Научная статья
УДК 332.12
DOI: 10.17586/2713-1874-2026-2-18-30

РАЙОНИРОВАНИЕ НА ОСНОВЕ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ РОССИЙСКИХ РЕГИОНОВ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Екатерина Андреевна Орлова[1], Светлана Владимировна Полянская[2]✉

[1,2]Северо-Западный институт управления, Санкт-Петербург, Россия
[1]eorlova-22@ranepa.ru
[2]polyanskaya-sv@ranepa.ru✉, https://orcid.org/0000-0002-2707-9746
Язык статьи – русский

Аннотация: Целью исследования является проведение сравнительного анализа эффективности методов понижения размерности с помощью различных алгоритмов машинного обучения, (PCA, t-SNE, UMAP) для выявления кластеров регионов России на основе многомерных социально-экономических показателей и обоснование практических рекомендаций для управленческого анализа. Исследование основано на гипотезе о том, что для выявления сходства необходимо выбрать не только метод, который способен выстроить нелинейные связи между показателями, но и метод нормализации данных. Для анализа были рассмотрены данные по 38 показателям из 6 ключевых сфер развития по 85 субъектам РФ за 2017–2023 гг. Применялись различные методы предобработки и понижения размерности. Для приведения распределения данных к нормальному использовался метод Бокса-Кокса. Качество методов оценивалось по нескольким метрикам качества. Установлено, что выбор метрики сходства кардинально меняет интерпретацию результатов. В частности, при использовании косинусной метрики (сравнение структурных профилей) метод UMAP с Z-стандартизацией данных показал максимальное качество кластеризации. Для задач мониторинга и сравнения абсолютных уровней развития рекомендован метод t-SNE с евклидовой метрикой и робастной нормировкой. Для стратегического анализа и выявления типовых моделей развития с целью разработки кластерно-ориентированной политики оптимальным является метод UMAP с косинусной метрикой и Z-стандартизацией данных. Таким образом, вместо поиска универсального алгоритма предлагается переход к дифференцированному управлению на основе эффективного кластерного анализа и выявленных структурных типов регионов.

Ключевые слова: кластеризация регионов**,** косинусное расстояние, машинное обучение, метрический анализ, понижение размерности, предобработка данных, региональное управление, социально-экономическое развитие, t-SNE, UMAP

Ссылка для цитирования: Орлова Е. А., Полянская С. В. Районирование на основе статистических показателей социально-экономического развития российских регионов: сравнительный анализ методов кластеризации // Экономика. Право. Инновации. – 2026. – Т. 14.– № 2.– С. 18–30. – DOI: 10.17586/2713-1874-2026-2-18-30.

COMPARATIVE ANALYSIS OF DIFFERENT APPROACHES TO CLUSTERING MULTIDIMENSIONAL DATA USING THE EXAMPLE OF RUSSIAN REGIONS

Ekaterina A. Orlova[1], Svetlana V. Polyanskaya[2]✉

[1,2]North-Western Institute of Management, Saint Petersburg, Russia
[1]eorlova-22@ranepa.ru
[2]polyanskaya-sv@ranepa.ru✉, https://orcid.org/0000-0002-2707-9746
Article in Russian

Abstract: The purpose of the study is to conduct a comparative analysis of the effectiveness of dimensionality reduction methods using various machine learning algorithms (PCA, t-SNE, UMAP) to identify clusters of Russian regions based on multidimensional socioeconomic indicators and to provide practical recommendations for management analysis. The study is based on the hypothesis that to identify similarities, it is necessary not only to use methods that can select nonlinear relationships between indicators, but also to select a data normalization method. The analysis included data on 38 indicators from 6 key areas of development for 85 Russian regions from 2017 to 2023. Various preprocessing and dimensionality reduction methods were used. The Box-Cox method was used to normalize data distribution. The quality of the methods was evaluated using several quality metrics. It was found that the choice of similarity metric

significantly affects the interpretation of the results. In particular, when using the cosine metric (comparing structural profiles), the UMAP with Z-standardization showed the highest clustering quality. For monitoring tasks and comparing absolute levels of development, the t-SNE method with Euclidean metric and robust normalization is recommended. For strategic analysis and identifying typical development models in order to develop a cluster-oriented policy, the UMAP method with cosine metric and Z-standardization is optimal. Thus, instead of searching for a universal algorithm, it is proposed to switch to differentiated management based on effective cluster analysis and identified structural types of regions.

Keywords: regional clustering, cosine distance, machine learning, metric analysis, dimensionality reduction, data preprocessing, regional management, socio-economic development, t-SNE, UMAP

For citation: Orlova E. A., Polyanskaya S. V. Comparative Analysis of Different Approaches to Clustering Multi-dimensional Data Using the Example of Russian Regions. *Ekonomika. Pravo. Innovacii*. 2026. Vol. 14. No. 2. pp. 18–30. (In Russ.). DOI: 10.17586/2713-1874-2026-2-18-30.

Введение. Активное внедрение цифровых технологий и накопление больших данных в государственном управлении создает новые возможности для принятия обоснованных решений на региональном уровне [1]. Ключевой задачей при этом становится анализ многомерных социально-экономических показателей, которые характеризуют комплексное развитие субъектов Российской Федерации [2]. Классические подходы к анализу, основанные на рассмотрении отдельных индикаторов или составлении интегральных рейтингов, зачастую не позволяют выявить скрытые структурные взаимосвязи и устойчивые типы региональных профилей. Это ограничивает возможности для разработки дифференцированной и адресной политики, учитывающей специфику групп регионов со схожими характеристиками.

Проблема усугубляется вычислительными сложностями, возникающими при работе с десятками взаимосвязанных показателей. Это приводит к необходимости применения методов понижения размерности и визуализации, которые способны преобразовать исходные высокоразмерные данные в форму, пригодную для содержательной интерпретации. В последние годы наряду с классическим линейным методом главных компонент (PCA) для уменьшения размерности данных стали использоваться алгоритмы машинного обучения, использующие нелинейные методы, такие как t-SNE и UMAP, каждый из которых основан на различных принципах и имеет свои преимущества. T-SNE (t-distributed Stochastic Neighbor Embedding) – метод стохастического вложения соседей с t-распределением. Этот метод машинного обучения предполагает вычисление для каждой пары точек в многомерном пространстве вероятности их сходства, а затем строит распределение вероятностей так, чтобы объекты, которые близки в исходном пространстве, имели высокие вероятности быть близкими и в низкоразмерном пространстве. UMAP (Uniform Manifold Approximation and Projection) также является методом машинного обучения для сжатия исходной информации, он использует четкий математический аппарат и имеет более высокую скорость выполнения по сравнению с t-SNE. Однако в экономических исследованиях эти методы пока не получили широкого применения.

Постановка задачи исследования. Целью данного исследования является сравнительный анализ эффективности методов понижения размерности (PCA, t-SNE, UMAP) для выявления содержательных кластеров российских регионов на основе многомерных социально-экономических показателей и получение практических рекомендаций для управленческого анализа.

Гипотеза исследования состоит в том, что эффективность методов кардинально зависит не только от выбора алгоритма, но и от осознанного подбора метрики сходства и методов предобработки данных, что в итоге определяет тип выявляемых закономерностей.

Новизна исследования заключается в адаптации передовых методов машинного обучения для обработки данных для решения конкретной управленческой задачи – выявления типовых профилей социально-экономического развития регионов. Практическая значимость исследования определяется тем, что его выводы позволяют перейти от мониторинга разрозненных показателей к

выделению устойчивых типов регионов, что является основой для разработки кластерно-ориентированной стратегии социально-экономического развития.

Для достижения поставленной цели был проведен анализ данных по 38 показателям, официально публикуемых Росстатом, сгруппированных в шесть ключевых сфер за период с 2017 по 2023 год: демография и трудовые ресурсы, здравоохранение, экономика и инновации, уровень жизни и доходы, образование и культура, услуги и транспорт. Такой комплексный набор показателей, охватывающий различные аспекты развития, позволяет проводить многомерную оценку, согласующуюся с подходами, описанными в работах по региональной кластеризации [3].

Литературный обзор. В научной литературе, в том числе в работах, посвященных кластеризации российских регионов [4–6], остается дискуссионным вопрос о выборе оптимального метода, метрики расстояния и предобработки данных для решения конкретных задач в сфере управления. В качестве методов для сжатия данных рассматриваются метод главных компонент (PCA) [7–9], метод t-SNE [10–12] и метод UMAP [13–15]. Данные методы широко используются в разных областях, но это применение имеет четкие ограничения, которые часто не отражаются в научных исследованиях.

Второй важный вопрос касается выбора показателей, характеризующих социально-экономическое состояние региона. В современных исследованиях рассматриваются разные методологии. Например, в работе [16] для сопоставления социально-экономического положения районов Крайнего Севера с другими территориями РФ используются показатели демографии, занятости, доходов и инфраструктуры. В работе [17] отмечается, что показатели демографии и трудовых ресурсов являются базовыми для оценки человеческого потенциала. В работе [18] рассматривается блок здравоохранения, который включает как ресурсные показатели (численность врачей, мощность медучреждений), так и результативные (ожидаемая продолжительность жизни, заболеваемость). В работе [19] выделяется группа экономических и инновационных показателей (ВРП, инвестиции, инновационная активность), которые позволяют оценить экономический потенциал и способность регионов к технологическому развитию. Показатели уровня жизни и доходов, отражающие социальное благополучие населения, выделены в [1]. Блок образования и культуры, который позволяет оценить уровень человеческого капитала и доступность культурных благ, рассмотрен в [5]. Группа услуг и транспорта как характеристика развития инфраструктуры и доступности услуг для населения, являющаяся важным аспектом регионального развития, рассмотрена в [20].

Методы и материалы исследования. Эмпирическую базу исследования составили официальные статистические данные по социально-экономическому развитию субъектов РФ за период 2017–2023 годы. Источником данных послужила официальная статистическая информация, публикуемая Федеральной службой государственной статистики. Единицей анализа выступили 85 субъектов РФ. В исследовании использовалось 38 показателей по всем субъектам РФ за период с 2017 по 2023 годы, сгруппированных по шести тематическим блокам: демография и трудовые ресурсы, здравоохранение, экономика и инновации, уровень жизни и доходы, образование и культура, услуги и транспорт. Такой подход позволяет комплексно оценить развитие регионов, учитывая как экономические, так и социальные аспекты (см. таблицу 1).

Для упрощения интерпретации и сглаживания годовых колебаний данные за семь лет были заменены средними значениями по каждому региону. Далее для обеспечения качества анализа и соответствия предположениям статистических методов был выполнен комплекс процедур предобработки [21, 22]. Обработка пропущенных значений осуществлялась путем заполнения медианным значением по соответствующему показателю для всех регионов. Для снижения скошенности распределений и сближения их с нормальным к исходным данным было применено преобразование Бокса-Кокса [23].

Таблица 1

Группы социально-экономических показателей

Источник: составлена авторами

Группа показателей	Показатели
Демография и трудовые ресурсы (7 показателей)	x_1 – численность населения; x_2 – коэффициент естественного прироста населения (на 1000 чел.); x_3 – коэффициент миграционного прироста (на 10000 чел.); x_4 – ожидаемая продолжительность жизни при рождении (в годах); x_5 – уровень занятости населения (в %); x_6 – уровень безработицы населения в возрасте 15 лет и старше (в %); x_7 – численность занятых, приходящихся на одного пенсионера (чел.)
Экономика и инновации (7 показателей)	x_8 – валовой региональный продукт; x_9 – инвестиции в основной капитал на душу населения; x_{10} – оборот розничной торговли в расчёте на душу населения; x_{11} – уровень инновационной активности организаций (в %); удельный вес организаций, x_{12} – осуществлявших технологические инновации (в %); x_{13} – доля продукции высокотехнологичных и наукоёмких отраслей в валовом региональном продукте (в %); x_{14} – доля внутренних затрат на исследования и разработки в ВРП (в %)
Здравоохранение (5 показателей)	x_{15} – мощность амбулаторно-поликлинических организаций (на 10 тыс. чел. в смену); x_{16} – численность врачей на 100 тыс. чел.; x_{17} – численность среднего медицинского персонала на 100тыс. чел.; x_{18} – численность населения на одну больничную койку (чел.). x_{19} – заболеваемость (на 1000 чел.)
Уровень жизни и доходов (7 показателей)	x_{20} – фактическое конечное потребление домашних хозяйств на душу населения; x_{21} – соотношение среднедушевых денежных доходов населения с величиной прожиточного минимума; x_{22} – потребительские расходы в среднем на душу населения; x_{23} – общая площадь жилых помещений, приходящаяся в среднем на одного жителя (кв. м); x_{24} – удельный вес расходов на оплату ЖКХ (в % от потребительских расходов); x_{25} – уровень бедности; x_{26} – число зарегистрированных преступлений на 100 тыс. чел.
Культура и образование (6 показателей)	x_{27} – численность учителей на 1000 человек населения; x_{28} – численность студентов СПО на 10 тыс. чел.; x_{29} – численность студентов высшего образования на 10 тыс. чел.; численность x_{30} – зрителей театров на 1 тыс. чел.; x_{31} – число посещений музеев на 1 тыс. чел.; x_{32} – библиотечный фонд на 1 тыс. чел.
Услуги и транспорт (6 показателей)	x_{33} – объём платных услуг на душу населения; x_{34} – объём бытовых услуг на душу населения; x_{35} – объём транспортных услуг на душу населения; x_{36} – объём телекоммуникационных услуг на душу населения; x_{37} – объём коммунальных услуг на душу населения; x_{38} – число автобусов общего пользования на 100 тыс. чел.

Для устранения влияния различий в масштабах и единицах измерения тестировались четыре метода нормировки: Z-score (Z-стандартизация) $x_i' = \frac{x_i - \mu_x}{\sigma_x}$ Min-Max нормализация (приведение к диапазону [0, 1]) $x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$ робастная стандартизация (центрирование по медиане и масштабирование по медианному абсолютному отклонению) $x_i' = \frac{x_i - x_{Me}}{Me|x_i - x_{Me}|}$ и нормировка на основе медианной скошенности (MedCouple), являющаяся усовершенствованным робастным методом, учитывающим асимметрию распределения и вместо медианы исходных данных использующая MC (MedCouple) – медиану функции точек x_i и x_j, находящихся по разные стороны от медианы [24, 25]

$$h(x_i, x_j) = \frac{(x_j - x_{Me}) - (x_{Me} - x_i)}{x_j - x_i} \quad .$$

В работе использовались технологии ИИ для обработки информации – методы уменьшения размерности данных (UMAP и TSNE) и иерархическая кластеризации для разделения регионов на группы. Для перехода от многомерного пространства к низкоразмерному сравнивались три алгоритма. Метод главных компонент (PCA) применялся для анализа линейной структуры данных и оценивался по количеству главных компонент, объясняющих не менее 85% дисперсии. Метод t-SNE использовался с автоматически устанавливаемым параметром perplexity в зависимости от размера выборки. Метод UMAP применялся с адаптивно устанавливаемым параметром n_neighbors. Для оценки влияния смысловой интерпретации расстояния между регионами был реализован двухфакторный подход, включающий вариант с евклидовым расстоянием и вариант с косинусным расстоянием.

После уменьшения размерности применялась иерархическая кластеризация с методом связи Уорда, число кластеров определялось, исходя из размера выборки. Комплексная оценка эффективности проводилась с использованием пяти метрик качества:

– коэффициент надежности (Trustworthiness), оценивающий сохранение локальной структуры;

– силуэтный коэффициент (Silhouette Score), отвечающий за четкость и разделенность кластеров;

– коэффициент корреляции Спирмена, показывающий степень сохранения глобальной структуры;

– индекс Калински-Харабаза, показывающий степень разброса данных внутри кластеров и между кластерами;

– индекс Дэвиса-Болдина, отвечающий за компактность кластеров.

Все этапы анализа были реализованы на языке R в среде RStudio.

Результаты. Применение PCA к шести тематическим группам показателей позволило количественно оценить внутреннюю организацию социально-экономической статистики регионов. Группы «Демография и трудовые ресурсы», «Уровень жизни и доходы» и «Экономика и инновации» продемонстрировали высокую степень линейной связности: первые две главные компоненты объясняли свыше 94% дисперсии в данных. Это свидетельствует о жесткой взаимосвязи ключевых показателей внутри этих сфер. Напротив, группа «Образование и культура» оказалась наиболее многомерной и структурно сложной. Для объяснения 87% дисперсии потребовалось четыре главные компоненты, что указывает на относительную независимость динамики показателей образования и культурной активности. По первым двум компонентам для этой группы было получено всего 48,65% объясненной дисперсии, что является крайне неудовлетворительным результатом. Группа «Услуги и транспорт» эффективно сжималась в одну главную компоненту, объясняющую более 81% дисперсии, что говорит о существовании сильного общего фактора, синхронизирующего развитие различных видов услуг в регионах. Для достижения не менее 85% объясненной дисперсии PCA позволил сократить исходное пространство из 38 показателей до тринадцати обобщенных компонент (см. таблицу 2).

Таблица 2

Результаты применения метода главных компонент ко всем группам социально-экономических показателей

Источник: составлена авторами

Компонента	1	2	3	4	5	6	7	8	9	10	11	12	13
Сумма квадр. нагрузок	9,09	4,26	2,89	2,62	3	2,45	1,86	1,69	1,61	1,43	1,34	1,28	1,25
% от дисперсии	23,3	10,93	7,4	6,72	7	6,28	4,78	4,33	4,13	3,65	3,44	3,27	3,21
Накопленный %	23,3	34,2	41,6	48,4	55	61,3	66	70,4	74,5	78,1	81,6	84,9	88,1

Результаты говорят о серьезных проблемах использования данного метода в задачах кластеризации регионов с большим количеством разных показателей. PCA хорошо подходит для небольшого количества экономических показателей, которые легко нормализуются, не содержат экстремальные значения и имеют линейные взаимосвязи. Однако, большое количество необходимых для полного анализа социально-экономических показателей плохо поддается нормализации; кроме того, между показателями, в основном, возникают нелинейные зависимости, с которыми PCA не справляется. а из-за сильной региональной дифференциации возникает высокий коэффициент вариации, который также ведет к искажению результатов. Все эти причины делают применение PCA для задач кластеризации регионов с использованием большого количества показателей малоэффективным.

В качестве альтернативы методу PCA были рассмотрены два метода, которые предпочтительны для данных с нелинейными зависимостями: t-SNE и UMAP. Данные методы сравнивались по нескольким характеристикам качества для двух разных метрик – евклидовой и косинусной. При использовании евклидовой метрики метод t-SNE показал высокую эффективность в сохранении локальной структуры (Trustworthiness в диапазоне 0,843–0,886). Однако низкие значения коэффициента корреляции Спирмена говорят о сильных искажениях расстояний между дальними соседями (изменение глобальной структуры). Поэтому в случае выбора евклидовой метрики метод UMAP можно признать оптимальным для кластеризации при Z-стандартизации данных – почти все метрики говорят и о сохранении локальной, и глобальной структуры данных (см. таблицу 3). Коэффициент корреляции 0,819 говорит о том, что 81,9% вариации глобальных расстояний сохранено. Низкие значения силуэтного коэффициента для UMAP (0,375–0,454) говорят о недостаточной однородности внутри кластеров. Также это может свидетельствовать о наличии размытых границ, когда некоторые регионы попадают на границу кластеров. Метод t-SNE показал лучшее сохранение глобальной структуры при использовании робастной стандартизации (коэффициент корреляции существенно выше, чем у UMAP)

Таблица 3

Сравнение метрик качества для UMAP и t-SNE при использовании евклидова расстояния.

Источник: составлена авторами

Метрика	Trustworthiness		Silhouette Score		Коэффициент корреляции Спирмена		Индекс Калински-Харабаза		Индекс Дэвиса-Болдина	
Метод	UMAP	T-SNE	UMAP	T-SNE	UMAP	T-SNE	UMAP	T-SNE	UMAP	T-SNE
Z score	0,862	0,898	0,374	0,401	0,819	<0,5	166	112,4	0,937	1,001
Минимакс-ная	0,8	0,872	0,443	0,464	<0,5	<0,5	180,1	96,3	1,033	0,937

Метрика	Trustworthiness		Silhouette Score		Коэффициент корреляции Спирмена		Индекс Калински-Харабаза		Индекс Дэвиса-Болдина	
Робастная	0,893	0,886	0,424	0,447	<0,5	0,523	111,7	90,4	1,3	0,995
MedCouple	0,844	0,829	0,454	0,388	<0,5	<0,5	118,3	140,8	1,05	0,87

При использовании косинусного расстояния лучший результат также был достигнут при использовании UMAP. В частности, при Z-score стандартизации UMAP показал высокую положительную корреляцию, что свидетельствует о практически полном сохранении глобальной структуры.

При переходе от евклидовой к косинусной метрике значения силуэтного коэффициента резко возросли, достигнув максимума 0,8143 при использовании UMAP с Z-score нормировкой (см. таблицу 4). Значения коэффициента силуэта Sihoutte Score выше 0,7 свидетельствует о высоком качестве кластеризации и говорят о наличии в данных четких, хорошо разделенных групп регионов, объединенных именно сходством структур их социально-экономических показателей.

Таблица 4

Сравнение метрик качества для UMAP и t-SNE при использовании косинусного расстояния

Источник: составлена авторами

Метрика	Trustworthiness		Silhouette Score		Коэффициент корреляции Спирмена		Индекс Калински-Харабаза		Индекс Дэвиса-Болдина	
Метод	UMAP	T-SNE	UMAP	T-SNE	UMAP	T-SNE	UMAP	T-SNE	UMAP	T-SNE
Z score	0,779	0,768	0,814	0,708	0,901	0,7537	116,5	60,3	0,938	1,455
Минимакс-ная	0,839	0,801	0,668	0,776	0,8883	0,515	73,2	63,4	1,238	1,606
Робастная	0,874	0,825	0,751	0,825	<0,5	<0,5	99,4	71,5	1,029	1,076
MedCouple	0,871	0,814	0,828	0,783	<0,5	0,5932	182,3	48,1	0,856	1,356

Анализ состава кластеров, включающих, в частности, Санкт-Петербург, выявил существенные различия в результатах в зависимости от выбранной метрики сходства и метода нормировки. При использовании евклидовой метрики, направленной на сравнение абсолютных уровней развития, Санкт-Петербург чаще всего оказывался в одном кластере с Москвой и группой регионов с высокими экономическими показателями (см. таблицу 5).

Таблица 5

Сравнение кластеров регионов, содержащих Санкт-Петербург, полученных при разных способах стандартизации исходных данных

Источник: составлена авторами

UMAP (Z-score)	UMAP (минимаксная)	t-SNE (Z-score)	t-SNE (минимаксная)
г. Москва	Белгородская обл.	Белгородская обл.	г.Москва
г. Санкт-Петербург	Владимирская обл.	Воронежская обл.	г. Санкт-Петербург
Краснодарский край	Воронежская обл.	г. Москва	Калининградская обл.
Красноярский край	г. Москва	г. Санкт-Петербург	Краснодарский край
Ненецкий авт. округ	г. Санкт-Петербург	Липецкая обл.	Ленинградская обл.

UMAP (Z-score)	UMAP (минимаксная)	t-SNE (Z-score)	t-SNE (минимаксная)
Нижегородская обл.	Калужская обл.	Нижегородская обл.	Московская обл.
Новосибирская обл.	Липецкая обл.	Р. Адыгея	Р. Татарстан
Омская обл.	Нижегородская обл.	Р. Татарстан	Ханты-Мансийский авт. округ-Югра
Пермский край	Новгородская обл.	Ростовская обл.	Ямало-Ненецкий авт. округ
Приморский край	Пензенская обл.	Свердловская обл.	
Р. Башкортостан	Р. Татарстан	Тюменская обл. (без авт.округов)	
Р. Саха (Якутия)	Тульская обл.		
Р.Татарстан	Ульяновская обл.		
Ростовская обл.	Ярославская обл.		
Самарская обл.			
Свердловская обл.			
Томская обл.			
Тюменская обл. (без авт. округов)			
Хабаровский край			
Ханты-Мансийский авт. округ-Югра			
Челябинская обл.			
Ямало-Ненецкий авт. округ			

При t-SNE с нормировкой Z-score в кластер вошли только одиннадцать регионов, включая ключевые экономические центры и ресурсодобывающие субъекты. Похожая картина была получена и при нормировке MedCouple (рисунок 1). Использование косинусной метрики, измеряющей структурное сходство профилей регионов, принципиально изменило круг аналогий. Метод T-SNE показал сохранение глобальной структуры при использовании нормировки MedCouple, в отличие от UMAP. Однако наиболее показательным является результат, полученный при использовании UMAP с косинусным расстоянием и Z-score нормировкой (рисунок 2).

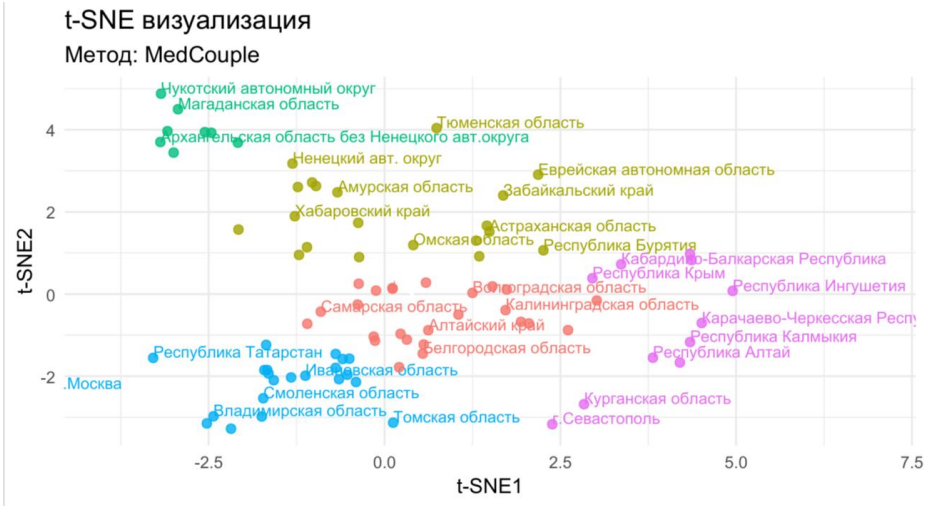

Рисунок 1 – Кластеризация после применения t-SNE с евклидовой метрикой и робастной нормировкой

Источник: составлено авторами по данным [26]

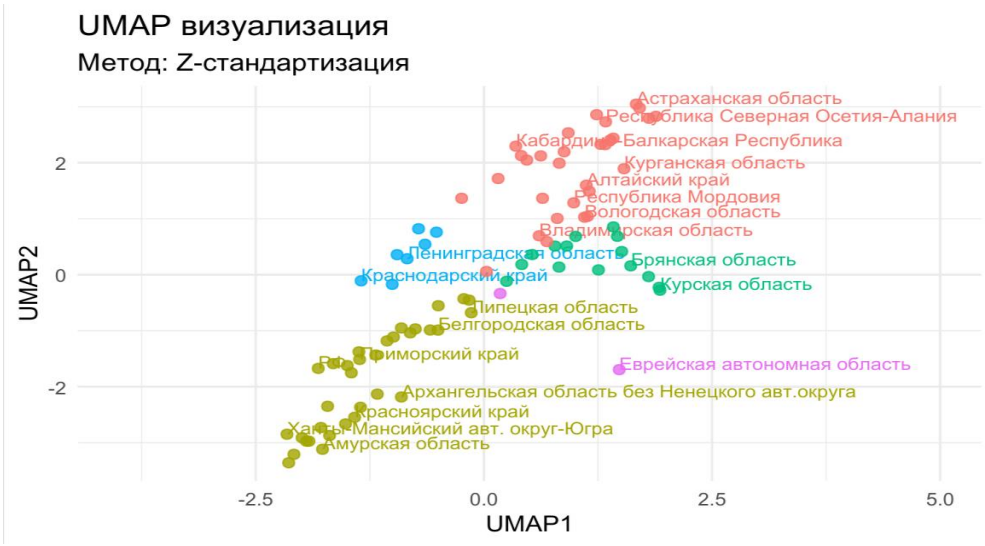

Рисунок 2 – Кластеризация после применения UMAP
с косинусной метрикой и Z-score нормировкой
Источник: составлено авторами по данным [26]

Санкт-Петербург был отнесен к обширному кластеру из 28 регионов. Помимо Москвы и традиционно развитых субъектов (Московская, Ленинградская, Свердловская области, Краснодарский край), в него вошли практически все приграничные и окраинные регионы страны: от Калининградской области и Республики Карелия на западе до Камчатского, Приморского, Хабаровского краев и Чукотки на востоке, а также северные территории (Ямало-Ненецкий, Ненецкий округа, Мурманская область, Республика Саха). Это указывает на то, что при анализе структурного сходства (пропорций и взаимосвязей показателей, а не их абсолютных величин) Санкт-Петербург имеет показатели, характерные для регионов с высокой значимостью внешнеэкономической и транзитной функции, особым геополитическим положением. То есть, будучи лидером по абсолютным показателям, Санкт-Петербург может разделять со специфической группой «окраинных» регионов общие структурные особенности развития, что требует учета при формировании направленной региональной политики, например, в сфере транспортной логистики. Таким образом, приоритетным для выявления таких содержательных структурных аналогий является использование именно косинусной метрики в связке с UMAP.

Выводы и рекомендации. Проведенное исследование подтвердило гипотезу о том, что эффективность методов анализа данных определяется не столько выбором конкретного алгоритма (t-SNE или UMAP), сколько подбором связки «метрика сходства – метод предобработки». Этот вывод является ключевым для преодоления подхода, когда выбор инструментария осуществляется по принципу универсальности без учета смысловой интерпретации расстояния между объектами анализа. Метод главных компонент, как метод, основанный на создании линейных комбинаций исходных данных, показал низкую эффективность при большом количестве разнородных показателей.

Значимым результатом исследования стало доказательство кардинального изменения результатов при переходе от евклидовой к косинусной метрике. При использовании евклидова расстояния выявляются группы, сформированные по принципу близости абсолютных уровней социально-экономического развития. В этом случае Санкт-Петербург закономерно оказывается в одном узком кластере с Москвой и ведущими регионами, такими как Московская область, Татарстан и нефтегазовые автономные округа. Однако такой подход, будучи полезным для ранжирования и оценки неравенства, маскирует более глубокие структурные сходства и различия.

Переход на косинусную метрику позволил выявить принципиально иную картину.

Обнаружение Санкт-Петербурга в одном крупном кластере с такими разнородными регионами, как приграничные и окраинные субъекты (Калининградская область, Камчатский край, Дальневосточные регионы), северные территории (Ямало-Ненецкий округ, Республика Саха) и другие, указывает на наличие общих структурных особенностей развития. Эти особенности могут проявляться в схожей значимости внешнеэкономической функции, особом геополитическом положении, в специфических демографических и инфраструктурных вызовах, связанных с удаленностью или особым статусом. Для целей управления это означает, что эффективные политические решения, апробированные для одного региона, после соответствующей адаптации масштаба и ресурсов, могут быть использованы для регионов со схожей структурой социально-экономического профиля, даже при сильном различии в абсолютных показателях.

Еще одним результатом стало опровержение превосходства UMAP над t-SNE в сохранении глобальной структуры. Анализ показал, что это справедливо лишь при определенных условиях. При использовании евклидовой метрики t-SNE продемонстрировал более предсказуемый результат в случае использования робастной нормировки, в то время как UMAP проявлял склонность к сильному искажению глобальных расстояний) для усиления визуального кластерного разделения. Неожиданно высокой оказалась эффективность робастных методов нормировки (RobustScaler, MedCouple), которые стабильно приводили к лучшим результатам по сравнению со стандартной Z-стандартизацией, особенно при работе с евклидовой метрикой. Это объясняется природой социально-экономических данных, часто содержащих выбросы и имеющих асимметричные распределения, к чему робастные методы менее чувствительны.

На основе сравнения методов понижения размерности (PCA, t-SNE, UMAP) с применением различных метрик сходства и методов предобработки был выработан практический инструмент для выявления содержательных кластеров регионов. Ключевым результатом работы является подтверждение гипотезы о том, что выбор метрики сходства является определяющим для интерпретации результатов кластеризации. Таким образом, исследование обосновывает следующий подход к анализу для целей управления:

1. Для оперативного мониторинга и оценки абсолютных уровней развития регионов рекомендуется использовать связку t-SNE (евклидово расстояние, Robust нормировка).

2. Для стратегического анализа, направленного на выявление типовых моделей развития и разработку адресной политики, оптимальным инструментом является применение метода UMAP с косинусным расстоянием и Z-score нормировкой.

Практическое применение полученных результатов позволяет перейти от мониторинга разрозненных индикаторов к выделению устойчивых типов регионов. Это создает основу для разработки кластерно-ориентированной политики, когда управленческие решения, доказавшие эффективность в одном регионе, могут быть осмысленно адаптированы для других субъектов, входящих в тот же структурный кластер, даже при значительном различии в абсолютных показателях. Для регионов это означает возможность обмена опытом и выработки совместных решений не только с традиционными регионами-лидерами, но и с географически и экономически удаленными субъектами, сталкивающимися со схожими структурными вызовами.

Декларация о применении ИИ. При подготовке текста статьи использовались технологии искусственного интеллекта (ИИ) в качестве инструментария для уменьшения размерности исходных данных и дальнейшей кластеризации. Все результаты исследования, выводы, данные, таблицы, графики и ссылки проверены и подтверждены авторами. Применение ИИ не повлияло на достоверность и научную ценность полученных результатов. Авторы полностью несут ответственность за содержание статьи и соблюдение принципов научной этики.

Список источников

1. Москвитина М. А. Теоретические подходы к классификации факторов развития региона и способам их анализа // Наука Красноярья. – 2020. – Т. 9. – № 4. – С. 296–312. – DOI: 10.12731/2070-7568-2020-4-296-312. – Текст: электронный.
2. Яковлев В. Б. Снижение размерности данных в региональной статистике российского образования // Вестник МГПУ. – 2016. – № 4 (38). – С. 61–69.
3. Логачева Н. М., Петрова А. К. Применение методов кластеризации в экономическом анализе регионов // Инновации. – 2021.– № 5(271). – С. 43–51.– DOI: 10.26310/2071-3010.2021.271.5.005. – Текст: электронный.
4. Блануца В. И. Пространственная алгоритмическая предвзятость в социально-экономической кластеризации российских регионов // Пространственная экономика. – 2024. – Т. 20. – № 2. – С. 71–92. – DOI: 10.14530/se.2024.2.071-092. – Текст: электронный.
5. Долгодворова Е. В. Кластерный анализ: базовые концепции и алгоритмы // Вопросы науки и образования. – 2018. – № 7 (19). – С. 73–76.
6. Кетова К. В., Касаткина Е. В., Вавилова Д. Д. Кластеризация регионов Российской Федерации по уровню социально-экономического развития с использованием методов машинного обучения // Экономические и социальные перемены: факты, тенденции, прогноз. – 2021. – Т. 14. – № 6. – С. 70–85. – DOI: 10.15838/esc.2021.6.78-4. – Текст: электронный.
7. Чавелипарамбил Д. Я. Сравнительное исследование методов PCA и Dynamic PCA на основе анализа случайных наборов данных // Неделя науки-2024: Сборник тезисов XIV научно-технической конференции студентов, аспирантов и молодых ученых имени А. С. Дудырева. – Санкт-Петербург: 2024. – 355 с.
8. Кенни Т., Гу Х., Хуан Т. Метод главных компонент Пуассона с применением к данным микробиома // Biometrics. – 2021. – Т. 77. – № 4. – С. 1369–1384. – DOI: 10.1111/biom.13384. – Текст: электронный. (In Eng.).
9. Марукатат С. Учебное пособие по PCA, приближенному PCA и приближенному ядерному PCA // Artificial Intelligence Review. – 2023. – Т. 56. – № 6. – С. 5445–5477. – DOI: 10.1007/s10462-022-10297-z. – Текст: электронный. (In Eng.).
10. Бо Канг, Д. Гарсия, Дж. Лиффийт и др. Условный t-SNE: более информативные t-SNE-встраивания. // Machine Learning. – 2021. – Т. 110. – № 10. – С. 2905–2940. – DOI: 10.1007/s10994-020-05917-0. – Текст: электронный. (In Eng.).
11. Озгоде Йигин Б., Сайгили Г. Оценка достоверности встраиваний t-SNE с использо- ванием случайного леса // International Journal of Machine

References

1. Moskvitina M. A. Theoretical Approaches to Classification of Regional Development Factors and Methods of their Analysis. *Nauka Krasnoyarya*. 2020. Vol. 9. No. 4. pp. 296–312. DOI: 10.12731/2070-7568-2020-4-296-312. (In Russ.).
2. Yakovlev V. B. Reducing the Dimension of Data in Regional Statistics of Russian Education. *Vestnik MGPU*. 2016. No. 4 (38). pp. 61–69. (In Russ.).
3. Logacheva N. M., Petrova A. K. Application of Clustering Methods in the Economic Analysis of Regions. *Innovatsii*. 2021. No. 5 (271). pp. 43–51. DOI: 10.26310/2071-3010.2021.271.5.005. (In Russ.).
4. Blanutsa V. I. Spatial Algorithmic Bias in Socio-Economic Clustering of Russian Regions. *Prostranstvennaya ekonomika*. 2024. Vol. 20. No. 2. pp. 71–92. DOI: 10.14530/se.2024.2.071-092. (In Russ.).
5. Dolgodvorova E. V. Cluster Analysis: Basic Concepts and Algorithms. *Voprosy nauki i obrazovaniya*. 2018. No. 7 (19). pp. 73–76. (In Russ.).
6. Ketova K. V., Kasatkina E. V., Vavilova D. D. Clusterization of Regions of the Russian Federation by the Level of Socio-Economic Development Using Machine Learning Methods. *Ekonomicheskiye i sotsial'nyye peremeny: fakty, tendentsii, prognoz*. 2021. Vol. 14. No. 6. pp. 70–85. DOI: 10.15838/esc.2021.6.78-4. (In Russ.).
7. Comparativestudy of PCA and Dynamic PCA methods based on the analysis of random data sets / D. Y. Chaveliparambil. *Science Week-2024: Collection of abstracts of the XIV Dudyrev Scientific and Technical Conference of Students, Postgraduates and Young Scientists*. Saint Petersburg: 2024, 355 p. (In Russ.).
8. Kenney T., Gu H., Huang T. Poisson PCA: Poisson Measurement Error Corrected PCA, with Application to Microbiome Data. *Biometrics*. 2021. Vol. 77. No. 4. pp. 1369–1384. DOI: 10.1111/biom.13384.
9. Marukatat S. Tutorial on PCA and Approximate PCA and Approximate Kernel PCA. *Artificial Intelligence Review*. 2023. Vol. 56. No. 6. pp. 5445–5477. DOI: 10.1007/s10462-022-10297-z.
10. Bo. Kang, D. García García, J. Lijffijt et al. Conditional t-SNE: More Informative t-SNE Embeddings. *Machine Learning*. 2021. Vol. 110. No. 10. pp. 2905–2940. DOI: 10.1007/s10994-020-05917-0.
11. Ozgode Yigin B., Saygili G. Confidence Estimation for t-SNE Embeddings Using Random Forest. *International Journal of Machine Learning and*

Learning and Cybernetics. – 2022. – Т. 13. – № 12. – С. 3981–3992. – DOI: 10.1007/s13042-022-01635-2. – Текст: электронный. (In Eng.).
12. Штайнербергер С., Чжан Ю. T-SNE, силовые раскраски и пределы среднего поля // Research in Mathematical Sciences. – 2022. – Т. 9. – № 3. – С. 1–30. – DOI: 10.1007/s40687-022-00340-4. – Текст: электронный. (In Eng.).
13. Амисса Д. К., Яокума В., Ансонг Э. Д. и др. Оценка методов снижения размерности при обнаружении программ Bitcoin: сравни тельный анализ инкрементального PCA и UMAP // Security and Privacy. – 2025. – Т. 8. – № 2. – DOI: 10.1002/spy2.70002. – Текст: электронный. (In Eng.).
14. Леон-Гомес Э. А., Альварес-Меса А. М., Кастельянос-Домингес Г. Расширение данных между наборами данных с использованием UMAP для прогнозирования скорости ветра на основе глубокого обучения // Computers. – 2025.–Т. 14. – № 4. – С. 123. – DOI: 10.3390/computers 14040123. – Текст: электронный. (In Eng.).
15. Мясников Е. Использование UMAP для снижения размерности гиперспектральных данных // FarEastCon 2020. – Владивосток, 2020. – С. 9271656. – DOI: 10.1109/FarEastCon50210.2020.9271656. – Текст: электронный. (In Eng.).
16. Куренков П. В., Карышев М.Ю. Сопоставление отдельных аспектов социально-экономического положения районов Крайнего Севера и приравненных к ним местностей с остальной территорией Российской Федерации // Социально-экономический и гуманитарный журнал. – 2025. – № 1. – С. 37–49. – DOI: 10.36718/2500-1825-2025-1-37-49. – Текст: электронный.
17. Королева У. Н., Лебедева А. А. Социально-экономическое развитие регионов в условиях внешних вызовов // Российские регионы в фокусе перемен: Сборник докладов XIX международной конференции студентов и молодых ученых, Екатеринбург, 14–16 ноября 2024 года. – Екатеринбург: «Ажур», 2025. – С. 169–171.
18. Заварухин В. П., Чинаева Т. И., Чурилова Э. Ю. Регионы России: результаты кластеризации на основе экономических и инновационных показателей // Статистика и экономика. – 2022. – Т. 19. – № 5. – С. 35–47. – DOI: 10.21686/2500-3925-2022-5-35-47. – Текст: электронный.
19. Тимофеева Ю. А. Построение модели рейтинга стран по параметрам «экономический рост», «кластеризация» в странах-лидерах по уровню кластеризации // Труды БГТУ. Серия 5: Экономика и управление. – 2021. – № 2 (250). – С. 134–137.
20. Минаков А. В. Проблемы сбалансированного социально-экономического развития регионов России // Вестник Алтайской академии экономики и права. – 2024. – № 3-3. – С. 420–427.

Cybernetics. 2022 Vol. 13. No. 12. - pp. 3981–3992. DOI: 10.1007/s13042-022-01635-2.

12. Steinerberger S., Zhang Yu. T-SNE, Forceful Colorings, and Mean Field Limits. *Research in Mathematical Sciences*. 2022. Vol. 9. No. 3. pp. 1–30. DOI: 10.1007/s40687-022-00340-4.

13. Amissah D. K., Yaokumah W., Ansong E. D., Appati Ju. K. Evaluating Dimensionality Reduction Techniques in Bitcoin Ransomware Detection: Comparative Analysis of Incremental PCA and UMAP. *Security and Privacy*. 2025. Vol. 8. No. 2. DOI: 10.1002/spy2.70002.
14. Leon-Gomez E. A., Álvarez-Meza A. M., Castellanos-Dominguez G. Cross-Dataset Data Augmentation Using UMAP for Deep Learning-Based Wind Speed Prediction. *Computers*. 2025. Vol. 14. No. 4. P. 123. DOI: 10.3390/computers 14040123.

15. Myasnikov, E. Using UMAP for Dimensionality Reduction of Hyperspectral Data. *FarEastCon 2020*. Vladivostok, 2020. P. 9271656. DOI: 10.1109/FarEastCon50210.2020.9271656.

16. Kurenkov P. V., Karyshev M. Yu. Comparison of Certain Aspects of The Socio-Economic Situation of the Far North Districts and Areas Equated to Them with the Rest of the Territory of the Russian Federation. *Sotsial'no-ekonomicheskiy i gumanitarniy zhurnal*. 2025. No. 1. pp. 37–49. DOI: 10.36718/2500-1825-2025-1-37-49. (In Russ.).

17. Koroleva U. N., Lebedeva A. A. Socio-Economic Development of Regions in the Context of External Challenges. *Russian Regions in the Focus of Change: Proceedings of the XIX International Conference of Students and Young Scientists, Yekaterinburg, November 14-16, 2024*. Yekaterinburg: Azhur Publishing House, 2025. pp. 169–171. (In Russ.).
18. Zavarukhin V. P., Chinaeva T. I., Churilova E. Yu. Regions of Russia: Clustering Results Based on Economic and Innovative Indicators. *Statistika i Economika*. 2022. Vol. 19. No. 5. pp. 35–47. DOI: 10.21686/2500-3925-2022-5-35-47. (In Russ.).

19. Timofeeva Yu. A. Building a Model of the Rating of Countries by the Parameters "Economic Growth", "Clustering" in the Leading Countries by the Level Of Clustering. *Trudy BSTU. Serija 5: Economika i Upravlenije*. 2021. No. 2 (250). pp. 134–137. (In Russ.).
20. Minakov A.V. Problems of Balanced Socio-Economic Development of Russian Regions. *Vestnik Altayskoy akademii ekonomiki i prava*. 2024. No. 3-3. pp. 420–427 (In Russ.).

21. Крыжко Д. А., Смирнова И. А., Конников Е. А., Унгвари Л. Методология снижения размерности в задачах анализа регионального инновационного потенциала: состязательный подход // Экономика Северо-Запада: проблемы и перспективы развития. – 2024. – № 2 (77). – С. 69–77. – DOI: 10. 52897/2411-4588-2024-2. – Текст: электронный.
22. Дастджерди Б. и др. Обзор применимых методов обнаружения выбросов при обработке геомеханических данных // Geotechnics. – 2023. – Т. 3. – № 2. – С. 375–396. – DOI: 10.3390/Geotech nics3020022. – Текст: электронный. (In Eng.).
23. Чжоу Хэ, Цзоу Х. Непараметрическая модель Бокса-Кокса для многомерного регрессионного анализа // Journal of Econometrics. – 2024. – Т. 239. – № 2. – С. 105419. – DOI: 10.1016/j.jeconom.2023. 01.025. – Текст: электронный. (In Eng.).
24. Старовойтов В. В., Голуб Ю. И. Нормализация данных в машинном обучении // Информатика. – 2021. – Т. 18. – № 3. – С. 83–96. – DOI: 10. 37661/1816-0301-2021-18-3-83-96. – Текст: электронный.
25. Письменский А. В. Подготовка данных для машинного обучения // Современные вызовы экономики и систем управления в России в условиях многополярного мира: Сборник статей Международной научно-практической конференции, приуроченной к 105-летию Финуниверситета. – Санкт-Петербург: ООО «Скифия-принт», 2024. – С. 196–204.
26. Федеральная служба государственной статистики. Официальный сайт [Электронный ресурс]. – URL: www.rosstat.gov.ru (дата обращения: 20.02.2026). – Текст: электронный.

21. Kryzhko D. A., Smirnova I. A., Konnikov E. A., Ungvari L. Methodology of Dimension Reduction in Problems of Analysis of Regional Innovation Potential: a Competitive Approach. *Ekonomika Severo-Zapada: problemy i perspektivy razvitiya*. 2024. No. 2 (77). pp. 69–77. DOI: 10. 52897/2411-4588-2024-2. (In Russ.).
22. Dastjerdy B., Saeidi B., Heidarzadeh Sh. Review of Applicable Outlier Detection Methods to Treat Geomechanical Data. *Geotechnics*. 2023. Vol. 3. No. 2. pp. 375–396. DOI: 10.3390/Geotech nics3020022.
23. Zhou He., Zou H. The Nonparametric Box–Cox Model for High-Dimensional Regression Analysis. *Journal of Econometrics*. 2024. Vol. 239. No. 2. P. 105419. DOI: 10.1016/j.jeconom.2023. 01.025.
24. Starovoitov V. V., Golub Yu. I. Data Normalization in Machine Learning. 2021. Vol. 18. No. 3. pp. 83–96. DOI: 10.37661/1816-0301-2021-18-3-83-96. (In Russ.).
25. Pisminskiy A.V. Preparing data for Machine Learning. *Modern Challenges of the Economy and management systems in Russia in a multipolar world: A collection of articles of the International Scientific and Practical Conference dedicated to the 105th anniversary of the Financial University*. Saint-Petersburg: OOO "Scythia-print". 2024. pp. 196–204. (In Russ.).
26. Federal State Statistics Service. Official website [Electronic resource]. URL: www.rosstat.gov.ru (Accessed: 20.02.2026). (In Russ.).

Орлова Екатерина Андреевна / Orlova Ekaterina A.
студент / student
Северо-Западный институт управления – филиал Российской академии народного хозяйства и государственной службы (РАНХиГС) / North-Western Institute of Management – a branch of the Russian Presidential Academy of National Economy and Public Administration (RANEPA)
Санкт-Петербург, Средний пр. В.О., д. 57/43, лит. А
E-mail: eorlova-22@ranepa.ru

Полянская Светлана Владимировна / Polyanskaya Svetlana V.
кандидат технических наук, доцент / PhD, Associate Professor
доцент кафедры бизнес-информатики / Associate Professor of the Department of Business Informatics
Северо-Западный институт управления – филиал Российской академии народного хозяйства и государственной службы (РАНХиГС) / North-Western Institute of Management – a branch of the Russian Presidential Academy of National Economy and Public Administration (RANEPA)
Санкт-Петербург, Средний пр. В.О., д. 57/43, лит. А
E-mail: polyanskaya-sv@ranepa.ru

Статья поступила в редакцию 12.01.2026; одобрена после рецензирования 15.04.2026; принята к публикации 24.06.2026.
The article was submitted 12.01.2026; approved after reviewing 15.04.2026; accepted for publication 24.06.2026.