



RESEARCH ARTICLE

Clustering of Regions of the Russian Federation Based on Intellectual Data Analysis

S.S. Mikhailova¹, N.V. Grineva^{1*}, A.A. Vilkul¹, U.A. Bachaev²^{1,2}Financial University under the Government of the Russian Federation, Moscow, Russia

ARTICLE INFO	ABSTRACT
Received: AUG 25, 2026	The high structural heterogeneity and spatial non-uniformity of the Russian economy call for an objective, mathematically grounded typology of regional economic systems (RES), whereas classical grouping methods rely on subjective expert weights; the incompleteness of the initial statistical arrays is an additional barrier. Aim - to develop and test a methodology for recovering missing data and objectively clustering the Russian regions using unsupervised machine-learning algorithms. The paper considers the stages of intelligent data pre-processing, dimensionality reduction of the feature space, and segmentation of the regions. Methods. Missing-value imputation by the k-nearest-neighbours method (KNNImputer), Z-standardization (StandardScaler), principal component analysis (PCA), and K-Means clustering, with quality verification by the silhouette coefficient, the Davies-Bouldin and the Calinski-Harabasz indices. Results. A three-cluster taxonomy of the Russian regions has been obtained (leading regions, regions with medium potential, and laggards territories), ranked along the first principal component; the clustering quality is confirmed by the metrics. The resulting typology can be used to design a differentiated regional policy.
Accepted: SEP 23, 2026	
Keywords	
Cluster analysis	
Regional economic systems	
Data recovery	
Missing-value imputation	
K-nearest neighbours	
Principal component analysis	
K-means	
Unsupervised machine learning	
Regional taxonomy	

Corresponding Author:ngrineva01@gmail.com

INTRODUCTION

The current stage of economic development in the Russian Federation is characterized by a high degree of structural heterogeneity, pronounced spatial unevenness, and growing socioeconomic asymmetry across its constituent regions. Major geopolitical challenges, the structural reconfiguration of markets, and the need to accelerate the achievement of technological sovereignty are driving a shift in governance paradigms: uniform regional policy is giving way to flexible, adaptive, and differentiated spatial development strategies. The effectiveness and precision of these strategies depend directly on the objectivity, relevance, and mathematical rigor of the taxonomy of regional economic systems (RES).

Conventional methods for grouping and ranking regions, traditionally used in Russian practice, often rely on expert-assigned weights, heuristic assumptions, or the analysis of isolated univariate indicators. These approaches are highly subjective and cannot fully capture the complex, nonlinear, and latent relationships within multidimensional statistical datasets, particularly under rapidly changing macroeconomic conditions. Implementing modern data analysis methods and unsupervised machine learning algorithms has therefore become an important scientific and practical task. These tools enable automated pattern recognition and the development of typologies for complex systems based solely on the intrinsic geometry and variance of the feature space.

The incompleteness of source statistical datasets is one of the central challenges in the quantitative analysis of regional system indicators and constitutes a distinct area of research (Little, Rubin, 2019; van Buuren, 2018). Studies on missing-data handling traditionally distinguish three groups

of approaches. The first involves removing observations or variables with missing values (listwise and pairwise deletion); when the population size is strictly limited, as in the case of the number of constituent regions of the Russian Federation, this approach leads to a loss of degrees of freedom and sample bias (Little, Rubin, 2019). The second is simple imputation using distributional statistics (the mean, median, or mode), which distorts the variance–covariance structure of the data (García-Laencina et al., 2010). The third and most appropriate approach is model-based imputation, including regression methods, the expectation–maximization (EM) algorithm, multiple imputation by chained equations (MICE), and machine learning algorithms (Emmanuel et al., 2021; van Buuren, 2018). Among the latter, the k -nearest neighbors method (KNN imputation), originally proposed for imputing missing values in multivariate data, is widely used (Troyanskaya et al., 2001); it preserves the local topology of the feature space and is therefore preferable when working with small samples of highly heterogeneous units (Emmanuel et al., 2021; García-Laencina et al., 2010). A conceptually similar approach to reconstructing incomplete economic time series using mathematical statistics has been applied to historical economic data (Grineva, Didenko, 2022).

The second methodological component concerns reducing the dimensionality of the feature space. Regional socioeconomic indicators exhibit high multicollinearity, which impairs the performance of distance-based algorithms (the “curse of dimensionality”). Principal component analysis (PCA) is a standard tool for addressing this problem, applying an orthogonal transformation to the original basis and extracting uncorrelated latent factors of regional development (Jolliffe, 2002; Jolliffe, Cadima, 2016).

The third component is clustering itself. The nonhierarchical k -means algorithm, originating in early studies (Lloyd, 1982; MacQueen, 1967), is the most common approach to constructing typologies of socioeconomic entities and, according to review studies, remains one of the most widely used methods of cluster analysis (Hartigan, Wong, 1979; Jain, 2010). Hierarchical and medoid-based algorithms are also employed (Kaufman, Rousseeuw, 1990). Clustering quality is assessed using a set of internal validation metrics: the silhouette coefficient (Rousseeuw, 1987), the Davies–Bouldin index (Davies, Bouldin, 1979), and the Caliński–Harabasz index (Caliński, Harabasz, 1974). Software implementations of all the algorithms listed above are available in the scikit-learn library (Pedregosa et al., 2011). In Russian practice, the grouping and ranking of regions have traditionally been addressed within regional economics and research on population quality of life (Aivazian, 2012; Granberg, 2003; Zubarevich, 2010). However, these approaches rely predominantly on expert-assigned weights, which motivates the shift toward objective unsupervised machine learning algorithms adopted in this study.

STATISTICAL DATA MINING

Characteristics of the Input Data

The empirical basis of the study consists of official data from the Federal State Statistics Service (Rosstat) (2025). The source data are organized as an observation–feature matrix, with the constituent regions of the Russian Federation as rows and a set of socioeconomic indicators as columns. Three temporal cross-sections—2015, 2020, and 2025—are examined to analyze changes in the taxonomy over time. The Republic of Crimea and the city of Sevastopol were excluded from the analysis because comparable data series for 2015 were unavailable. The analyzed population comprised 84 regions.

The feature space comprises thirteen indicators covering three substantive dimensions of regional economic systems: the economic dimension—gross regional product (GRP), fixed capital investment, average monthly wages, industrial output, retail trade turnover, and regional government debt; the sociodemographic dimension—the unemployment rate and related social indicators, natural population growth, life expectancy, and housing completions; and the infrastructure dimension—road network length, electricity generation and consumption, and railway network length.

Summary descriptive statistics for the indicators at the beginning and end of the study period (2015 and 2025) are presented in Table 1. These statistics allow an assessment of both the initial magnitude of each indicator and its pattern of change over the period under study.

Table 1: Descriptive statistics for the indicators (raw data), 2015 and 2025

Indicator	Unit	2015 (median; range)	2025 (median; range)
Gross regional product	million RUB	328,064; 39,192–12,779,526	760,915; 79,000–31,848,048
Fixed capital investment	million RUB	16,015; 8,834–54,814	31,347; 18,573–104,323
Average monthly wages	RUB	26,614; 19,239–79,531	66,298; 39,964–184,970
Industrial production index	% of previous year	102.9; 79.2–138.5	104.0; 84.1–124.9
Retail trade turnover	million RUB	163,262; 64,228–351,042	350,134; 0–8,287,865
Housing completions	million m ²	0.70; 0.0–9.6	0.74; 0.01–12.5
Regional government debt	thousand RUB	20,692,700; 0–145,264,406	25,243,989; 665,000–283,341,220
Unemployment rate	%	5.9; 1.8–30.6	1.9; 0.8–25.5
Natural population growth	persons	–1,900; –18,363 to 189,754	–5,075; –21,984 to 124,482
Life expectancy	years	70.4; 63.1–78.5	Data not available*
Road network length	km	16,578; 2,427–55,633	17,012; 2,849–53,341
Electrification level	arbitrary units	215; 0–6,100	141; 1–3,332
Railway network length	km	920; 39–3,524	922; 39–3,511

Life expectancy data for 2025 had not yet been published by Rosstat when the dataset was compiled; data for 2015 and 2020 are almost complete.

The descriptive statistics indicate substantial changes in the key indicators over the decade. Median gross regional product increased approximately 2.3-fold (from RUB 328 billion to RUB 761 billion in nominal terms), average monthly wages rose 2.4-fold (from RUB 26.6 thousand to RUB 66.3 thousand), and median fixed capital investment nearly doubled. At the same time, registered unemployment declined markedly, with its median falling from 5.9% to 1.9%. Nevertheless, enormous disparities persist across regions: the maximum gross regional product exceeds the minimum by more than a factor of 300. This directly indicates the structural heterogeneity of the regional economic landscape and provides a substantive justification for developing a regional typology.

Importantly, the indicators differ not only in scale but also in the direction of desirable change. For most indicators, higher values correspond to an improvement in a region's position (benefit indicators), whereas lower values are preferable for unemployment and government debt (cost indicators). Before observations are compared, cost indicators are transformed to align their direction with that of benefit indicators, while strongly right-skewed variables (monetary aggregates and network lengths) are first log-transformed and winsorized at the 2nd and 98th percentiles. Since the absolute values are not directly comparable, each feature is then centered and scaled. Without this procedure, Euclidean distances would be determined almost entirely by gross regional product and government debt, which have the largest absolute variances, while demographic and infrastructure characteristics would make a negligible contribution.

The raw datasets, which aggregate indicators from various statistical reporting forms, contain missing values arising from the heterogeneity of sources, changes in Rosstat's statistical methodology, and shifts in data publication schedules. The total number of missing values and their distribution across indicators and temporal cross-sections are presented in Table 2. For each year, the matrix analyzed comprises 84 regions and 13 indicators, corresponding to 1,092 cells.

Table 2: Percentage of missing values by indicator (%)

Indicator	Dimension	2015	2020	2025
Gross regional product	Economic	3.6	3.6	2.4
Fixed capital investment	Economic	6.0	3.5	4.8
Average monthly wages	Economic	2.4	3.6	2.4
Industrial production index	Economic	3.6	3.6	2.4
Retail trade turnover	Economic	13.1	13.1	13.1
Housing completions	Economic	3.6	3.6	2.4
Regional government debt	Economic	3.6	3.6	2.4
Unemployment rate	Social	3.6	3.6	3.6
Natural population growth	Demographic	1.2	1.2	1.2
Life expectancy	Demographic	1.2	1.2	—
Road network length	Infrastructure	13.1	13.1	13.1
Electrification level	Infrastructure	27.4	40.5	48.8
Railway network length	Infrastructure	10.7	10.7	10.7
Overall dataset		7.1	7.8	15.9

Missing values are distributed highly unevenly. Data for the core economic indicators are almost complete: the proportion of missing values for gross regional product, wages, the industrial production index, and housing completions does not exceed 4% in any temporal cross-section. Most missing values occur in the infrastructure dimension. The electrification indicator has the poorest data coverage: its proportion of missing values increases steadily from 27.4% in 2015 to 40.5% in 2020 and 48.8% in 2025, reflecting changes in reporting forms and differences in the level of detail of published data across years. Approximately 11–13% of values are consistently missing for retail trade turnover, road network length, and railway network length.

Across regions, the largest numbers of missing values are found in northern and Far Eastern constituent regions with limited statistical coverage: the Khanty-Mansi and Nenets Autonomous Okrugs, Magadan Oblast, Kamchatka Krai, and the republics of Altai and Tuva.

The objective of this stage is not merely to automate the development of a typology of regional economic systems, but to construct a mathematically grounded taxonomy of the constituent regions of the Russian Federation. This objective is achieved by combining feature engineering methods with unsupervised machine learning algorithms. The application of this analytical framework minimizes the subjectivity inherent in conventional expert assessments and provides a rigorous mathematical formalization of the underlying topological structures in the multidimensional space of socioeconomic indicators, creating opportunities to develop more adaptive and differentiated regional policy.

Data Preprocessing: Addressing Information Asymmetry and Normalizing the Feature Space

Constructing a representative dataset is the initial step that determines the validity of all subsequent modeling. Given the strictly limited size of the population (the number of constituent regions of the Russian Federation), the use of conventional heuristic approaches, such as removing observations (regions) with incomplete data, is methodologically inappropriate. Such a reduction results in the irreversible loss of valuable structural information, a critical decrease in the number of degrees of freedom, and substantial statistical bias in the analyzed sample. To address this issue, the present study implemented an intelligent missing-value imputation method based on the nonparametric k -nearest neighbors local approximation algorithm (KNNImputer).

The mathematical formulation of the method is based on calculating distances (primarily using the Euclidean metric) between feature vectors in an n -dimensional space. A missing feature value for a given region is estimated as the weighted mean of the corresponding values for the k neighboring regions whose profiles are most structurally similar to the target observation. This differentiated approach guarantees the preservation of endogenous relationships, the local topology of the data, and the internal variance–covariance structure of the sample, minimizing the risk of introducing artificial statistical noise.

Formally, for each missing entry, the k regions with the smallest Euclidean distances, calculated using the available (nonmissing) features, are selected. The missing value is then estimated as the weighted mean of the corresponding observed values, with weights inversely proportional to the distance to each neighbor. In this study, the number of neighbors was set to five ($k = 5$), and inverse-distance weights were used to increase the contribution of regions with the most similar profiles. This approach imputes values based on the local neighborhood of each observation and thus preserves the covariance structure of the data. Given the small population size (84 regions) and the moderate overall proportion of missing values (ranging from 7% in 2015 to 16% in 2025), it does not introduce noticeable shifts in the distributions. Plots of the indicator distributions for 2020 before and after KNN imputation are presented in Figure 1.

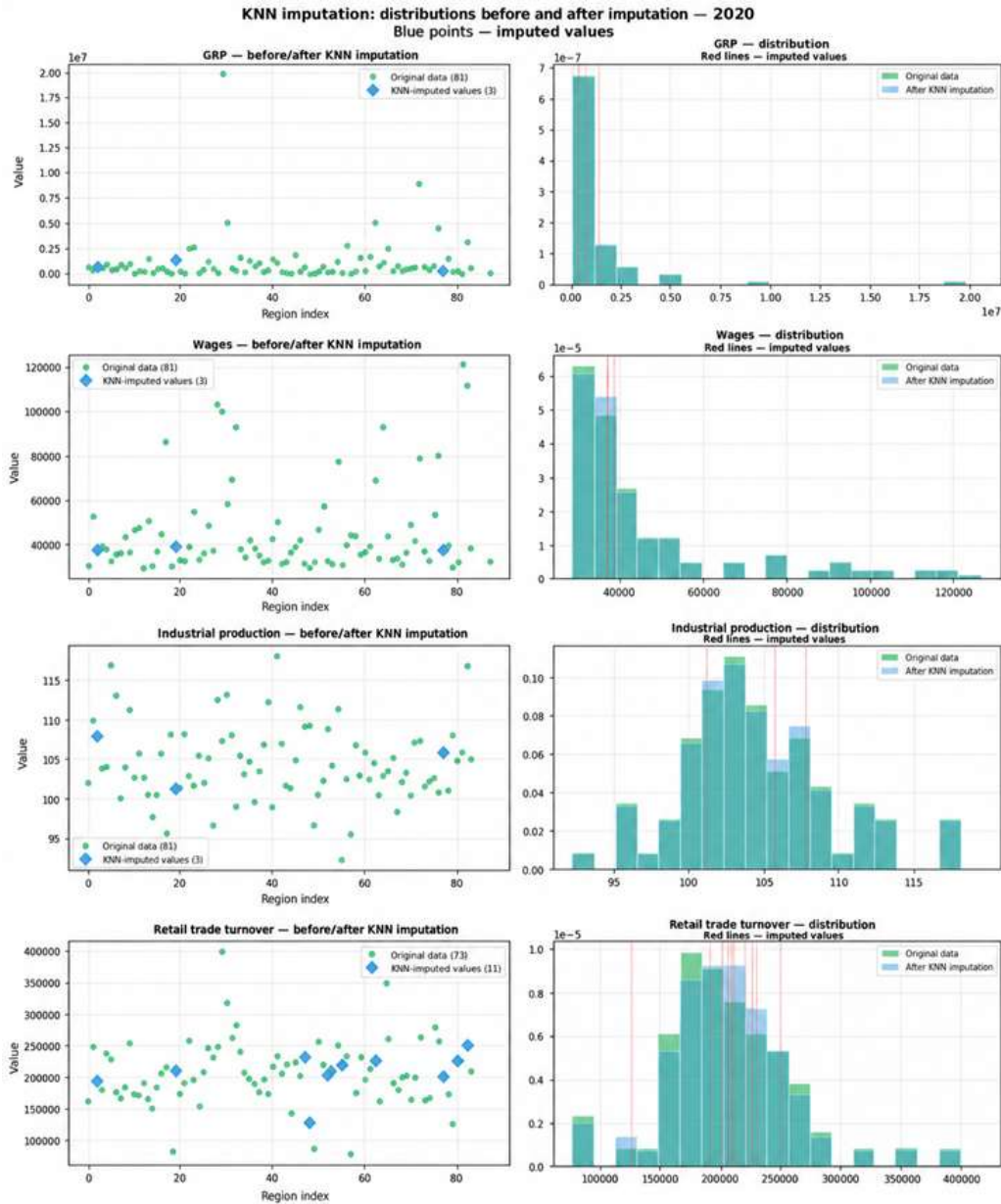


Figure 1: Distributions of the indicators (2020) before and after KNN imputation

The next stage of the algorithmic data transformation addresses the problem of incomparable measurement scales. The predictors under analysis differ substantially in their physical and economic dimensions, ranging from absolute financial values measured in hundreds of billions of rubles to relative demographic rates, percentage shares, and structural indices. These differences create a highly anisotropic (heterogeneous) feature space.

To render the vector space metrically homogeneous and eliminate scale effects, standardization, or Z-score scaling (StandardScaler), is applied. This procedure involves linear centering and scaling of each feature according to the following formula:

$$z = \frac{x - \mu}{\sigma}$$

where: μ - the mathematical expectation (empirical mean) of the feature; σ - its standard deviation.

As a result of this mathematical transformation, all variables under investigation are mapped onto a common dimensionless scale with a zero mean and unit variance. Such data homogenization is an essential mathematical requirement that ensures equal treatment of all socioeconomic factors. It guarantees that the optimization objective in clustering algorithms (such as K-means) and

dimensionality reduction methods (PCA) is based exclusively on the structural variance of the features rather than on artifacts arising from their original units of measurement.

The combined effect of logarithmic transformation and winsorization on the shapes of the indicator distributions is illustrated in Figure 2. For monetary aggregates and network lengths, the skewness coefficient, which initially exceeded one, approaches zero after transformation, making the features suitable for Euclidean distance calculations and improving the stability of the clustering solution.

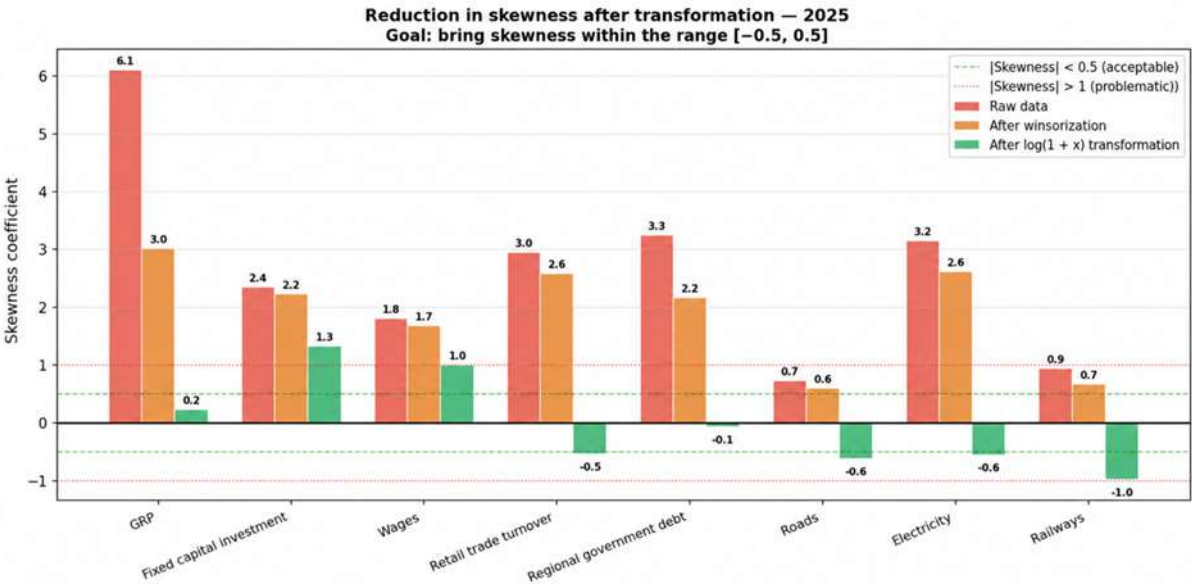


Figure 2: Reduction in the skewness of indicator distributions following winsorization and logarithmic transformation, 2025

Dimensionality Reduction of the Feature Space: Orthogonal Transformation and Principal Component Extraction (PCA)

The analysis of socioeconomic systems at the regional level inevitably encounters multicollinearity—a high degree of linear dependence among explanatory variables. In regional economics, many indicators, such as gross regional product, fixed capital investment, and household income levels, are strongly correlated because they reflect related aspects of a common macroeconomic process. Multicollinearity in the dataset substantially distorts the geometry of the multidimensional space: distance-based algorithms repeatedly account for the same underlying factors, leading to model overfitting and reduced stability of cluster structures (the “curse of dimensionality”).

To overcome this limitation and identify the fundamental, mutually uncorrelated drivers of regional development, principal component analysis (PCA) is applied. Replacing the original set of socioeconomic indicators with a limited number of principal components serves two purposes.

Information Compression and Noise Reduction

Discarding components with the smallest eigenvalues removes stochastic noise and minor fluctuations from the data while preserving the dominant macroeconomic signals (latent variables).

Metric Optimization

Constructing a strictly orthogonal basis ensures the proper functioning of distance metrics, particularly Euclidean distance, which is an essential prerequisite for the subsequent application of centroid-based clustering algorithms.

CLUSTER ANALYSIS

Regional systems are partitioned in the transformed and reduced principal component space using the K-means partitioning clustering algorithm. This algorithm was selected for its high computational efficiency, robustness when processing dense multidimensional datasets, and ability

to produce convex, compact clusters, making it well suited to developing a typology of the constituent regions of the Russian Federation.

From a mathematical perspective, K-means addresses an NP-hard optimization problem involving the minimization of within-cluster variance, known as inertia or the within-cluster sum of squares (WCSS). The objective function is expressed as follows:

$$f = \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 \rightarrow \min,$$

where: k - the specified number of clusters; S_i - the set of observations (constituent regions of the Russian Federation) assigned to the i -th cluster; x - the row vector of characteristics of a particular region in the principal component space; μ_i - the centroid of the i -th cluster (the vector of mean values across all observations within the cluster); $\|x - \mu_i\|^2$ - the squared Euclidean distance between an observation and its cluster centroid.

The objective function f is minimized through an iterative process (Lloyd's algorithm) comprising two alternating steps.

Expectation Step

Each constituent region of the Russian Federation is assigned exclusively to the cluster whose centroid is closest to it.

Maximization step

The centroid coordinates are recalculated as the geometric centers of mass (mean values) of the clusters formed in the preceding step.

The process is repeated until convergence, when the centroid coordinates stabilize and regions no longer move between clusters.

In this study, the target number of clusters was set to $= 3$, based on a preliminary analysis of the data topology using the elbow method and silhouette analysis. This model configuration allows three macroeconomic regional profiles to be identified, such as leading regions, regions with moderate potential, and lagging regions.

To ensure scientific rigor and full reproducibility of the empirical results, and to eliminate randomness in the initial selection of centroids, the pseudorandom number generator parameter is strictly fixed at the hardware level. Each resulting cluster is described by its centroid vector, which, when expressed in the coordinates of the original features, serves as an "ideal type"—a reference macroeconomic profile representing the average characteristics of the constituent regions within that cluster.

Clustering Validation and Quality Assessment

To objectively assess the quality and mathematical validity of the resulting three-cluster structure of regional economic systems, a set of complementary metrics is used:

Silhouette Coefficient

This metric assesses cluster compactness and separation. The average silhouette coefficient provides an indication of whether clusters are free of substantial overlap and whether regions have been appropriately assigned to the identified groups.

Davies–Bouldin Index

This index measures similarity between clusters by calculating the ratio of the sum of within-cluster distances to between-cluster distances. Lower index values indicate better clustering quality, showing that the clusters are compact and well separated.

Calinski–Harabasz Index (Variance Ratio Criterion)

This index represents the ratio of between-cluster variance to within-cluster variance, adjusted for the number of degrees of freedom. High values confirm the distinctness and geometric clarity of the boundaries in the resulting cluster structure.

The results obtained by applying these criteria to the three-cluster solution for all three temporal cross-sections are summarized in Table 3. The preliminary selection of the number of clusters was based on the combined assessment of the metrics for k values ranging from two to eight (Figure 3).

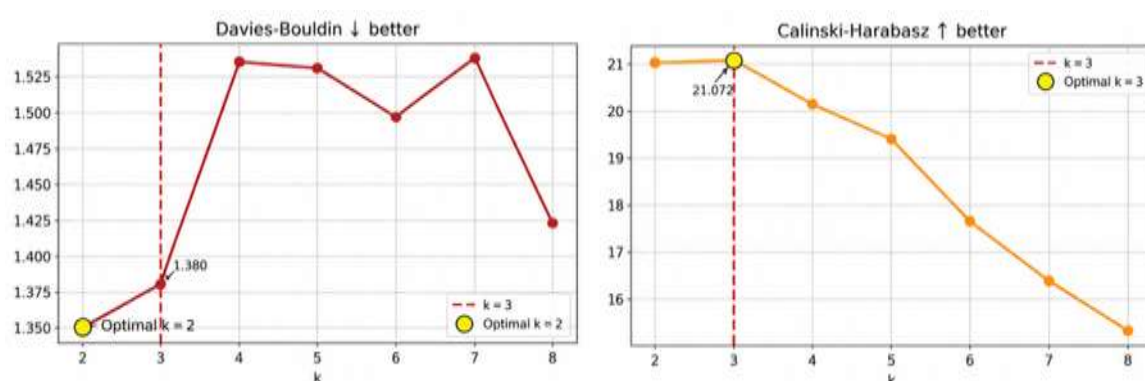


Figure 3: Justification for the number of clusters based on a combination of metrics (the elbow method, silhouette coefficient, Davies–Bouldin index, and Caliński–Harabasz index)

Table 3: Clustering quality metrics and cluster sizes ($k = 3$)

Year	Silhouette coefficient	Davies–Bouldin index	Caliński–Harabasz index	Leading regions	Mid-level regions	Lagging regions
2015	0.310	1.516	20.1	13	60	11
2020	0.310	1.380	21.1	9	62	13
2025	0.289	1.468	21.2	14	57	13

Across all temporal cross-sections, the silhouette coefficient remains stable at approximately 0.29–0.31, the Davies–Bouldin index does not exceed 1.52, and the Caliński–Harabasz index ranges from 20 to 21. Given the substantial heterogeneity of the original data, silhouette values above 0.29 indicate the absence of substantial overlap between groups, while the low Davies–Bouldin index confirms their compactness and separation. Taken together, these metrics and the substantive interpretability of the groups justify the selection of three clusters as the working solution.

Interpretation of the Segments and Analysis of Development Trajectories

Since the underlying algorithm assigns arbitrary numerical identifiers to the clusters, principal component analysis (PCA) was used to facilitate their substantive interpretation. Projecting the multidimensional space onto the first principal component established a single axis of overall development. Ordering the clusters along this axis provided a mathematically grounded basis for unambiguously ranking the groups as leading regions, regions with an intermediate level of development, and lagging regions.

Projecting the multidimensional feature space onto the first principal component (PC1) made it possible to construct a single axis of overall development. Ordering the resulting clusters along this axis enabled a mathematically rigorous ranking of the groups. The clusters were consequently interpreted as follows

Cluster 0 (core regions): Federal cities serving as Russia’s political and cultural capitals exhibit the highest values for such indicators as population size, GRP, living standards, and household income. At the same time, they have low unemployment and infant mortality rates.

Cluster 1 (leading regions): As shown by the radar chart (Figure 4), this group exhibits high values across all key economic determinants. These are primarily reflected in high wage levels, strong investment activity, and substantial industrial output. Such indicators point to high economic resilience, well-developed infrastructure, and considerable potential within these regional systems.

Cluster 2 (regions with moderate potential): This group occupies an intermediate position within the resulting taxonomy. It is characterized by moderate values, close to the sample average, for most of the macroeconomic and social indicators analyzed. The group forms a stable foundation of the regional economy but requires additional incentives to move into Cluster 1.

Cluster 3 (lagging regions): This segment comprises regions with the lowest values for the core economic indicators (GRP, fixed capital investment, and wages). The cluster is characterized by adverse social conditions, particularly higher unemployment. This signals deep-seated structural constraints, low spatial connectivity, and the need for targeted government support measures.

Comparing the three temporal cross-sections makes it possible to trace not only the static structure but also the trajectories of regions moving between typological groups. The cluster membership of each constituent region in 2024 is presented in Table 4, while the aggregate flows between groups are shown in the transition matrices (Figure 4).

Table 4: Regional cluster membership in 2024

Cluster	Regions
3	Kabardino-Balkarian Republic; Karachay-Cherkess Republic; Republic of Adygea; Altai Republic; Republic of Dagestan; Republic of Ingushetia; Republic of Kalmykia; Republic of North Ossetia–Alania; Republic of Tuva; Chechen Republic
2	Amur Oblast; Arkhangelsk Oblast; Kamchatka Krai; Krasnoyarsk Krai; Magadan Oblast; Murmansk Oblast; Nenets Autonomous Okrug; Primorsky Krai; Republic of Buryatia; Republic of Karelia; Komi Republic; Sakha Republic (Yakutia); Republic of Khakassia; Sakhalin Oblast; Tyumen Oblast; Khabarovsk Krai; Khanty-Mansi Autonomous Okrug–Yugra; Chukotka Autonomous Okrug; Yamalo-Nenets Autonomous Okrug; Yaroslavl Oblast; Altai Krai
1	Altai Krai; Astrakhan Oblast; Belgorod Oblast; Bryansk Oblast; Vladimir Oblast; Volgograd Oblast; Vologda Oblast; Voronezh Oblast; Jewish Autonomous Oblast; Zabaykalsky Krai; Ivanovo Oblast; Irkutsk Oblast; Kaliningrad Oblast; Kaluga Oblast; Kemerovo Oblast–Kuzbass; Kirov Oblast; Kostroma Oblast; Krasnodar Krai; Kurgan Oblast; Kursk Oblast; Leningrad Oblast; Lipetsk Oblast; Moscow Oblast; Nizhny Novgorod Oblast; Novgorod Oblast; Novosibirsk Oblast; Omsk Oblast; Orenburg Oblast; Oryol Oblast; Penza Oblast; Perm Krai; Pskov Oblast; Republic of Bashkortostan; Mari El Republic; Republic of Mordovia; Republic of Tatarstan; Rostov Oblast; Ryazan Oblast; Samara Oblast; Saratov Oblast; Sverdlovsk Oblast; Smolensk Oblast; Stavropol Krai; Tambov Oblast; Tver Oblast; Toms Oblast; Tula Oblast; Udmurt Republic; Ulyanovsk Oblast; Chelyabinsk Oblast; Chuvash Republic; Yaroslavl Oblast
0	Moscow; St. Petersburg

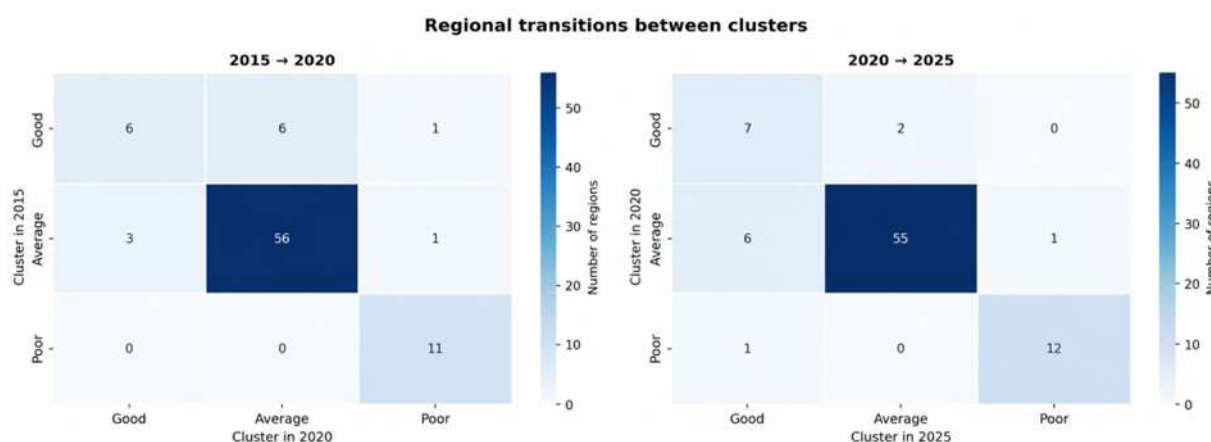


Figure 4: Transition matrices showing regional movements between clusters

The analysis of changes in the cluster structure of Russian regions over 2015–2024 reveals substantial redistribution during the initial period, followed by high cluster stability. In 2015, 53 regions belonged to the moderate-potential cluster, 16 to the leading cluster, and 13 to the economically depressed cluster. By 2020, the structure had changed considerably: the number of leading regions had increased to 49, while the moderate-potential cluster had contracted to 21 regions and the lagging cluster to 12. In 2024, the distribution remained virtually unchanged: 49 regions belonged to the leading cluster, 22 to the moderate-potential cluster, and 11 to the economically depressed cluster.

The most intensive movement between clusters occurred during 2015–2020. The predominant transition was from the moderate-potential cluster to the leading group, with 36 constituent regions of the Russian Federation moving in this direction. At the same time, three regions—

Krasnoyarsk Krai, Tyumen Oblast, and the Khanty-Mansi Autonomous Okrug–Yugra—moved from the leading cluster to the moderate-potential cluster.

The economically depressed cluster was generally more stable. Of the 13 regions assigned to it in 2015, 12 remained in this group in 2020. The only exception was the Nenets Autonomous Okrug, which moved into the moderate-potential cluster. Thus, unlike the middle segment, the economically depressed regions did not exhibit widespread movement into higher-ranked groups, indicating strong persistence at the lower end of the regional hierarchy.

During 2020–2024, the cluster structure became almost entirely stable. Only one transition between clusters occurred during this period: the Jewish Autonomous Oblast left the economically depressed cluster and was assigned to the moderate-potential group. All 49 regions belonging to the leading cluster in 2020 retained their membership in 2024. Similarly, all 21 regions in the moderate-potential cluster retained their status, while the economically depressed group decreased from 12 to 11 regions following the transition of the Jewish Autonomous Oblast.

CONCLUSION

This study successfully designed, implemented, and validated an integrated mathematical framework for the multivariate clustering of socioeconomic systems in the constituent regions of the Russian Federation.

The proposed methodology provides a complete processing cycle for heterogeneous regional statistics: missing-value imputation using the *k*-nearest neighbors method, transformation of cost indicators to align their direction with that of benefit indicators, logarithmic transformation and winsorization of skewed features, standardization, dimensionality reduction through principal component analysis, and nonhierarchical clustering using K-means. The methodology was tested on a dataset comprising 84 constituent regions of the Russian Federation and thirteen socioeconomic indicators for 2015, 2020, and 2024. The overall proportion of imputed values ranged from 7% to 16%, depending on the temporal cross-section.

The quality of the resulting three-cluster solution was confirmed by a set of internal validation metrics: the silhouette coefficient was approximately 0.29–0.31 across all temporal cross-sections, the Davies–Bouldin index did not exceed 1.52, and the Caliński–Harabasz index remained at approximately 20–21. The resulting taxonomy consistently separates regions into a leading group (9–14 regions), a large middle group (57–62 regions), and a lagging periphery (11–13 regions). Ranking the groups along the first principal component gives the partition a substantive economic interpretation.

The analysis of changes over the decade revealed considerable inertia in the regional structure: the vast majority of regions retained their group membership, six regions formed a stable core within the leading group, and the lagging periphery of eleven regions, mostly ethnic republics, remained virtually unchanged. These findings indicate entrenched spatial inequality and limited upward mobility among the most disadvantaged territories.

Promising directions for further methodological development include expanding the set of indicators, applying alternative imputation and clustering algorithms, including hierarchical and density-based methods, and assessing the stability of the cluster solution using bootstrapping.

ACKNOWLEDGMENTS

The article was prepared based on the results of research carried out at the expense of budgetary funds on the state assignment of the Financial University.

REFERENCES

- Aivazian SA. Analysis of the Population's Quality of Life and Lifestyle: An Econometric Approach. Moscow: Nauka; 2012. 432 p.
- Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 1974;3:1-27. <https://doi.org/10.1080/03610927408827101>
- Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell* 1979;PAMI-1:224-7.
- Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in

- machine learning. *J Big Data* 2021,8:140. <https://doi.org/10.1186/s40537-021-00516-9>
- Federal State Statistics Service. Regions of Russia. Socioeconomic Indicators. 2025: Statistical Compendium. Moscow: Rosstat; 2025. 1035 p.
- García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR. Pattern classification with missing data: A review. *Neural Comput Appl* 2010,19:263-82. <https://doi.org/10.1007/s00521-009-0295-6>
- Granberg AG. Fundamentals of Regional Economics. Moscow: State University–Higher School of Economics Publishing House; 2003. 495 p.
- Grineva NV, Didenko DV. Methods of mathematical statistics in reconstruction of historical data on economic growth factors of the former USSR Republics. *J Econ Hist Hist Econ* 2022,23:44-81. [https://doi.org/10.17150/2308-2488.2022.23\(1\).44-81](https://doi.org/10.17150/2308-2488.2022.23(1).44-81)
- Hartigan JA, Wong MA. Algorithm AS 136: A K-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* 1979,28:100-8. <https://sci-hub.al/10.2307/2346830>
- Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognit Lett* 2010,31:651-66. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Jolliffe IT. Principal Component Analysis. 2nd ed. New York: Springer; 2002. 487 p.
- Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. *Phil Trans R Soc A* 2016,374:20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons; 1990. 342 p.
- Little RJA, Rubin DB. Statistical Analysis with Missing Data. 3rd ed. Hoboken: John Wiley & Sons; 2019. 462 p.
- Lloyd SP. Least squares quantization in PCM. *IEEE Trans Inf Theory* 1982,28:129-37.
- MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1967. vol. 1, p. 281-97.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D. Scikit-learn: Machine learning in Python. *J Mach Learn Res* 2011,12:2825-30.
- Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987,20:53-65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001,17:520-5. <https://doi.org/10.1093/bioinformatics/17.6.520>
- van Buuren S. Flexible Imputation of Missing Data. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2018. 444 p. <https://doi.org/10.1201/9780429492259>
- Zubarevich NV. Regions of Russia: Inequality, Crisis, and Modernization. Moscow: Independent Institute for Social Policy; 2010. 160 p.