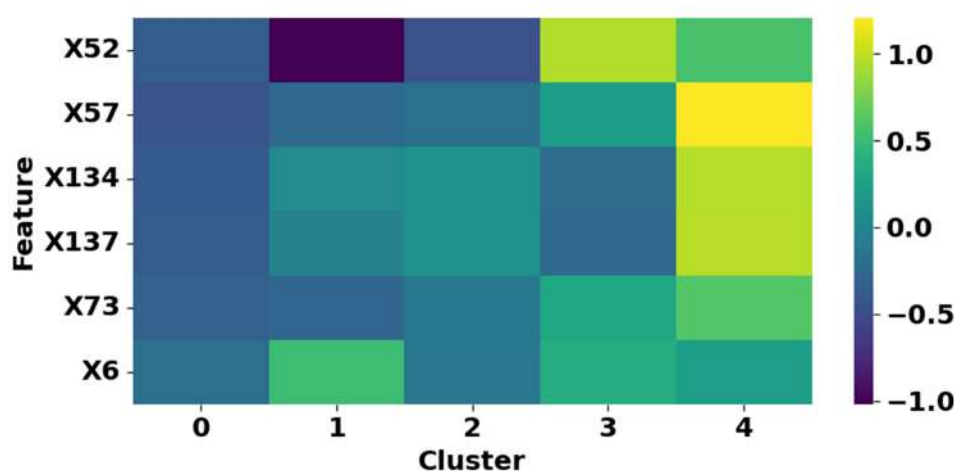


Figure 8 Heatmap of mean standardized values of key indicators across clusters. Uses zero-based labels 0–4, which correspond to clusters 1–5, respectively, in the text and tables

4.9 Cluster Validation, Robustness, and Boundary Interpretation

Validation combines statistical and geometric diagnostics. We performed pairwise comparisons across clusters on core indicators using Welch's t-test with false discovery rate adjustment; effect sizes are summarized with Hedges' g . Robustness to random k-means initialization is assessed via repeated restarts: agreement with the final partition reaches $\text{ARI} \approx 0.49$ and $\text{NMI} \approx 0.43$. Geometric diagnostics rely on pairwise inter-centroid distances in the standardized space (Table 5). The smallest centroid distance occurs between clusters 2 and 5 (0.771), consistent with a smooth boundary between stable mid-sized firms and a large mass of small enterprises. The largest distance appears between clusters 3 and 5 (2.610), aligning growth profiles with the most distinct operating–financial regime relative to the small-business segment's baseline.

Table 5 Pairwise distances between the centroids of the identified clusters

	μ_1	μ_2	μ_3	μ_4	μ_5
μ_1	0	1.132	1.155	1.039	1.495
μ_2	1.132	0	2.089	1.775	0.771
μ_3	1.155	2.089	0	1.826	2.610
μ_4	1.039	1.775	1.826	0	1.693
μ_5	1.495	0.771	2.610	1.693	0

Anomaly screening uses the Mahalanobis distance with the Ledoit–Wolf covariance and a chi-squared threshold at the 99.9th percentile, producing 1,015 candidate outliers. These cases serve as a dedicated review set because extreme financial patterns can distort centroid geometry and weaken interpretability. Overall, the validation results support a reproducible segmentation with soft boundaries, where adjacent profiles overlap but remain stable under repeated estimation. When stronger separation of overlapping clouds is required, mixture models may be used as a complementary check.

5. Conclusions

This study presents reproducible segmentation of Russian manufacturing enterprises using a unified panel of 6,479 firms (2015–2024) with 29,460 firm–year records. PCA is used to process standardized indicators and cluster them via k-means with $k = 5$, followed by firm-level label consolidation (modal label; ties resolved by mean distance to the centroid). The resulting five-profile map has cluster sizes of 45, 1,746, 993, 1,286, and 2,409. Repeated initializations yielded $\text{ARI} \approx 0.49$ and $\text{NMI} \approx 0.43$ relative to the final partition, indicating moderate agreement across solutions, while the low mean silhouette indicates substantial overlap between adjacent profiles. The inter-centroid distances ranged from 0.77 to 2.61, and the Mahalanobis outlier diagnostics (Ledoit–Wolf covariance; 99.9% χ^2 threshold) flagged 1,015 candidate outliers. For monitoring and peer benchmarking, a composite index was computed from six standardized indicators, and within-profile min–max normalization was applied on a 0–1 scale. The score ranks a firm relative to its closest peer group using indicators of workforce scale, revenue, payroll expenses, sales profit, return on sales, and return on costs. Program targeting is a simple application: a policymaker can select firms in the lower quantiles of the index within each profile, while a firm can track whether operational changes move its peer-relative position across reporting years. The limitations of this study are attributed to heterogeneous reporting, missingness-driven feature selection, and smooth boundaries between adjacent profiles. Further work should compare k-means with Gaussian Mixture Models on the top-15 feature set where the first two principal components explain 64.45% of variance and evaluate rolling-window reclustering to capture structural shifts. Adding operational telemetry alongside financial statements may improve boundary contrast and strengthen managerial interpretation.

Acknowledgements

The research was financed as part of the project “Development of a methodology for instrumental base formation for analysis and modelling of the spatial socio-economic development of systems based on internal reserves in the context of digitalization” (FSEG-2023-0008, 2023-2025).

Conflict of Interest

The authors declare no conflicts of interest.

References

- Ahn, S., & Chang, S. (2019). A similarity-based hierarchical clustering method for manufacturing process models. *Sustainability*, 11(9), 2560. <https://doi.org/10.3390/su11092560>
- Arruda, H. M., Bavaresco, R. S., Kunst, R., Bugs, E. F., Pesenti, G. C., & Barbosa, J. L. V. (2023). Data science methods and tools for Industry 4.0: A systematic literature review and taxonomy. *Sensors*, 23(11), 5010. <https://doi.org/10.3390/s23115010>
- Blachowicz, T., Wylezek, J., Sokol, Z., & Bondel, M. (2025). Real-time analysis and classification of welding parameters using HDBSCAN and k-means clustering. *Information*, 16(2), 79. <https://doi.org/10.3390/info16020079>
- Buggineni, V., Chen, C., & Camelio, J. (2024). Enhancing manufacturing operations with synthetic data: A systematic framework for data generation, accuracy, and utility. *Frontiers in Manufacturing Technology*, 4, 1320166. <https://doi.org/10.3389/fmtec.2024.1320166>
- Chen, M.-K., Wu, S.-W., Huang, Y.-P., & Chang, F.-J. (2022). The key success factors for the operation of SME cluster business ecosystem. *Sustainability*, 14(14), 8236. <https://doi.org/10.3390/su14148236>
- Chen, T., Sampath, V., May, M. C., Shan, S., Jorg, O. J., Aguilar Martín, J. J., & Calaon, M. (2023). Machine learning in manufacturing towards Industry 4.0. From “For Now” to “Four-Know”. *Applied Sciences*, 13(3), 1903. <https://doi.org/10.3390/app13031903>
- Chen, X., Wang, E., Miao, C., Ji, L., & Pan, S. (2020). Industrial clusters as drivers of sustainable regional economic development. *Sustainability*, 12(7), 2848. <https://doi.org/10.3390/su12072848>
- Chiang, T.-A., Che, Z.-H., Lee, C.-H., & Liang, W.-C. (2021). Applying clustering methods to develop an optimal storage-location planning-based consolidated picking methodology and operational sequencing. *Applied Sciences*, 11(21), 9895. <https://doi.org/10.3390/app11219895>
- Cottineau, C., & Arcaute, E. (2020). The nested structure of urban business clusters. *Applied Network Science*, 5, 2. <https://doi.org/10.1007/s41109-019-0246-9>
- Cui, Y., Niu, Y., Ren, Y., Zhang, S., & Zhao, L. (2024). A model to analyze industrial clusters to measure land use efficiency in China. *Land*, 13(7), 1070. <https://doi.org/10.3390/land13071070>
- Dhondt, N., Mendez Alva, F., & Van Eetvelde, G. (2024). Introducing industrial clusters in multi-node energy system modelling for neutrality and security. *Sustainability*, 16(6), 2585. <https://doi.org/10.3390/su16062585>
- Dmitriev, N., Zaytsev, A., Kichigin, O., & Yashchenko, E. (2022). Factor analysis of fixed capital investments: Regional aspect. *TEM Journal*, 11(3), 1108–1118. <https://doi.org/10.18421/TEM113-16>
- Dmitriev, N. D., & Satmurzaev, M. M. (2025). The impact of corporate social responsibility on enhancing the intellectual potential of organizations. *Economic Consultant*, 3(49), 65–79. <https://doi.org/10.46224/ecoc.2025.3.5>
- Eremina, I., & Rodionov, D. (2023). The special aspects of devising a methodology for predicting economic indicators in the context of situational response to digital transformation.

- International Journal of Technology*, 14(8), 1653–1662. <https://doi.org/10.14716/ijtech.v14i8.6839>
- Faizullin, R. V., Ototsky, P. L., & Goriacheva, E. N. (2025). Assessing the impact of artificial intelligence on Russian labor market development scenarios: Industry analysis. *Economic and Social Changes: Facts, Trends, Forecast*, 18(1), 170–189. <https://doi.org/10.15838/esc.2025.1.97.10>
- Gargalo, C. L., Malanca, A. A., Aouichaoui, A. R. N., Huusom, J. K., & Gernaey, K. V. (2024). Navigating Industry 4.0 and 5.0: The role of hybrid modelling in (bio)chemical engineering's digital transition. *Frontiers in Chemical Engineering*, 6, 1494244. <https://doi.org/10.3389/fceng.2024.1494244>
- Giordano, D., Mellia, M., & Cerquitelli, T. (2021a). A data-driven unsupervised multivariate segmentation and clustering method for anomaly detection in industrial contexts. *Electronics*, 10(10), 1166. <https://doi.org/10.3390/electronics10101166>
- Giordano, L., Messina, F., Santoro, C., Scata, M., & Longo, F. (2021b). K-MDTSC. k-means-based multi-dimensional time series clustering for anomaly detection. *Electronics*, 10(14), 1728. <https://doi.org/10.3390/electronics10141728>
- Glukhov, V., Shchepinin, V., Lyubek, Y., Babkin, I., & Karimov, D. (2023). Assessment of the impact of services and digitalization level on the infrastructure development in oil and gas regions. *International Journal of Technology*, 14(8), 1810–1820. <https://doi.org/10.14716/ijtech.v14i8.6855>
- Herrero, Á. C., Sangüesa, J. A., Garrido, P., Martínez, F. J., & Calafate, C. T. (2023). Mo-BiSea: A mobility-based binary search to cluster product flows in e-commerce. *Electronics*, 12(15), 3262. <https://doi.org/10.3390/electronics12153262>
- Hou, Z., Yan, R., & Wang, S. (2022). On the k-means clustering model for performance enhancement of port state control. *Journal of Marine Science and Engineering*, 10(11), 1608. <https://doi.org/10.3390/jmse10111608>
- Hu, Q., Qi, H., Huang, W., & Liu, M. (2023). A method to recommend cloud manufacturing service based on spectral clustering and improved Slope One algorithm. *Journal of Cloud Computing*, 12, 115. <https://doi.org/10.1186/s13677-023-00489-5>
- Ikotun, A. M., & Ezugwu, A. E. (2022). Boosting k-means clustering with symbiotic organisms search for automatic clustering problems. *PLOS ONE*, 17(8), e0272861. <https://doi.org/10.1371/journal.pone.0272861>
- Kannan, R., Abdul Halim, H. A., Ramakrishnan, K., Ismail, S., & Wijaya, D. R. (2022). Machine learning approach for predicting production delays: A quarry company case study. *Journal of Big Data*, 9, 94. <https://doi.org/10.1186/s40537-022-00644-w>
- Khan, A. A., & Abonyi, J. (2022). Simulation of sustainable manufacturing solutions. Tools for enabling circular economy. *Sustainability*, 14(15), 9796. <https://doi.org/10.3390/su14159796>
- Kılıç, D. K., & Nielsen, P. (2022). Comparative analyses of unsupervised PCA k-means change detection algorithm from the viewpoint of follow-up plan. *Sensors*, 22(23), 9172. <https://doi.org/10.3390/s22239172>
- Liu, H., & Ren, J. (2025). Intelligent manufacturing policy, ESG performance, and total factor productivity. *PLOS ONE*, 20(2), e0311369. <https://doi.org/10.1371/journal.pone.0311369>
- Lu, Q., Wang, S., Jiang, M., Li, Y., & Dong, K. (2021). Main control factors and sensitivity analysis of an improved k-means clustering algorithm. *PLOS ONE*, 16(5), e0248840. <https://doi.org/10.1371/journal.pone.0248840>
- Molinié, D., Madani, K., & Amarger, V. (2022). Clustering at the disposal of Industry 4.0. Automatic extraction of plant behaviors. *Sensors*, 22(8), 2939. <https://doi.org/10.3390/s22082939>
- Mourtzis, D. (2022). Advances in adaptive scheduling in Industry 4.0. *Frontiers in Manufacturing Technology*, 2, 937889. <https://doi.org/10.3389/fmtec.2022.937889>

- Pérez-Ortega, J., Almanza-Ortega, N. N., & Romero, D. (2018). Balancing effort and benefit of k-means clustering algorithms in big data realms. *PLOS ONE*, 13(9), e0201874. <https://doi.org/10.1371/journal.pone.0201874>
- Pittino, F., Puggl, M., Moldaschl, T., & Hirschl, C. (2020). Automatic anomaly detection on in-production manufacturing machines using statistical learning methods. *Sensors*, 20(8), 2344. <https://doi.org/10.3390/s20082344>
- Sandoyan, E., Rodionov, D., Voskanyan, M., & Galstyan, A. (2025). The role of digital technologies in capital market development: A pathway to economic growth in developing countries. *International Journal of Technology*, 16(2), 573–584. <https://doi.org/10.14716/ijtech.v16i2.7421>
- Scharmer, V. M., Vernim, S., Horsthofer-Rauch, J., Jordan, P., Maier, M., Paul, M., & Zaeh, M. F. (2024). Sustainable manufacturing: A review and framework derivation. *Sustainability*, 16(1), 119. <https://doi.org/10.3390/su16010119>
- Seo, D., Kim, S., Oh, S., & Kim, S.-H. (2022). K-means clustering-based safety system using wireless sensor networks in industrial sites. *Sensors*, 22(8), 2897. <https://doi.org/10.3390/s22082897>
- Shirasawa, N., & Seo, Y. (2025). The role of institutional and geographic proximity in enhancing Creating Shared Value initiatives in an industrial cluster setting. *Sustainability*, 17(6), 2410. <https://doi.org/10.3390/su17062410>
- Wang, T., Jing, Z., Zhang, S., & Qiu, C. (2023). Utilizing principal component analysis and hierarchical clustering to develop driving cycles. A case study in Zhenjiang. *Sustainability*, 15(6), 4845. <https://doi.org/10.3390/su15064845>
- Wang, X., Xie, Z., Yan, F., Wang, J., Fan, J., Zeng, Z., Lu, J., Zhang, H., & Zeng, N. (2025). Towards more accurate industrial anomaly detection: A component-level feature-enhancement approach. *Electronics*, 14(8), 1613. <https://doi.org/10.3390/electronics14081613>
- Yuan, C., & Yang, H. (2019). Research on K-value selection method of k-means clustering algorithm. *J*, 2(2), 226–235. <https://doi.org/10.3390/j2020016>
- Zaytsev, A., & Dmitriev, N. (2025). Automated collection and processing of spatiotemporal data for the analysis of sustainable development in industrial systems. *2025 International Russian Smart Industry Conference (SmartIndustryCon)*, 1043–1050. <https://doi.org/10.1109/SmartIndustryCon65166.2025.10986205>
- Zaytsev, A., Mihel, E., Dmitriev, N., Alferyev, D., & Laszlo, U. (2024). Optimization of interaction with counterparties: Selection game algorithm under uncertainty. *Mathematics*, 12, 2079. <https://doi.org/10.3390/math12132079>
- Zhang, G., Li, Y., & Deng, X. (2020). K-means clustering-based electrical equipment identification for smart building application. *Information*, 11(1), 27. <https://doi.org/10.3390/info11010027>
- Zhang, W., Bao, X., Hao, X., & Gen, M. (2025). Metaheuristics for multi-objective scheduling in smart manufacturing: A survey. *Frontiers in Industrial Engineering*, 3, 1540022. <https://doi.org/10.3389/fie.2025.1540022>
- Zhao, C., Ma, C., & Zhang, H. (2022). Modeling manufacturing resources based on manufacturability features. *Scientific Reports*, 12, 10775. <https://doi.org/10.1038/s41598-022-15072-2>
- Zhu, Y., Wang, R., Feng, M., Qin, L., Shia, B.-C., & Chen, M.-C. (2024). Supply chain analysis based on community detection of multi-layer weighted networks. *Mathematics*, 12(22), 3606. <https://doi.org/10.3390/math12223606>