








# Identifying Patterns of Regional Polarization in Russia: A Machine Learning Approach

Daniel M. Balungu  , Anna V. Rozanova , Kristina A. Andreeva ,  
Anastasia V. Solod , Yutong Chen 

*Ural Federal University  
named after the First President of Russia B. N. Yeltsin,  
Yekaterinburg, Russia  
 danielbal03db@gmail.com*

**Abstract.** The study of regional polarization is critically relevant for Russia, as pronounced socio-economic disparities threaten national economic stability, social cohesion, and the strategic goal of technological sovereignty. These imbalances hinder efficient resource allocation and create vulnerabilities amid global realignment. This study analyzes patterns and drivers of regional polarization in Russia by leveraging machine learning to model the complex interplay of economic and non-economic factors. The research is guided by three hypotheses: polarization is driven by multi-dimensional disparities (H1); processes of convergence and divergence coexist (H2); and machine learning can effectively identify latent polarization structures overlooked by traditional methods (H3). K-means clustering identified regional typologies, with optimal cluster count validated by the Elbow Method and Silhouette Score. Predictive power and key drivers were analyzed using ensemble methods, Random Forest and XGBoost, with performance evaluated by Mean Squared Error (MSE). Results confirm a deeply polarized landscape with four distinct socio-economic clusters: resource-rich regions plagued by inequality and outmigration; leading hubs facing over-centralization risks; industrial-agrarian regions with persistent poverty; and lagging republics trapped in subsidy dependence. Analysis validated H2, showing simultaneous catch-up growth in some areas and entrenched divergence in others. Random Forest showed superior predictive accuracy (MSE = 0.132), confirming H3 and identifying key drivers (H1) beyond economic metrics. Theoretical significance lies in its novel, data-driven typology of Russian regions, highlighting the superiority of ML ensemble methods for modeling complex spatial inequality. Practically, findings provide policymakers with an evidence-based roadmap for differentiated regional strategies and a predictive tool for proactive intervention vital to balanced national development.

**Key words:** regional polarization; interregional differentiation; machine learning; cluster analysis; time series forecasting.

JEL R12, C61

## 1. Introduction

A defining feature of Russia's post-Soviet economic transformation has been the pronounced divergence in socio-economic development across its regions. The transition to a market economy, while driving aggregate growth, has also exacerbated long-standing interregional inequalities. Kuzin [1] attributes this to market mechanisms that have failed to mitigate disparities in economic efficiency and social well-being, leading to a self-reinforcing cycle where economic imbalances slow overall progress, reduce production, and increase unemployment.

This phenomenon is a manifestation of spatial economic polarization, the uneven development resulting from the asymmetric redistribution of resources, which creates significant differences in regional economic and social progress. Understanding these disparities is critical for fostering equitable development, enabling efficient policymaking, and maintaining social stability. However, the consequences of polarization extend beyond the economic sphere. As noted by Conover et al. [2], severe regional divergence can erode social cohesion and political stability. This is evident in governance challenges, where, as Silva [3] argues, deep-seated regional divides can fuel conflicting political agendas, leading to legislative gridlock and eroding public trust in institutions.

While complete societal homogeneity is neither attainable nor desirable, targeted interventions are necessary to mitigate polarization's most divisive effects. A growing body of research, including that of Levy [4], suggests that such strategies must address the underlying structural drivers of division rather than merely its symptoms. To date, the literature on Russian regional disparities has primarily focused on economic determinants, such as resource wealth and federal subsidies, or broad political dynamics. A significant gap remains in the systematic analysis of non-economic drivers including migration patterns, educational access, and digital infrastructure; and in the application of advanced analytical methods to uncover latent patterns of regional divergence.

This study seeks to address these gaps by employing data-driven techniques, including machine learning, to analyze the multidimensional dynamics of polarization in Russia. By moving beyond traditional economic explanations, this research offers a more comprehensive understanding of the underlying forces shaping the country's regional inequalities and provides a foundation for more sustainable policy solutions.

To guide this investigation, the study explores *three key questions (RQ)*:

*RQ1*: What are the principal economic and non-economic factors driving regional polarization in Russia?

*RQ2*: How do convergence and divergence processes manifest simultaneously across Russian regions?

*RQ3*: And can machine learning models effectively identify clusters of polarized regions?

These questions are examined through three corresponding *hypotheses*:

*H1*: Regional polarization is driven not only by traditional economic factors like GDP and investment but also by social and infrastructural disparities, including access to education and healthcare.

*H2*: Convergence (evident in catch-up growth) and divergence (seen in core-periphery divides) coexist, influenced by federal policies and regional adaptability.

*H3*: Machine learning techniques such as clustering and dimensionality reduction can reveal polarization patterns that conventional econometric methods might overlook.

*The purpose of this study* is to identify and analyze patterns of regional polarization in Russia using machine learning, with the broader goal of informing policies that mitigate spatial inequalities.

*Article structure:* The paper begins with a review of existing literature on regional polarization and methodological gaps, followed by a detailed explanation of the analytical approach, which combines clustering and predictive modeling. The findings are then presented alongside their policy implications, and the study concludes by outlining directions for future research.

## 2. Literature Review

Uneven development of regions, known as regional polarization, leads to the emergence of both prosperous and lagging areas. This imbalance is formed under the influence of many forces, among which are the concentration of production capacities, the direction of investment flows, the dynamics of labor supply and government decision-making.

Let us elaborate on the definition of key terms and concepts related to regional polarization such as regional divergence, polarization, and spatial inequality.

Regional divergence is the unevenness of economic, social, and environmental conditions in different areas [5]. These divergences are often reflected in employment levels, earnings, access to education, and health care [6].

Polarization is the process by which groups within a society become increasingly distinct and separate because of divergent economic, social or political interests. In the context of regions, this often means that wealthy areas are getting richer and poorer areas are falling behind.

Spatial inequality is a broader term that encompasses the unequal distribution of resources and opportunities, resulting in different outcomes in the level and quality of life in different areas [7].

The analysis of regional inequalities reveals a few reasons, among which globalization, rapid technological progress and specific features of local policy decisions stand out. Many studies have examined the drivers and outcomes of regional inequality and polarization. For example, Hansen [8] emphasizes the role of capital accumulation in deepening income inequality, which can lead to regional polarization. Similarly, Marco et al. [9] highlighted that capital investment in education and infrastructure is necessary to mitigate regional inequality in the EU.

In addition, Glaeser [10] point to urbanization as a significant driver of regional polarization. They argue that cities attract talent and investment, creating a feedback loop that further concentrates resources. However, these studies also demonstrate methodological limitations, often relying on classical regression models that may not capture the complex, non-linear relationships inherent in regional data.

## ***2.1. Regional Polarization in Russia***

Regional polarization in Russia has emerged as a significant area of study in recent decades, especially considering the country's complex socio-political landscape. Tebekin [11] believes that polarization refers to the growing divergence between different regions, often in terms of political alignment, economic development, and social identities. In the context of Russia, regional disparities have been shaped by historical, economic, and political factors. Altunina [12] suggested that the legacy of Soviet centralization and its dissolution created a landscape of uneven development, with Moscow and St. Petersburg historically benefiting from political power and economic resources, while peripheral regions lagged.

Taymaz [13] has documented how Russia's vast territorial expanse has given rise to distinct regional identities, often influenced by local economic conditions, ethnic compositions, and varying degrees of state control. For example, the northern regions like Siberia and the Far East have struggled with population decline and economic stagnation, while the southern Caucasus and Volga regions have seen political instability due to ethnic and nationalistic movements. Additionally, the rise of regional elites has often further exacerbated these disparities, with local leaders seeking to consolidate power and influence at the expense of national unity.

Polarization in Russia is not merely an economic or political phenomenon but also a cultural one. Gluschenko [14] highlights how cultural and social norms vary widely between regions, influencing public opinion and political preferences. For instance, urban centers like Moscow tend to exhibit more liberal values, while rural and remote regions may retain more conservative or traditional ideologies. This dichotomy has played a key role in shaping regional voting behavior, contributing to an overall sense of division within the Russian Federation.

## ***2.2. Machine Learning in Social Science Research***

The study of regional polarization has traditionally been dominated by econometric and statistical methods. Foundational approaches, such as the use of the Lorenz curve and Gini coefficient for measuring income inequality [15], have provided valuable high-level insights but often fail to capture the nuanced, non-linear, and spatial complexities inherent in regional data. While regression analysis and Geographic information system (GIS) tools have been combined to map disparities [16], these linear models are limited in their ability to identify latent patterns and complex interactions within large, multidimensional datasets.

In recent years, machine learning (ML) has emerged as a powerful paradigm to overcome these limitations. Unlike traditional methods that often test pre-defined hypotheses, ML is adept at uncovering hidden structures and non-linear relationships directly from the data. This is particularly valuable for polarization research, where the drivers are often complex and interlinked. As noted by Amiri et al. [17], the full potential of these advanced algorithms in this field remains underexplored.

Two families of ML techniques have shown significant promise in economic and social research:

- 1) Clustering Algorithms (e. g., k-means, hierarchical clustering) are used to group regions based on shared socio-economic characteristics, moving beyond simple geographical proximity to identify “convergence clubs” or polarized clusters. For instance, Bergal [18] demonstrated the efficacy of k-means in categorizing regions for targeted policymaking.
- 2) Dimensionality Reduction Techniques (e. g., PCA, t-SNE) simplify complex datasets by distilling numerous indicators into a manageable set of core components, thereby clarifying the fundamental axes of disparity, as exemplified by Rogot’s [19] application of PCA.

The power of this ML-driven approach is demonstrated by several international case studies. Research on Ukraine by Ageyev et al. [20] effectively combined Ward’s method, k-means, and factor analysis to reveal a stable two-cluster structure of regional development that persisted for nearly a decade, a deep polarization that traditional  $\beta$ -convergence models averaged out. Similarly, studies on the United States have used ML to decode the socio-political drivers of polarization. Viale & Binns [21] identified key factors in rural-urban voting divides, while Terlizzi & Cohen [22] and Oberlander [23] highlighted how divergence in social policy (e. g., Medicaid expansion) creates self-reinforcing cycles of inequality in access to healthcare and other key resources.

Even in contexts of overall convergence, such as within the European Union, ML analyses have revealed a more complex reality. Studies by Giannini & Martini [24] and Lang [25] confirmed a trend of absolute convergence at the macro level. However, by applying advanced techniques like the XGBoost algorithm, Lang [25] was able to model non-linearities and spatial dependencies, showing that this convergence was driven by investments in human and physical capital and did not result in fixed “clubs”, but rather a dynamic core-periphery structure.

These international examples underscore a critical insight: regional polarization is a multidimensional phenomenon driven by a confluence of economic, social, and infrastructural factors. They also collectively demonstrate that machine learning is uniquely capable of detecting the latent patterns and structural drivers that traditional econometric approaches often miss.

### ***2.3. Gap in Literature and Contribution of the Study***

While there is a growing body of research on regional polarization in Russia and the use of machine learning in political science, there remains a significant gap in studies that combine these two areas in the context of Russia. Previous studies have largely focused on either identifying the socio-political divides or employing traditional methods like regression analysis to understand these patterns. Few have applied advanced machine learning techniques to systematically

identify and map regional polarization in Russia using a wide range of social, economic, and political indicators.

This study aims to fill this gap by employing a machine learning approach to analyze regional polarization in Russia. By using clustering algorithms and supervised learning techniques, the research will identify latent patterns of polarization across Russia's regions, accounting for a variety of factors, including demographic variables, economic disparities, political affiliations, and cultural divisions. The use of machine learning allows for a more nuanced understanding of how these factors interact and contribute to the regional divides, enabling policymakers and researchers to better understand the underlying forces shaping Russia's socio-political landscape.

### 3. Materials and Methods

This study is based on the use of approaches from the theory of dynamical systems and machine learning. The region's state (determined per year) is characterized by the value of numerous socio-economic indicators: gross regional product ( $x_1$ ), unemployment rate ( $x_2$ ), Gini index ( $x_3$ ), poverty rate ( $x_4$ ), hospitals number ( $x_5$ ), vehicles available ( $x_6$ ), population size ( $x_7$ ), and migration trends ( $x_8$ ).

The eight-dimensional vector of the region's state ( $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_8$ ) can be projected onto one- and two-dimensional spaces. The set of states of all regions over the entire study period can be divided into several subsets (classes, clusters), in each of which the vectors of states will be closest to each other. All states are considered independent (the "ergodic hypothesis").

The transition of a region from one cluster to another means that there is a process of divergence with the remaining regions of the first cluster and a process of convergence with the regions of the second cluster.

This study includes the following steps: (1) Data preparation; (2) Determining the optimal number of clusters in the state space; (3) Classification of the states of regions in eight-dimensional spaces; and (4) Predictive modeling for future polarization trends.

To carry out the necessary calculations, the following application software packages (libraries) were used — *Pandas*<sup>1</sup> (for data analysis), *NumPy*<sup>2</sup> (for mathematical operations on multidimensional data arrays performance), *Sklearn*<sup>3</sup> (for machine learning), and others.

#### 3.1. Data Collection, Integration and Preparation

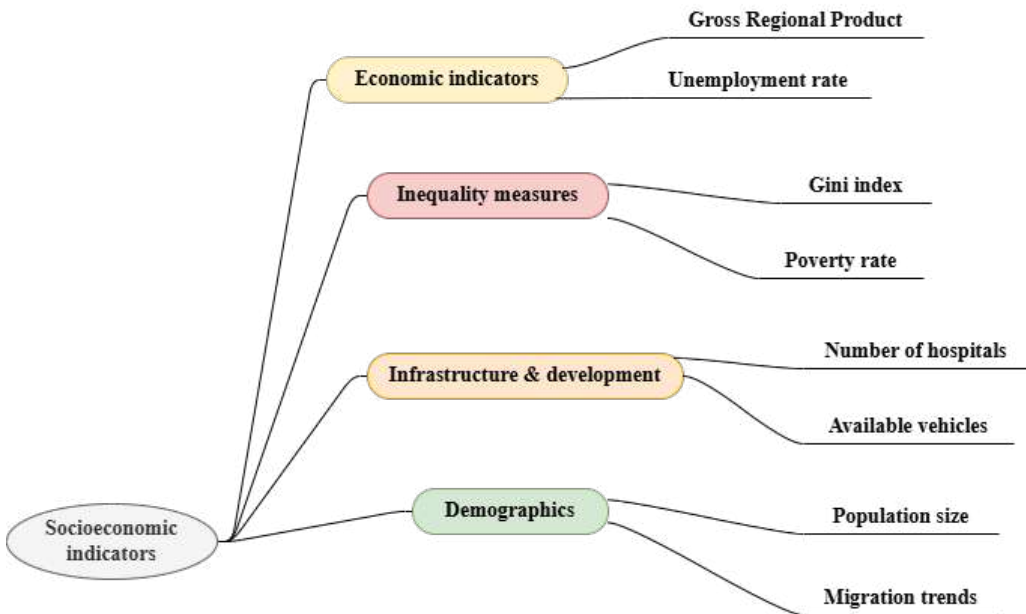
This study utilizes panel data from 85 regions across 2018–2024, covering key socio-economic indicators as showcased by Figure 1. Data is compiled according to statistical reports from the Federal state of statistics<sup>4</sup>.

<sup>1</sup> Pandas. Data analysis tool. URL: <https://pandas.pydata.org/> (Date of access: February, 2025)

<sup>2</sup> NumPy. The fundamental package for scientific computing with Python. URL: <https://numpy.org/> (Date of access: February, 2025)

<sup>3</sup> Scikit-learn. Efficient tools for predictive data analysis. URL: <https://scikit-learn.org/stable/index.html> (Date of access: February, 2025)

<sup>4</sup> Federal State Statistics Service. URL: <https://rosstat.gov.ru/> (Date of access: February 2025)



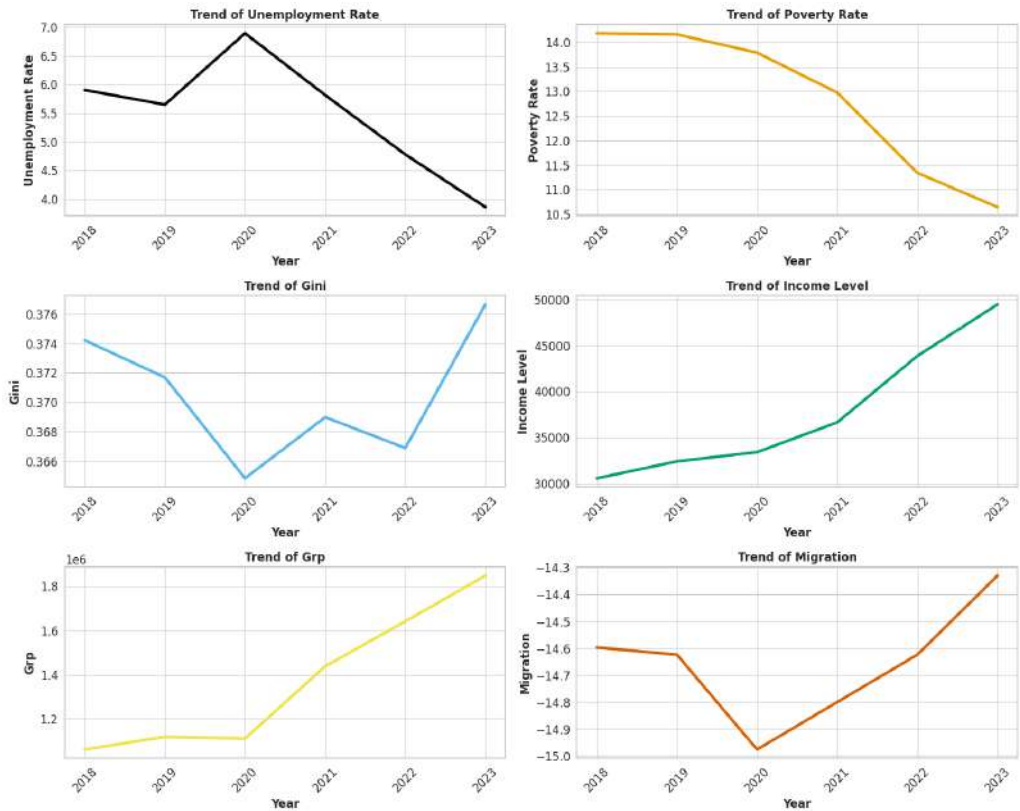
**Figure 1.** Regional Socioeconomic indicators

*Source:* compiled by the authors

To analyze the multi-year dataset, we first consolidated all annual files into a single structured dataset in long format, appending a ‘Year’ column during the merging process while standardizing variable names to account for administrative changes over time. Missing values were addressed through linear interpolation to ensure data continuity.

To capture polarization dynamics, we derived two sets of features: (1) growth rates, including annualized changes in Gini coefficients and GRP, and (2) volatility measures calculated as the standard deviation of unemployment and poverty rates over time [26]. This approach enables systematic tracking of economic disparities and stability patterns across the study period. Figure 2 depicts socioeconomic trends observed in Russia over the years.

An analysis of Russia’s socio-economic indicators for the period from 2018 to 2023 identifies key trends that explain changes in regional polarization and point to their causes. The increase in unemployment in 2020 was caused by the COVID-19 pandemic, which led to massive job cuts and business shutdowns, but the subsequent decline in unemployment is due to economic recovery and government support measures. The poverty rate began to decrease after 2021 due to economic growth, social programs and an increase in household incomes, which contributed to an improvement in living conditions in the regions. Nevertheless, the Gini coefficient remains high, indicating persistent income inequality, widening the gap between rich and poor regions. The increase in the income level of the population, although it helps to reduce poverty, at the same time highlights



**Figure 2.** Socioeconomic trends in Russia

*Source:* compiled by the authors

the problem of uneven distribution of wealth. The increase in population outflow observed since 2020 is associated with economic instability, the search for better living conditions, and socio-political factors, which increases polarization as less developed regions lose their workforce.

According to machine learning best practices, it is better to normalize data before training the model [27]. To do so, we use the function *StandardScaler* from *sklearn*, which is mathematically defined as:

$$X_{scaled} = \frac{X - \bar{X}}{\sigma(X)}, \quad (1)$$

where  $X$  — the original feature,  $\bar{X}$  — the mean of  $X$ ,  $\sigma(X)$  — the standard deviation of  $X$ , and  $X_{scaled}$  — the scaled feature.

### 3.2. Optimal Cluster Determination

When clustering data, it is necessary to specify in advance the number of clusters into which the initial data set will be divided. This can be done using an expert

method or by comparing classification results for different numbers of specified clusters. In this paper, we use the elbow and silhouette methods to determine the optimal number of clusters (Figure 3).

### 3.2.1 Elbow Method

Thorndike [28] proposed the elbow method. The essence of this method is to classify the source data for a different number of clusters  $k$ , for example, from 2 to 8. For each value of  $k$ , the sum of error squares is calculated and a graph of the sum of error squares versus the number of clusters is displayed. If we represent such a graph in the form of an arm, then the point where the curve of the graph most resembles the bend of the elbow will give us the best value of  $k$ .

In other words, it is necessary to determine the smallest number of clusters in which the sum of error squares remains small. This error is calculated by the formula:

$$WCSS(k) = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2, \quad (2)$$

where  $C_i$  — cluster  $i$ ,  $\mu_i$  — centroid of  $C_i$ ;  $\|x - \mu_i\|$  — Euclidean distance from point  $x$  to its centroid.

The interpretation of this method is the following — as the number of clusters  $k$  increases,  $WCSS$  decreases. The optimal  $k$  is where the rate of decrease sharply slows (the “elbow”).

### 3.2.2. Silhouette Score

The Silhouette Score measures how well each data point fits its assigned cluster compared to other clusters. It ranges from  $-1$  (worst) to  $+1$  (best). For each data point  $i$ , we must:

1. Calculate the average intra-cluster distance (cohesion):

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j), \quad (3)$$

where  $C_i$  — cluster containing point  $i$ ,  $d(i, j)$  — distance between points  $i$  and  $j$ ,  $|C_i|$  — number of points in  $C_i$ .

2. Calculate the smallest average inter-cluster distance (separation):

$$b(i) = \min_{k \neq C_i} \left( \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \right), \quad (4)$$

where  $C_k$  — any other cluster.

3. Compute the Silhouette Coefficient for point  $i$ :

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (5)$$

4. Calculate the overall Silhouette Score (average across all points):

$$S = \frac{1}{N} \sum_{i=1}^N s(i). \quad (6)$$

The silhouette score can have the following interpretation — if  $S \approx 1$ , then we can conclude that data points are well-clustered (points are close to their cluster and far from others), if  $S \approx 0$ , then we have overlapping clusters, if  $S \approx -1$ , then data points are misclassified.

The determination of the optimal number of clusters is presented in Figure 3. Based on this figure we can conclude that it's better to classify our regions into four groups.

### 3.3. Clustering Method (*k*-means)

Among the various clustering algorithms, the *K-means* method is one of the most widely adopted due to its simplicity, computational efficiency, and scalability. The algorithm operates by iteratively assigning data points to the nearest cluster centroid and updating the centroids based on the current cluster assignments, with the objective of minimizing the within-cluster variance [29].

The K-means algorithm begins by selecting an initial set of centroids, typically either randomly or through a more sophisticated initialization method such as *K-means++*, which helps improve convergence by spreading the initial centroids apart [30]. Once the centroids are initialized, the algorithm proceeds in an iterative manner, alternating between two key steps: assignment and update. In the assignment step, each data point is assigned to the nearest centroid based on Euclidean distance, effectively forming clusters [31]. Subsequently, in the update step, the centroids are recalculated as the “mean” of all points within their respective clusters. This process continues until a stopping criterion is met, such as minimal changes in centroid positions or a predefined number of iterations [32].

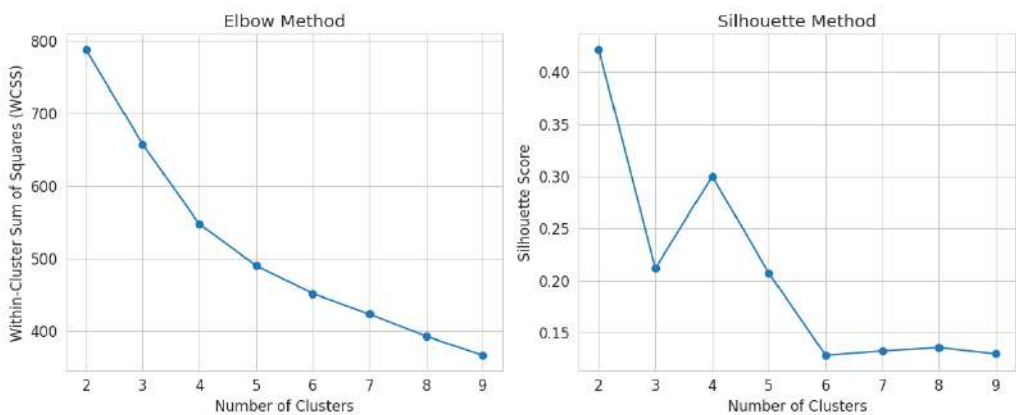


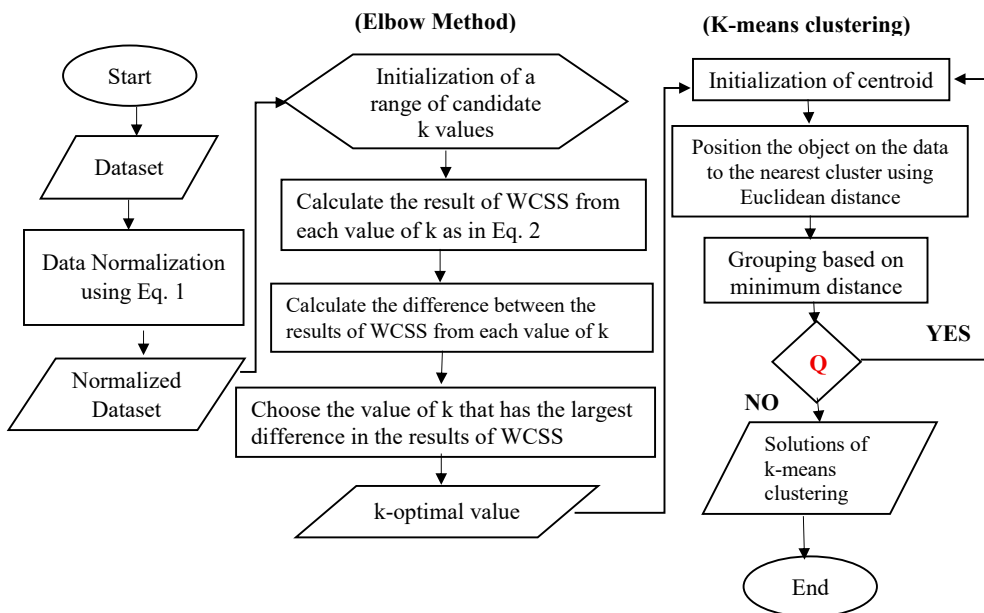
Figure 3. Optimal number of clusters determination

Source: compiled by the authors

A significant advantage of the *K-means* method lies in its computational efficiency, making it particularly suitable for large datasets. The algorithm's time complexity is linear with respect to the number of data points, dimensions, and iterations, ensuring scalability. Additionally, its straightforward implementation and interpretability contribute to its widespread use in practical applications. However, the method is not without limitations. One notable drawback is its sensitivity to the initial placement of centroids, which can lead to suboptimal clustering solutions.

Another limitation of *K-means* is the requirement to predefine the number of clusters,  $K$ , which may not always be known in advance. Heuristic approaches, such as the Elbow method or Silhouette analysis, are often employed to estimate an appropriate value for  $K$ . Furthermore, the algorithm assumes that clusters are spherical and equally sized, which may not hold true for datasets with irregular or non-convex cluster shapes. Additionally, *K-means* is sensitive to outliers, as the mean-based centroid computation can be heavily influenced by extreme values.

Despite these limitations, *K-means* remains a foundational tool in unsupervised learning with diverse applications. It is commonly used in customer segmentation to group users with similar purchasing behaviors, in image compression for reducing color space by clustering pixel intensities, and in document clustering to organize textual data into thematic categories. Variants of the algorithm, such as *Fuzzy C-means* and *Mini-Batch K-means*, have been developed to address specific challenges, including soft clustering and large-scale data processing. Figure 4 depicts the steps to follow when implementing the *k-means* method.



**Figure 4.** K-means clustering with optimal k selection. Q represents the criteria whether objects are still switching clusters

Source: compiled by the authors

### 3.4. Predictive Modeling for Future Polarization Trends

To predict future regional polarization trends, we employ two advanced ensemble learning techniques: Random Forest (RF) and eXtreme Gradient Boosting (XGBoost). Both methods leverage multiple decision trees to enhance predictive accuracy while mitigating overfitting.

#### 3.4.1. Random forest (RF) Model

Random Forest is a bagging (bootstrap aggregating) method that constructs numerous decision trees during training and outputs the mean prediction (for regression) of individual trees [33]. The model introduces randomness by: (1) Bootstrap sampling. Each tree is trained on a random subset of the data  $D_{d^p}$  sampled with replacement from the original dataset  $D$ ; (2) Feature randomness. At each split, only a random subset of features  $m$  (where  $m < p$  with  $p$  being the total features) is considered, reducing correlation among trees.

The prediction for a new input  $x$  is given by:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x), \quad (7)$$

where  $B$  is the number of trees and  $T_b(x)$  is the prediction of the  $b^{\text{th}}$  tree.

#### 1. XGBoost Model

XGBoost — is a gradient boosting framework that optimizes a differentiable loss function by iteratively adding weak learners [34]. Unlike RF, XGBoost builds trees sequentially, where each new tree corrects errors from previous ones. The objective function consists of:

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad (8)$$

where  $l(y_i, \hat{y}_i)$  is the loss function (e. g., squared error for regression),

$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2$  is the regularization term (penalizing tree complexity),

$T$  is the number of leaves, and  $w$  is the leaf weights.

At each iteration  $t$ , the model adds a new tree  $f_t$  to minimize [35]:

$$L^{(t)} \approx \sum_{i=1}^n \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \quad (9)$$

where  $g_i = \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1})$  and  $h_i = \partial_{\hat{y}^{t-1}}^2 l(y_i, \hat{y}^{t-1})$  are the first and second-order gradients of the loss function.

## 2. Model evaluation metric — Mean Squared Error (MSE)

The performance of both models is evaluated using Mean Squared Error (MSE), which measures the average squared difference between predicted ( $\hat{y}_i$ ) and actual ( $y_i$ ) polarization values:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (10)$$

Lower MSE values indicate better predictive accuracy. By comparing the MSE of RF and XGBoost, we determine which model better captures regional polarization dynamics for future trend forecasting [36, 37]. These methods provide robust, interpretable, and high-performance solutions for analyzing complex socio-political trends, enabling policymakers to anticipate and mitigate rising polarization [20, 38–40].

## 4. Results

The application of the elbow method and subsequent validation using the Fisher criterion indicated that the optimal number of clusters for the Russian regions is four ( $k = 4$ ). The sharp decrease in the total within-cluster sum of squares when moving from  $k = 3$  to  $k = 4$ , followed by a plateau for  $k > 4$ , confirms this choice (Figure 3). The resulting clustering of regions is presented in Table 1, and the characteristic socio-economic profiles of each cluster are visualized in Figure 5.

Cluster 0 comprises predominantly remote republics and autonomous okrugs (e. g., Altai Republic, Chukotka, Yamalo-Nenets). Cluster 1 includes the most developed economic centers, such as Moscow, St. Petersburg, and resource-rich regions like Khanty-Mansi Autonomous Okrug. Cluster 2 represents a middle-tier group with a mix of moderately developed industrial regions and struggling territories from Siberia and the North Caucasus. Cluster 3 is dominated by the least developed regions, particularly in the North Caucasus (e. g., Chechnya, Ingushetia, Dagestan).

Table 1. Clustering of regions according to the level of socio-economic development, considering regional polarization

Cluster	Regions
0	Altai Republic, Chechen Republic, Chukotka Autonomous Okrug, Kabardino-Balkarian Republic, Karachay-Cherkess Republic, Nenets Autonomous Okrug, Republic of Buryatia, Republic of Crimea, Republic of Ingushetia, Republic of North Ossetia–Alania, Sakhalin Oblast, Tuva Republic, Yamalo-Nenets Autonomous Okrug, Zabaykalsky Krai
1	Altai Krai, Amur Oblast, Arkhangelsk Oblast (excluding autonomous okrug), Astrakhan Oblast, Belgorod Oblast, Bryansk Oblast, Chelyabinsk Oblast, Chukotka Autonomous Okrug, Chuvash Republic, Irkutsk Oblast, Ivanovo Oblast, Jewish Autonomous Oblast, Kaliningrad Oblast, Kaluga Oblast, Kamchatka Krai,

Continuation of table 1

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

End of table 1

Cluster	Regions
3	Republic of Ingushetia, Republic of Kalmykia, Republic of Karelia, Republic of Khakassia, Republic of Mordovia, Republic of North Ossetia–Alania, Republic of Tatarstan, Rostov Oblast, Ryazan Oblast, Saint Petersburg, Sakha Republic (Yakutia), Samara Oblast, Saratov Oblast, Sevastopol, Smolensk Oblast, Stavropol Krai, Sverdlovsk Oblast, Tambov Oblast, Tomsk Oblast, Tula Oblast, Tuva Republic, Tver Oblast, Tyumen Oblast (excluding autonomous okrugs), Udmurt Republic, Ulyanovsk Oblast, Vladimir Oblast, Volgograd Oblast, Vologda Oblast, Voronezh Oblast, Yaroslavl Oblast, Zabaykalsky Krai

Source: compiled by the authors.

The analysis of cluster characteristics (Figure 5) reveals stark contrasts. Cluster 0 has the highest GRP per capita (4,086,395 rubles) but suffers from high inequality (Gini 0.39), unemployment (8.83 %), and severe outmigration (−24.48). Conversely, Cluster 1 exhibits balanced development with low unemployment (3.96 %) and poverty (11.26 %), and high infrastructure development. Cluster 2 shows average economic indicators but the highest poverty rate (14.80 %), while Cluster 3 is distinctive for its positive net migration (+10.15) coupled with weak economic performance and high reliance on federal subsidies.

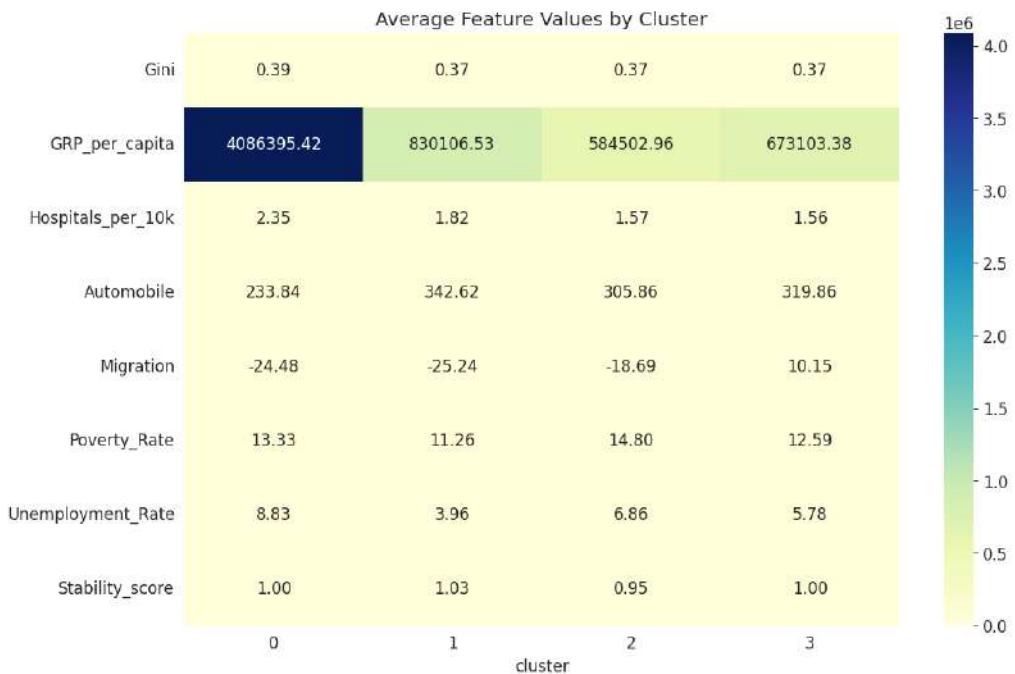


Figure 5. Clusters characteristics

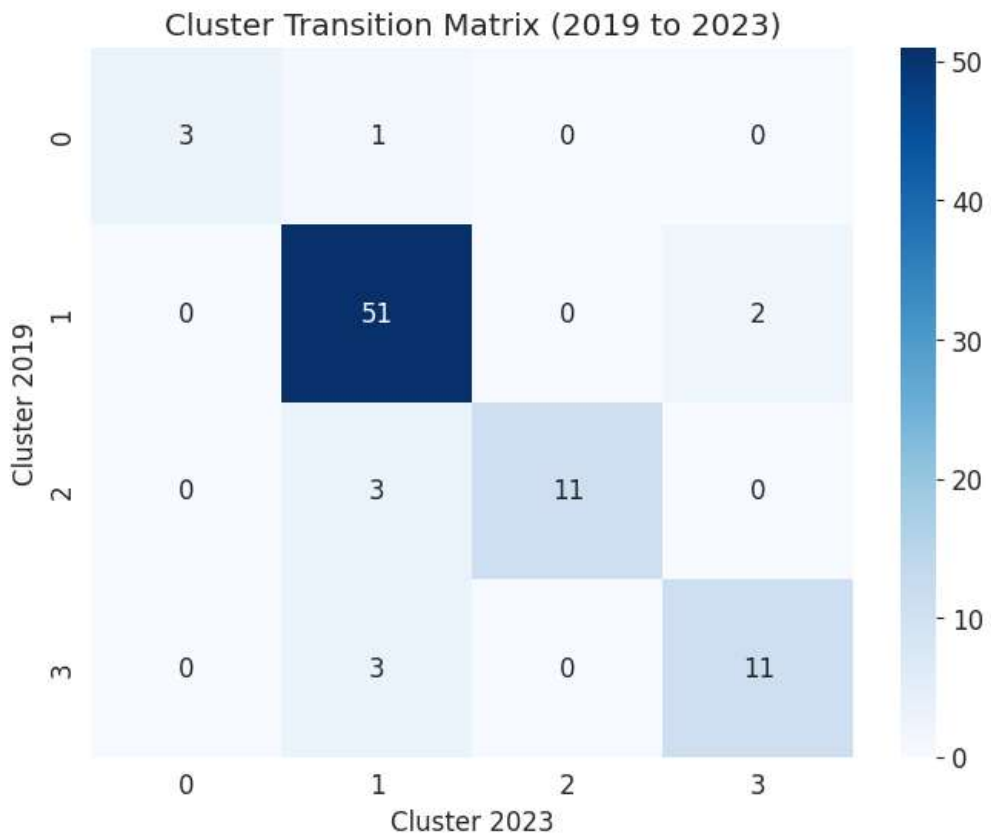
Source: compiled by the authors

To understand the dynamics of polarization, we analyzed cluster transitions between 2019 and 2023 (Figure 6). The transition matrix and the visualization of regional trajectories (Figure 7) indicate fluid movement between clusters, with some regions demonstrating convergent (catch-up) paths and others following divergent (falling-behind) trajectories.

For predictive analysis, we employed Random Forest and XGBoost algorithms. The Random Forest model demonstrated superior performance with a lower Mean Squared Error ( $MSE = 0.132$ ) compared to XGBoost ( $MSE = 0.146$ ), confirming its robustness for this type of socio-economic data.

## 5. Discussion

This study set out to map and forecast the complex landscape of regional polarization in Russia. Our findings not only provide a detailed snapshot of regional disparities but also offer insights into the underlying dynamics and drivers. The results largely confirm our initial hypotheses and align with, yet also complicate, existing literature on regional science.



**Figure 6.** Cluster transition matrix

*Source:* compiled by the authors

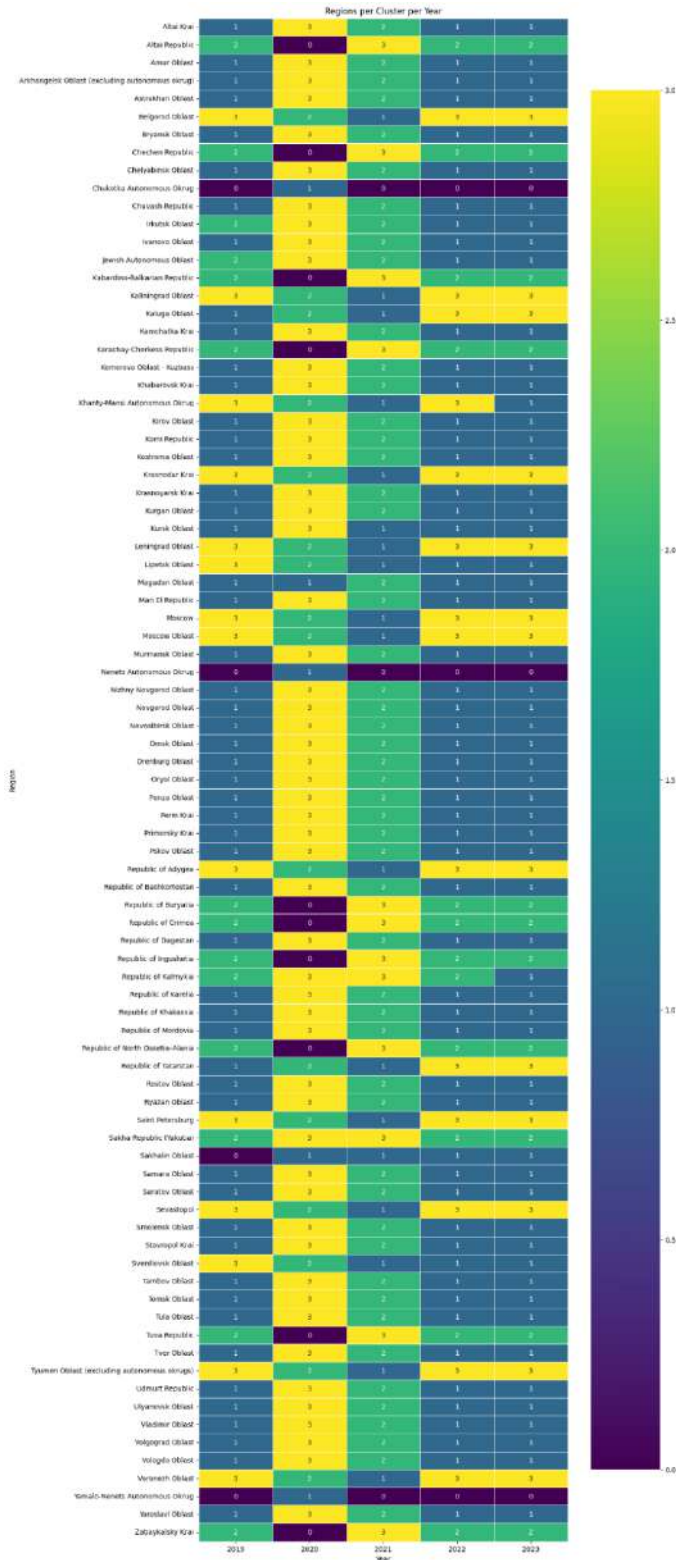


Figure 7. Convergence and divergence trajectories of some regions

Source: compiled by the authors

### 5.1. Interpretation of Findings and Hypothesis Verification

*H1: The Multidimensional Nature of Polarization.* Our hypothesis that polarization is driven by social and infrastructural factors alongside economic ones is strongly supported. The distinct profiles of the four clusters cannot be explained by GRP alone. For instance, Cluster 0 exemplifies the “resource curse,” where extreme economic wealth coexists with severe social challenges like high inequality, outmigration, and relatively poor infrastructure (as seen in low automobile ownership). This aligns with studies by Ahrend [41] and Maximova [42], who noted similar disconnects between resource wealth and human development in Russian regions. Conversely, Cluster 3’s positive migration is a social phenomenon not reflected in its weak economic indicators, suggesting that factors like ethnic homogeneity, cultural ties, or public sector employment drive demographic trends independently of economic performance. This finding supports the conclusion made by Sitkevich [43], who argued that non-economic factors are crucial for understanding demographic dynamics in the North Caucasus.

*H2: The Coexistence of Convergence and Divergence.* The analysis of cluster transitions (Figures 5 and 6) provides clear evidence for H2. We observe simultaneous processes of convergence and divergence. The movement of some regions from lower-tier clusters (e. g., Cluster 3) to more developed ones (e. g., Cluster 2) signals catch-up growth, potentially driven by targeted federal investment or local policy improvements. However, the stability of the core-periphery divides with Clusters 0 and 1 remaining largely distinct highlights persistent divergence. This core-periphery pattern is a classic finding in regional studies [44], but our results show that the Russian periphery is itself highly differentiated between a resource-rich, sparsely populated Arctic/Siberian group (Cluster 0) and a demographically growing but economically stagnant Southern group (Cluster 3). The trajectories suggest that federal policies have been more successful at fostering convergence among mid-tier regions than at bridging the fundamental gap between the core and the most challenging peripheries.

*H3: The Value of Machine Learning.* Our study confirms H3, demonstrating that ML techniques can uncover nuanced patterns that might be oversimplified by conventional methods. Traditional regression analysis might identify correlations between variables, but the clustering approach revealed four distinct, non-linear combinations of these variables. For example, it clearly separated high-GRP/high-inequality regions (Cluster 0) from moderate-GRP/low-inequality regions (Cluster 1), a distinction crucial for policy. The predictive power of the Random Forest model further underscores this value. Its ability to forecast cluster transitions based on a multivariate set of indicators provides a dynamic tool that goes beyond the static snapshots typically produced by econometric models. This aligns with a growing body of literature [45, 46] advocating for ML in regional studies due to its ability to handle complex, interacting variables.

## 5.2. *Limitations of the Study*

While this research provides valuable insights, several limitations should be acknowledged.

First, the analysis relies on official regional statistics, which can vary in reliability and may not fully capture informal economic activities or subjective well-being.

Second, our clustering is based on data from a specific period (2019–2023), and the models assume a continuation of past trends. Structural breaks caused by major geopolitical events, drastic policy shifts, or economic shocks could alter these trajectories in ways the model cannot anticipate.

Third, the selection of variables, while comprehensive, is not exhaustive. Factors such as institutional quality, social capital, or environmental conditions were not included but could influence polarization.

Finally, while the ML models show good predictive accuracy, they operate as “black boxes” to some extent. The complex interplay of variables making a region prone to transition requires further qualitative investigation to be fully understood and actionable for policymakers.

## 6. Conclusion

This study has validated a multidimensional and dynamic approach to understanding regional polarization in Russia, demonstrating that economic, social, and infrastructural disparities are deeply intertwined. By employing a suite of machine learning techniques including k-means clustering, Random Forest, and XGBoost, we have moved beyond diagnosis to forecast future trends, confirming the efficacy of these data-driven methods for analyzing complex socio-economic systems. The analysis reveals a Russian landscape characterized by stark and persistent divides, organized into four distinct regional clusters.

The theoretical significance of this research lies in its contribution to the methodological framework for studying regional development. It demonstrates the superior capability of machine learning ensemble methods, particularly Random Forest, over traditional econometric models in capturing the non-linearities and complex interactions that define regional polarization. By successfully identifying a stable, four-cluster structure of Russia’s regions, this study provides a novel, data-driven typology that challenges simpler core-periphery models and offers a more nuanced theoretical lens for understanding spatial inequality in resource-based federal states.

The practical significance of the findings is direct and substantial for policymakers. The validated cluster typology provides a rigorous foundation for moving beyond ineffective one-size-fits-all strategies toward precisely differentiated regional policies. For instance, the model clarifies that policies for resource-rich Cluster 0 must focus on transforming resource wealth into sustainable local development and reversing outmigration, while strategies for lagging Cluster 3

must prioritize economic diversification to break the cycle of federal dependency. Furthermore, the predictive accuracy of the Random Forest model (MSE = 0.132) offers an actionable tool for proactive governance, allowing authorities to identify regions at risk of further divergence and allocate resources with greater efficiency and foresight.

## References

1. Kuzin, V.Yu. (2023). An evaluation of the Russian Far East spatial polarization in the post-Soviet period. *Vestnik of North-Eastern Federal University Series "Earth Sciences"*, No. 2, 102–113. (In Russ.). <https://doi.org/10.25587/svfu.2023.30.2.009>
2. Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., Flammini, A. (2021). Political polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5, No. 1, 89–96. <https://doi.org/10.1609/icwsm.v5i1.14126>
3. Da Silva, T.P. (2023). Learning beyond the spatial autocorrelation structure: A machine learning- based approach to discovering new patterns and relationships in the context of spatially contextualized modeling of voting behavior. *Doctoral Thesis*. Sao Carlos. <https://doi.org/10.11606/t.55.2023.tde-15012024-174102>
4. Levy, R. (2021). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review*, Vol. 111, No. 3, 831–870. <https://doi.org/10.1257/aer.20191777>
5. Aharon-Gutman, M., Schaap, M., Lederman, I. (2018). Social topography: Studying spatial inequality using a 3D regional model. *Journal of Rural Studies*, Vol. 62, 40–52. <https://doi.org/10.1016/j.jrurstud.2018.06.010>
6. Martin, R., Sunley, P. (2020). Regional economic resilience: evolution and evaluation. In: *Handbook on Regional Economic Resilience*. Edited by G. Bristow, A. Healy. Edward Elgar Publishing, 10–35. <https://doi.org/10.4337/9781785360862.00007>
7. Aharon-Gutman, M., Burg, D. (2019). How 3D visualization can help us understand spatial inequality: On social distance and crime. *Environment and Planning B: Urban Analytics and City Science*, Vol. 48, Issue 4, 793–809. <https://doi.org/10.1177/2399808319896524>
8. Stephens, J.D. (2015). Thomas Piketty (2014), *Capital in the Twenty-First Century*. Translated by Arthur Goldhammer. Cambridge, Massachusetts: Belknap Press of Harvard University Press, 685 pp., *Journal of Social Policy*, Vol. 45, Issue 1, 172–173. <https://doi.org/10.1017/s0047279415000616>
9. Marco, C., Lenka, J., Christa, K.S. (2024). *The Future of EU Cohesion: Scenarios and Their Impacts on Regional Inequalities*. Belgium: European Parliamentary Research Service (EPRS), 61 p. Available at: [https://www.europarl.europa.eu/RegData/etudes/STUD/2024/762854/EPRS\\_STU\(2024\)762854\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2024/762854/EPRS_STU(2024)762854_EN.pdf)
10. Glaeser, E.L. (2021). Urban resilience. *Urban Studies*, Vol. 59, Issue 1, 3–35. <https://doi.org/10.1177/00420980211052230>
11. Tebekin M. V. (2023). Methodological approach to the assessment of spatial polarization of regions. *Applied Economic Research*, No. 4, 86–94. (In Russ.). [https://doi.org/10.47576/2949-1908\\_2023\\_4\\_86](https://doi.org/10.47576/2949-1908_2023_4_86)
12. Altunina, V., Anuchina, D. (2023). Assessment of the level of spatial polarization of Russian regions. *Journal of Economics Entrepreneurship and Law*, Vol. 13, No. 5, 1319–1340. (In Russ.). <https://doi.org/10.18334/epp.13.5.117516>
13. Taymaz, E. (2022). Regional convergence or polarization: the case of the Russian Federation. *Regional Research of Russia*, Vol. 12, 469–482. <https://doi.org/10.1134/s2079970522700198>
14. Gluschenko, K.P. (2023). Regional inequality in Russia: Anatomy of convergence. *Regional Research of Russia*, Vol. 13, Suppl. 1, S1–S12. <https://doi.org/10.1134/s207997052360004x>

15. Atkinson, A.B. (1970). On the measurement of inequality. *Journal of Economic Theory*, Vol. 2, Issue 3, 244–263. [https://doi.org/10.1016/0022-0531\(70\)90039-6](https://doi.org/10.1016/0022-0531(70)90039-6)
16. Chi, S., Grigsby-Toussaint, D.S., Bradford, N., Choi, J. (2013). Can Geographically Weighted Regression improve our contextual understanding of obesity in the US? Findings from the USDA Food Atlas. *Applied Geography*, Vol. 44, 134–142. <https://doi.org/10.1016/j.ap-geog.2013.07.017>
17. Amiri, M., Pourghasemi, H.R., Ghanbarian, G.A., Afzali, S.F. (2019). Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. *Geoderma*, Vol. 340, 55–69. <https://doi.org/10.1016/j.geoderma.2018.12.042>
18. Bergal, B. (2022). Audit of the effectiveness of cluster policy implementation in the region. *Pskov Region-logical Journal*, Vol. 18, No. 2, 154–167. <https://doi.org/10.37490/S221979310019289-0>
19. Rogot, A. (2023). Dimensionality reduction techniques in Macroeconomic Analysis. *CUNY Academic Works*. Available at: [https://academicworks.cuny.edu/bb\\_etds/166](https://academicworks.cuny.edu/bb_etds/166)
20. Chagovets, L., Chahovets, V., Chernova, N. (2020). Machine Learning Methods Applications for Estimating Unevenness Level of Regional Development. In: *Data-Centric Business and Applications. Evolvments in Business Information Processing and Management*. Vol. 3. Edited by D. Ageyev, T. Radivilova, N. Kryvinska. Springer Cham, 115–139. [https://doi.org/10.1007/978-3-030-35649-1\\_6](https://doi.org/10.1007/978-3-030-35649-1_6)
21. Veale, M., Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, Vol. 4, Issue 2, 205395171774353. <https://doi.org/10.1177/2053951717743530>
22. Terlizzi, E.P., Cohen, R.A. (2023). *Geographic Variation in Health Insurance Coverage: United States, 2022*. National Health Statistics Reports. <https://doi.org/10.15620/cdc:133320>
23. Oberlander, J. (2024). Polarization, partisanship, and health in the United States. *Journal of Health Politics Policy and Law*, Vol. 49, Issue 3, 329–350. <https://doi.org/10.1215/03616878-11075609>
24. Giannini, M., Martini, B. (2024). Regional disparities in the European Union. A machine learning approach. *Papers of the Regional Science Association*, Vol. 103, Issue 4, 100033. <https://doi.org/10.1016/j.pirs.2024.100033>
25. Lange, T. (2015). Socio-economic and political responses to regional polarisation and socio-spatial peripheralisation in Central and Eastern Europe: A research agenda. *Hungarian Geographical Bulletin*, Vol. 64, No. 3, 171–185. <https://doi.org/10.15201/hungeobull.64.3.2>
26. Druckman, J.N., Levendusky, M.S. (2024). Correction to: What Do We Measure When We Measure Affective Polarization? *Public Opinion Quarterly*, Vol. 88, Issue 3, 1095–1096. <https://doi.org/10.1093/poq/nfae051>
27. Jo, J. (2019). Effectiveness of normalization Pre-Processing of big data to the machine learning performance. *The Journal of the Korea Institute of Electronic Communication Sciences*, Vol. 14, Issue 3, 547–552. <https://doi.org/10.13067/jkiecs.2019.14.3.547>
28. Thorndike, R.L. (1953). Who belongs in the family? *Psychometrika*, Vol. 18, Issue 4, 267–276. <https://doi.org/10.1007/bf02289263>
29. Eltibi, M.F., Ashour, W.M. (2011). Initializing KMeans Clustering Algorithm using Statistical Information. *International Journal of Computer Applications*, Vol. 29, No. 7, 51–55. <https://doi.org/10.5120/3573-4930>
30. Currin, C.B., Vera, S.V., Khaledi-Nasab, A. (2022). Depolarization of echo chambers by random dynamical nudge. *Scientific Reports*, Vol. 12, 9234. <https://doi.org/10.1038/s41598-022-12494-w>
31. Behrens, T., Schmidt, K., Rossel, R.V., Gries, P., Scholten, T., MacMillan, R.A. (2018). Spatial modelling with Euclidean distance fields and machine learning. *European Journal of Soil Science*, Vol. 69, Issue 5, 757–770. <https://doi.org/10.1111/ejss.12687>

32. Likas, A., Vlassis, N., Verbeek, J.J. (2002). The global k-means clustering algorithm. *Pattern Recognition*, Vol. 36, Issue 2, 451–461. [https://doi.org/10.1016/s0031-3203\(02\)00060-2](https://doi.org/10.1016/s0031-3203(02)00060-2)
33. Breiman, L. (2001). Random forests. *Machine Learning*, Vol. 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
34. Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Vol. 29, No. 5, 1189–1232. <https://doi.org/10.1214/aos/1013203451>
35. Friedman, J., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, Vol. 28, No. 2, 337–407. <https://doi.org/10.1214/aos/1016218223>
36. Biau, G., Scornet, E. (2016). A random forest guided tour. *Test*, Vol. 25, 197–227. <https://doi.org/10.1007/s11749-016-0481-7>
37. Wright, M.N., Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, Vol. 77, Issue 1, 1–17. <https://doi.org/10.18637/jss.v077.i01>
38. Maksubova, D.M., Umargadzhieva, N.M., Aripova, P.G. (2024). Methodological approaches to assessing regional development. In: *Computational and Strategic Business Modelling*. Edited by D. P. Sakas, D. K. Nasiopoulos, Yu. Taratuhina. Springer Cham, 543–555. [https://doi.org/10.1007/978-3-031-41371-1\\_45](https://doi.org/10.1007/978-3-031-41371-1_45)
39. Amarasinghe, K., Rodolfa, K., Lamba, H., Ghani, R. (2020). Explainable Machine learning for public Policy: use cases, gaps, and research directions. *Data & Policy*, Vol. 5, e5. <https://doi.org/10.1017/dap.2023.2>
40. Kamal, S., Gullic, J., Bagavathi, A. (2022). Modeling Polarization on Social Media posts: A Heuristic approach using media Bias. In: *Foundations of Intelligent Systems. Proceedings of 26th International Symposium, ISMIS 2022*. Edited by M. Ceci, S. Flesca, E. Masciari, G. Manco, Z. W. Raś. Springer Cham, 35–43. [https://doi.org/10.1007/978-3-031-16564-1\\_4](https://doi.org/10.1007/978-3-031-16564-1_4)
41. Ahrend, R. (2005). Can Russia Break the “Resource Curse”? *Eurasian Geography and Economics*, Vol. 46, Issue 8, 584–609. <https://doi.org/10.2747/1538-7216.46.8.584>
42. Maximova, S.G., Omelchenko, D.A., Noyanzina, O.E. (2022). Human development, satisfaction with human capital and security in the Siberian and Far Eastern border regions. *RUDN Journal of Sociology*, Vol. 22, No. 3, 646–660. <https://doi.org/10.22363/2313-2272-2022-22-3-646-660>
43. Sitkevich, D.A. (2023). Economic and sociocultural factors of migration attitudes of residents of the North Caucasus. *Regional Research of Russia*, Vol. 13, Suppl 1, S78–S88. <https://doi.org/10.1134/s2079970523600166>
44. Berg, D.B., Balungu, D.M., Shelomentsev, A.G., Goncharova, K.S. (2024). Experimental trajectories of convergence and divergence processes of Russian regions population incomes inequality. *Journal of Applied Economic Research*, Vol. 23, No. 2, 364–393. (In Russ.). <https://doi.org/10.15826/vestnik.2024.23.2.015>
45. Balungu, D.M., Kumar, A. (2024). Forecasting the economic growth of Sverdlovsk Region: A comparative analysis of machine learning, linear regression and autoregressive models. *Journal of Applied Economic Research*, Vol. 23, No. 3, 674–695. <https://doi.org/10.15826/vestnik.2024.23.3.027>
46. Ketova, K., Kasatkina, E., Vavilova, D. (2021). Clustering Russian Federation Regions According to the Level of Socio-Economic Development with the Use of Machine Learning Methods. *Economic and Social Changes: Facts, Trends, Forecast*, Vol. 14, No. 6, 70–85. <https://doi.org/10.15838/esc.2021.6.78.4>

## INFORMATION ABOUT AUTHORS

### Daniel Musafiri Balungu

Post-Graduate Student, Assistant, Department of Big Data Analytics and Video Analysis Methods, Institute of Radio Electronics and Information Technologies, Ural Federal University named after the first President of Russia B. N. Yeltsin, Yekaterinburg, Russia (620002, Yekaterinburg, Mira street, 19); ORCID <https://orcid.org/0009-0001-5098-7603> e-mail: [danielbal03.db@gmail.com](mailto:danielbal03.db@gmail.com)

### **Anna Vyacheslavovna Rozanova**

Master Student, Department of Big Data Analytics and Video Analysis Methods, Institute of Radio Electronics and Information Technologies, Ural Federal University named after the first President of Russia B. N. Yeltsin, Yekaterinburg, Russia (620002, Yekaterinburg, Mira street, 19); ORCID <https://orcid.org/0009-0003-9803-0848> e-mail: [rozanna221132@icloud.com](mailto:rozanna221132@icloud.com)

### **Kristina Aleksandrovna Andreeva**

Master Student, Department of Big Data Analytics and Video Analysis Methods, Institute of Radio Electronics and Information Technologies, Ural Federal University named after the first President of Russia B. N. Yeltsin, Yekaterinburg, Russia (620002, Yekaterinburg, Mira street, 19); ORCID <https://orcid.org/0009-0009-5345-6160> e-mail: [kristinalezhnina88@gmail.com](mailto:kristinalezhnina88@gmail.com)

### **Anastasia Vasil'evna Solod**

Master Student, Department of Big Data Analytics and Video Analysis Methods, Institute of Radio Electronics and Information Technologies, Ural Federal University named after the first President of Russia B. N. Yeltsin, Yekaterinburg, Russia (620002, Yekaterinburg, Mira street, 19); ORCID <https://orcid.org/0009-0007-8795-1640> e-mail: [nsolodv@mail.ru](mailto:nsolodv@mail.ru)

### **Yutong Chen**

Master Student, Department of Big Data Analytics and Video Analysis Methods, Institute of Radio Electronics and Information Technologies, Ural Federal University named after the first President of Russia B. N. Yeltsin, Yekaterinburg, Russia (620002, Yekaterinburg, Mira street, 19); ORCID <https://orcid.org/0009-0001-7684-1530> e-mail: [kirito200207@gmail.com](mailto:kirito200207@gmail.com)







### **FOR CITATION**


Balungu, D.M., Rozanova, A.V., Andreeva, K.A., Solod, A.V., Chen, Yu. (2026). Identifying Patterns of Regional Polarization in Russia: A Machine Learning Approach. *Journal of Applied Economic Research*, Vol. 25, No. 1, 135–162. <https://doi.org/10.15826/vestnik.2026.25.1.005>

### **ARTICLE INFO**

Received May 9, 2025; Revised October 6, 2025; Accepted November 5, 2025.

## Выявление закономерностей региональной поляризации в России: подход машинного обучения

Д. М. Балунгу  , А. В. Розанова , К. А. Андреева ,  
А. В. Солод , Ю. Чэнь 

Уральский федеральный университет  
имени первого Президента России Б. Н. Ельцина,  
г. Екатеринбург, Россия  
 danielbal03db@gmail.com

**Аннотация.** Исследование региональной поляризации критически актуально для России, так как ярко выраженные социально-экономические диспропорции угрожают национальной экономической стабильности, социальной сплоченности и стратегической цели технологического суверенитета. Эти дисбалансы препятствуют эффективному распределению ресурсов и создают уязвимости в условиях глобальной перестройки. В данном исследовании анализируются закономерности и движущие силы региональной поляризации в России с использованием машинного обучения для моделирования сложного взаимодействия экономических и неэкономических факторов. В основе исследования лежат три гипотезы: поляризация обусловлена многомерными диспропорциями (H1); процессы сходимости и дивергенции сосуществуют (H2); а машинное обучение может эффективно выявлять скрытые поляризационные структуры, упускаемые из виду традиционными методами (H3). Кластеризация K-средних определила региональные типологии, при этом оптимальное количество кластеров было подтверждено методами «локтя» и «оценка силуэта». Прогностическая мощность и ключевые факторы были проанализированы с использованием ансамблевых методов, Random Forest и XGBoost, а производительность оценивалась по среднеквадратичной ошибке (MSE). Результаты подтверждают глубоко поляризованный ландшафт с четырьмя различными социально-экономическими кластерами: богатые ресурсами регионы, страдающие от неравенства и оттока; ведущие хабы, сталкивающиеся с рисками чрезмерной централизации; индустриально-аграрные регионы с устойчивой бедностью; и отстающие республики, попавшие в ловушку зависимости от субсидий. Анализ подтвердил H2, показав одновременный догоняющий рост в одних областях и укоренившуюся дивергенцию в других. Random Forest продемонстрировал превосходную точность прогнозирования (MSE = 0,132), подтвердив H3 и определив ключевые драйверы (H1), выходящие за рамки экономических показателей. Теоретическая значимость заключается в его новой, основанной на данных типологии регионов России, что подчеркивает превосходство методов ML-ансамбля для моделирования сложных пространственных неравенств. На практике полученные результаты предоставляют директивным органам основанную на фактических данных дорожную карту для дифференцированных региональных стратегий и прогностический инструмент для упреждающего вмешательства, имеющего жизненно важное значение для сбалансированного национального развития.

**Ключевые слова:** региональная поляризация; межрегиональная дифференциация; машинное обучение; кластерный анализ; прогнозирование временных рядов.

## Список использованных источников

1. Кузин В. Ю. Оценка пространственной поляризации Дальнего Востока России в постсоветский период // Вестник Северо-Восточного федерального университета им. М. К. Аммосова. Серия «Науки о Земле». 2023. № 2. С. 102–113. <https://doi.org/10.25587/svfu.2023.30.2.009>
2. Conover M., Ratkiewicz J., Francisco M., Goncalves B., Menczer F., Flammini A. Political polarization on Twitter // Proceedings of the International AAAI Conference on Web and Social Media. 2021. Vol. 5, No. 1. Pp. 89–96. <https://doi.org/10.1609/icwsm.v5i1.14126>
3. Da Silva T. P. Learning beyond the spatial autocorrelation structure: A machine learning-based approach to discovering new patterns and relationships in the context of spatially contextualized modeling of voting behavior. Doctoral Thesis. Sao Carlos, 2023. <https://doi.org/10.11606/t.55.2023.tde-15012024-174102>
4. Levy R. Social Media, News Consumption, and Polarization: Evidence from a Field Experiment // American Economic Review. 2021. Vol. 111, No. 3. Pp. 831–870. <https://doi.org/10.1257/aer.20191777>
5. Aharon-Gutman M., Schaap M., Lederman I. Social topography: Studying spatial inequality using a 3D regional model // Journal of Rural Studies. 2018. Vol. 62. Pp. 40–52. <https://doi.org/10.1016/j.jrurstud.2018.06.010>
6. Martin R., Sunley P. Regional economic resilience: evolution and evaluation // Handbook on Regional Economic Resilience. Edited by G. Bristow, A. Healy. Edward Elgar Publishing, 2020. Pp. 10–35. <https://doi.org/10.4337/9781785360862.00007>
7. Aharon-Gutman M., Burg D. How 3D visualization can help us understand spatial inequality: On social distance and crime // Environment and Planning B: Urban Analytics and City Science. 2019. Vol. 48, Issue 4. Pp. 793–809. <https://doi.org/10.1177/2399808319896524>
8. Stephens J. D. Thomas Piketty (2014), Capital in the Twenty-First Century. Translated by Arthur Goldhammer. Cambridge, Massachusetts: Belknap Press of Harvard University Press, 685 Pp. // Journal of Social Policy. 2015. Vol. 45, Issue 1. Pp. 172–173. <https://doi.org/10.1017/s0047279415000616>
9. Marco C., Lenka J., Christa K. S. The Future of EU Cohesion: Scenarios and Their Impacts on Regional Inequalities. Belgium: European Parliamentary Research Service (EPRS), 2024. 61 p. URL: [https://www.europarl.europa.eu/RegData/etudes/STUD/2024/762854/EPRS\\_STU\(2024\)762854\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2024/762854/EPRS_STU(2024)762854_EN.pdf)
10. Glaeser E. L. Urban resilience // Urban Studies. 2021. Vol. 59, Issue 1. Pp. 3–35. <https://doi.org/10.1177/00420980211052230>
11. Тебекин М. В. Методический подход к оценке пространственной поляризации регионов // Прикладные экономические исследования. 2023. № 4. С. 86–94. [https://doi.org/10.47576/2949-1908\\_2023\\_4\\_86](https://doi.org/10.47576/2949-1908_2023_4_86)
12. Алтунина В. В., Анучина Д. А. Оценка уровня пространственной поляризации российских регионов // Экономика, предпринимательство и право. 2023. Т. 13, № 5. С. 1319–1340. <https://doi.org/10.18334/epp.13.5.117516>
13. Taymaz E. Regional convergence or polarization: the case of the Russian Federation // Regional Research of Russia. 2022. Vol. 12. Pp. 469–482. <https://doi.org/10.1134/s2079970522700198>
14. Gluschenko K. P. Regional inequality in Russia: Anatomy of convergence // Regional Research of Russia. 2023. Vol. 13, Suppl. 1. Pp. S1–S12. <https://doi.org/10.1134/s207997052360004x>
15. Atkinson A. B. On the measurement of inequality // Journal of Economic Theory. 1970. Vol. 2, Issue 3. Pp. 244–263. [https://doi.org/10.1016/0022-0531\(70\)90039-6](https://doi.org/10.1016/0022-0531(70)90039-6)
16. Chi S., Grigsby-Toussaint D. S., Bradford N., Choi J. Can Geographically Weighted Regression improve our contextual understanding of obesity in the US? Findings from the USDA Food Atlas // Applied Geography. 2013. Vol. 44. Pp. 134–142. <https://doi.org/10.1016/j.ap-geog.2013.07.017>

17. *Amiri M., Pourghasemi H. R., Ghanbarian G. A., Afzali S. F.* Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms // *Geoderma*. 2019. Vol. 340. Pp. 55–69. <https://doi.org/10.1016/j.geoderma.2018.12.042>
18. *Bergal B.* Audit of the effectiveness of cluster policy implementation in the region // *Pskov Region-logical Journal*. 2022. Vol. 18, No. 2. Pp. 154–167. <https://doi.org/10.37490/S221979310019289-0>
19. *Rogot A.* Dimensionality reduction techniques in Macroeconomic Analysis // *CUNY Academic Works*. 2023. URL: [https://academicworks.cuny.edu/bb\\_etds/166](https://academicworks.cuny.edu/bb_etds/166)
20. *Chagovets L., Chahovets V., Chernova N.* Machine Learning Methods Applications for Estimating Unevenness Level of Regional Development // *Data-Centric Business and Applications. Evolutions in Business Information Processing and Management*. Vol. 3 / edited by D. Ageyev, T. Radivilova, N. Kryvinska. Springer Cham, 2020. Pp. 115–139. [https://doi.org/10.1007/978-3-030-35649-1\\_6](https://doi.org/10.1007/978-3-030-35649-1_6)
21. *Veale M., Binns R.* Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data // *Big Data & Society*. 2017. Vol. 4, Issue 2. 205395171774353. <https://doi.org/10.1177/2053951717743530>
22. *Terlizzi E. P., Cohen R. A.* Geographic Variation in Health Insurance Coverage: United States, 2022. National Health Statistics Reports. 2023. <https://doi.org/10.15620/cdc:133320>
23. *Oberlander J.* Polarization, partisanship, and health in the United States // *Journal of Health Politics Policy and Law*. 2024. Vol. 49, Issue 3. Pp. 329–350. <https://doi.org/10.1215/03616878-11075609>
24. *Giannini M., Martini B.* Regional disparities in the European Union. A machine learning approach // *Papers of the Regional Science Association*. 2024. Vol. 103, Issue 4. 100033. <https://doi.org/10.1016/j.pirs.2024.100033>
25. *Lange T.* Socio-economic and political responses to regional polarisation and socio-spatial peripheralisation in Central and Eastern Europe: a research agenda // *Hungarian Geographical Bulletin*. 2015. Vol. 64, No. 3. Pp. 171–185. <https://doi.org/10.15201/hungeobull.64.3.2>
26. *Druckman J. N., Levendusky M. S.* Correction to: What Do We Measure When We Measure Affective Polarization? // *Public Opinion Quarterly*. 2024. Vol. 88, Issue 3. Pp. 1095–1096. <https://doi.org/10.1093/poq/nfae051>
27. *Jo J.* Effectiveness of normalization Pre-Processing of big data to the machine learning performance // *The Journal of the Korea Institute of Electronic Communication Sciences*. 2019. Vol. 14, Issue 3. Pp. 547–552. <https://doi.org/10.13067/jkiecs.2019.14.3.547>
28. *Thorndike R. L.* Who belongs in the family? // *Psychometrika*. 1953. Vol. 18, Issue 4. Pp. 267–276. <https://doi.org/10.1007/bf02289263>
29. *Eltibi M. F., Ashour W. M.* Initializing KMeans Clustering Algorithm using Statistical Information // *International Journal of Computer Applications*. 2011. Vol. 29, No. 7. Pp. 51–55. <https://doi.org/10.5120/3573-4930>
30. *Currin C. B., Vera S. V., Khaledi-Nasab A.* Depolarization of echo chambers by random dynamical nudge // *Scientific Reports*. 2022. Vol. 12. 9234. <https://doi.org/10.1038/s41598-022-12494-w>
31. *Behrens T., Schmidt K., Rossel R. V., Gries P., Scholten T., MacMillan R. A.* Spatial modelling with Euclidean distance fields and machine learning // *European Journal of Soil Science*. 2018. Vol. 69, Issue 5. Pp. 757–770. <https://doi.org/10.1111/ejss.12687>
32. *Likas A., Vlassis N., Verbeek J. J.* The global k-means clustering algorithm // *Pattern Recognition*. 2002. Vol. 36, Issue 2. Pp. 451–461. [https://doi.org/10.1016/s0031-3203\(02\)00060-2](https://doi.org/10.1016/s0031-3203(02)00060-2)
33. *Breiman L.* Random forests // *Machine Learning*. 2001. Vol. 45. Pp. 5–32. <https://doi.org/10.1023/A:1010933404324>
34. *Friedman J. H.* Greedy function approximation: A gradient boosting machine // *The Annals of Statistics*. 2001. Vol. 29, No. 5. Pp. 1189–1232. <https://doi.org/10.1214/aos/1013203451>
35. *Friedman J., Hastie T., Tibshirani R.* Additive logistic regression: a statistical view of boosting // *The Annals of Statistics*. 2000. Vol. 28, No. 2. Pp. 337–407. <https://doi.org/10.1214/aos/1016218223>

36. Biau G., Scornet E. A random forest guided tour // *Test*. 2016. Vol. 25. Pp. 197–227. <https://doi.org/10.1007/s11749-016-0481-7>
37. Wright M. N., Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R // *Journal of Statistical Software*. 2017. Vol. 77, Issue 1. Pp. 1–17. <https://doi.org/10.18637/jss.v077.i01>
38. Maksubova D. M., Umargadzhieva N. M., Aripova P. G. Methodological approaches to assessing regional development // *Computational and Strategic Business Modelling* / edited by D. P. Sakas, D. K. Nasiopoulos, Yu. Taratuhina. Springer Cham, 2024. Pp. 543–555. [https://doi.org/10.1007/978-3-031-41371-1\\_45](https://doi.org/10.1007/978-3-031-41371-1_45)
39. Amarasinghe K., Rodolfa K., Lamba H., Ghani R. Explainable Machine learning for public Policy: use cases, gaps, and research directions. *Data & Policy* 2023. Vol. 5. e5. <https://doi.org/10.1017/dap.2023.2>
40. Kamal S., Gullic J., Bagavathi A. Modeling Polarization on Social Media posts: A Heuristic approach using media Bias // *Foundations of Intelligent Systems. Proceedings of 26th International Symposium, ISMIS 2022* / edited by M. Ceci, S. Flesca, E. Masciari, G. Manco, Z. W. Raś. Springer Cham, 2022. Pp. 35–43. [https://doi.org/10.1007/978-3-031-16564-1\\_4](https://doi.org/10.1007/978-3-031-16564-1_4)
41. Ahrend R. Can Russia Break the “Resource Curse”? // *Eurasian Geography and Economics*. 2005. Vol. 46, Issue 8. Pp. 584–609. <https://doi.org/10.2747/1538-7216.46.8.584>
42. Maximova S. G., Omelchenko D. A., Noyanzina O. E. Human development, satisfaction with human capital and security in the Siberian and Far Eastern border regions // *RUDN Journal of Sociology*. 2022. Vol. 22, No. 3. Pp. 646–660. <https://doi.org/10.22363/2313-2272-2022-22-3-646-660>
43. Sitkevich D. A. Economic and sociocultural factors of migration attitudes of residents of the North Caucasus // *Regional Research of Russia*. 2023. Vol. 13, Suppl 1. Pp. S78–S88. <https://doi.org/10.1134/s2079970523600166>
44. Берг Д. Б., Балунгу Д. М., Шеломенцев А. Г., Гончарова К. С. Экспериментальные траектории процессов конвергенции и дивергенции неравномерности доходов населения регионов России // *Journal of Applied Economic Research*. 2024. Т. 23, № 2. С. 364–393. <https://doi.org/10.15826/vestnik.2024.23.2.015>
45. Balungu D. M., Kumar A. Forecasting the economic growth of Sverdlovsk Region: A comparative analysis of machine learning, linear regression and autoregressive models // *Journal of Applied Economic Research*. 2024. Vol. 23, No. 3. Pp. 674–695. <https://doi.org/10.15826/vestnik.2024.23.3.027>
46. Ketova K., Kasatkina E., Vavilova D. Clustering Russian Federation Regions According to the Level of Socio-Economic Development with the Use of Machine Learning Methods // *Economic and Social Changes: Facts, Trends, Forecast*. Vol. 14, No. 6. Pp. 70–85. <https://doi.org/10.15838/esc.2021.6.78.4>

## ИНФОРМАЦИЯ ОБ АВТОРАХ

### Балунгу Даниель Мусафири

Аспирант, ассистент базовой кафедры аналитики больших данных и методов видеоанализа Института радиоэлектроники и информационных технологий Уральского федерального университета имени первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия (620002, г. Екатеринбург, ул. Мира, 19); ORCID <https://orcid.org/0009-0001-5098-7603> e-mail: [danielbal03.db@gmail.com](mailto:danielbal03.db@gmail.com)

### Розанова Анна Вячеславовна

Магистрант базовой кафедры аналитики больших данных и методов видеоанализа Института радиоэлектроники и информационных технологий Уральского федерального университета имени первого Президента России Б. Н. Ельцина г. Екатеринбург, Россия

(620002, г. Екатеринбург, ул. Мира, 19); ORCID <https://orcid.org/0009-0003-9803-0848> e-mail: [rozanna221132@icloud.com](mailto:rozanna221132@icloud.com)

### **Андреева Кристина Александровна**

Магистрант базовой кафедры аналитики больших данных и методов видеоанализа Института радиоэлектроники и информационных технологий Уральского федерального университета имени первого Президента России Б. Н. Ельцина г. Екатеринбург, Россия (620002, г. Екатеринбург, ул. Мира, 19); ORCID <https://orcid.org/0009-0009-5345-6160> e-mail: [kristinalezhnina88@gmail.com](mailto:kristinalezhnina88@gmail.com)

### **Солод Анастасия Васильевна**

Магистрант базовой кафедры аналитики больших данных и методов видеоанализа Института радиоэлектроники и информационных технологий Уральского федерального университета имени первого Президента России Б. Н. Ельцина г. Екатеринбург, Россия (620002, г. Екатеринбург, ул. Мира, 19); ORCID <https://orcid.org/0009-0007-8795-1640> e-mail: [nsolodv@mail.ru](mailto:nsolodv@mail.ru)

### **Чэнь Юйтун**

Магистрант базовой кафедры аналитики больших данных и методов видеоанализа Института радиоэлектроники и информационных технологий Уральского федерального университета имени первого Президента России Б. Н. Ельцина г. Екатеринбург, Россия (620002, г. Екатеринбург, ул. Мира, 19); ORCID <https://orcid.org/0009-0001-7684-1530> e-mail: [kirito200207@gmail.com](mailto:kirito200207@gmail.com)

### **ДЛЯ ЦИТИРОВАНИЯ**

Балунгу Д. М., Розанова А. В., Андреева К. А., Солод А. В., Чэнь Ю. Выявление закономерностей региональной поляризации в России: подход машинного обучения // *Journal of Applied Economic Research*. 2026. Т. 25, № 1. С. 135–162. <https://doi.org/10.15826/vestnik.2026.25.1.005>

### **ИНФОРМАЦИЯ О СТАТЬЕ**

Дата поступления 9 мая 2025 г.; дата поступления после рецензирования 6 октября 2025 г.; дата принятия к печати 5 ноября 2025 г.

