DOI 10.19181/demis.2026.6.2.14
EDN JYOUPV

Научная статья

ФАКТОРЫ ДЕМОГРАФИЧЕСКОГО БЛАГОПОЛУЧИЯ РОССИЙСКОГО СЕЛА: ПРЕДВАРИТЕЛЬНЫЕ ГИПОТЕЗЫ ДЛЯ ML-МОДЕЛИРОВАНИЯ

Дождиков А. В.
Институт социально-политических исследований ФНИСЦ РАН, Москва, Россия
E-mail: antondnn@yandex.ru

Для цитирования: *Дождиков, А. В.* Факторы демографического благополучия российского села: предварительные гипотезы для ML-моделирования // ДЕМИС. Демографические исследования. 2026. Т. 6, № 2. С. 240–262. DOI 10.19181/demis.2026.6.2.14. EDN JYOUPV.

Аннотация. *Исследование посвящено формулированию предварительных гипотез и подбору методов машинного обучения для количественного анализа демографо-миграционной динамики сельского населения регионов России и ее связи с инфраструктурно-экономическими характеристиками. Рассматриваются зависимости между миграционным и естественным движением населения и уровнем инфраструктуры, инвестициями, жилищным строительством, доступом к услугам и экономической активностью. Цель – построить модель для выявления устойчивых связей между факторами, кластерной типологией регионов и предварительно проверить первичные гипотезы с применением методов машинного обучения. Используется ограниченный набор данных Росстата, агрегированных в панель «регион–год» с очисткой, нормировкой и построением синтетических индексов. Применяются методы EDA, PCA + K-means, регрессии, деревья решений и Random Forest. Для продолжения исследования и последующей корректировки гипотез нами установлено, что: пока не выявлена связь инвестиций с ростом миграционного сальдо, тогда как налицо признаки ассоциации жилищного строительство с замедлением убыли населения; эффекты влияния инфраструктуры ограничены; что необходим больший временной лаг для выявления связей между управленческими решениями и результатами; кроме того, зафиксированы кластерная неоднородность регионов и ограниченная предсказательная способность моделей в отношении миграции. Практическая значимость — улучшение таргетирования политики развития села. Ограничения исследования включают короткую панель, агрегирование и шоки 2020–2023 гг. Перспективы исследования – проверка гипотез с использованием задействованных методов на большом наборе данных регион-год.*

Ключевые слова: *сельское население, внутренняя миграция, демографические процессы, сельские территории, инфраструктура, жилищное строительство, социальная инфраструктура, панельные данные, кластерный анализ, машинное обучение*

Введение

Сельские территории России в XXI в. переживают трансформацию, проявляющуюся в сокращении численности населения, изменении возрастной структуры и миграционном оттоке. В 2023 г. среднегодовая численность сельского населения составила 36,7 млн человек, что на 0,2 млн меньше, чем в 2022 г.; с 2019 г. наблюдается ускорение снижения. К 1 января 2025 г. городское население достигло 109,8 млн человек, сельское – 36,3 млн человек. Рост городского населения обеспечивается миграцией, тогда как сельские территории продолжают его терять. Сохраняется тенденция концентрации населения в крупных городах и «пространственного сжатия» сельских территорий в 2022–2024 гг.[1]

[1] Демография // Федеральная служба государственной статистики : [сайт]. URL: https://rosstat.gov.ru/folder/12781 (дата обращения: 13.03.2026).

Депопуляция сельских территорий носит системный характер: отток и высокая смертность приводят к деформации возрастной структуры. Молодежь мигрирует в города с более высокими доходами и качеством жизни. Государственная реакция выражена в программе «Комплексное развитие сельских территорий» (2019–2030)[2], направленной на улучшение условий жизни; к 2024 г. целевой показатель доли сельского населения (24,9%) формально достигнут.

Научная проблема заключается в недостаточной изученности количественных связей между инфраструктурно-экономическими характеристиками и демографо-миграционными процессами. Требуется комплексный анализ влияния инфраструктуры, инвестиций и жилищного строительства с применением методов машинного обучения, определения временных лагов и построением гипотез.

Актуальность определяется усилением депопуляции, необходимостью доказательной базы для политики и ограниченным использованием ML-подходов. *Объект исследования* – зависимости между демографо-миграционными показателями сельского населения регионов РФ и уровнем развития территорий, *предмет* – методы и инструменты установления зависимостей на основе статистических данных и связанные с ними первичные гипотезы исследования.

Цель – выявить и количественно оценить устойчивые связи, а также сформировать предварительную типологию регионов. *Задачи* включают: разведочный анализ, построение панели «регион–год», кластеризацию, моделирование (регрессии, деревья, ансамбли) и первичную проверку гипотез. Задачи исследования ограничены созданием и предварительным тестированием пилотного набора методов и инструментов.

Научная новизна – формирование воспроизводимой панели данных и применение ML-пайплайна (PCA + K-means, Ridge, Lasso, Random Forest) с предварительной проверкой девяти гипотез. *Практическая значимость* – создание инструментария для обоснования дифференцированной политики развития сельских территорий и повышения эффективности бюджетных вложений.

Предварительные (первичные) гипотезы исследования.

1. Гипотеза инвестиционной нагрузки: более высокий уровень инвестиций в сельские территории ассоциирован с благоприятным миграционным сальдо сельского населения.

2. Гипотеза влияния жилищного строительства: рост объемов ввода жилья в сельской местности сопровождается замедлением убыли сельского населения и/или смягчением миграционного оттока.

3. Гипотеза влияния доступа к услугам: высокая обеспеченность сельского населения объектами торговли, бытовых и социальных услуг связана с численностью сельского населения и меньшим оттоком.

4. Гипотеза влияния обеспеченности транспортом и связью: качество транспортной доступности и развитость почтово-коммуникационной инфраструктуры могут способствовать удержанию населения.

[2] Постановление Правительства РФ от 31 мая 2019 г. № 696 «Об утверждении государственной программы Российской Федерации “Комплексное развитие сельских территорий” и о внесении изменений в некоторые акты Правительства Российской Федерации» // Гарант : [сайт]. URL: https://base.garant.ru/72260516/ (дата обращения: 13.03.2026).

5. Гипотеза влияния социальной инфраструктуры: высокая обеспеченность села объектами здравоохранения, спорта и туризма ассоциирована с лучшим миграционным сальдо.

6. Гипотеза кластерной неоднородности регионов: существует несколько устойчивых кластеров регионов с принципиально разными моделями сельской демографической динамики.

7. Гипотеза неравномерной отдачи: прирост инфраструктурных показателей (например, услуг или дорог) в регионах с очень низкой исходной обеспеченностью дает более сильный положительный эффект на демографию и миграцию, чем такой же прирост в уже относительно развитых сельских территориях (нелинейный/пороговый эффект).

8. Гипотеза скорой отдачи: влияние инфраструктурно-экономических изменений проявляется сразу, то есть улучшение инфраструктуры в сельских территориях быстро (за год) отражается в демографо-миграционных показателях.

9. Гипотеза структурных разрывов: в периоды шоков (например, COVID-19) связи между экономико-инфраструктурными показателями и демографией временно изменяются (структурный сдвиг), что можно выявить как смену параметров моделей или кластерной структуры.

Обзор научной литературы

В современных исследованиях принято выделять концепцию «природных удобств», согласно которой природно-рекреационный потенциал территории способен выступать драйвером притока. Параллельно развивается подход «нового экономического роста в деревне» [1], рассматривающий инфраструктуру [2] как предпосылку удержания людей и формирования устойчивых сообществ.

По данным Всероссийской переписи населения 2020 г., четверть населения страны все еще проживает в сельской местности. Между переписями 2010 и 2020 гг. численность сельского населения росла лишь в 20 регионах. В большинстве регионов (53 из 89) сельское население сокращалось за счет миграционного оттока[3].

Пространственная неоднородность демографической динамики сельских территорий хорошо задокументирована в российской литературе: субурбанизационные и периферийные различия в миграционном балансе сельских территорий выражены сильнее, чем пространственно-географические (зональные). Пригородные села интенсивно прирастают за счет миграции, тогда как периферия испытывает сильный миграционный отток [3]. Эти выводы получили подтверждение в работе [4] установившей, что сельско-городская миграция в России рациональна и экономически обоснована: переезд в город улучшает положение на рынке труда и ведет к росту доходов.

Вместе с тем укажем на ускорение депопуляции с 2019 г. и существенное пространственное расхождение демографических показателей по группам регионов.

[3] Переписи населения // Федеральная служба государственной статистики : [сайт]. URL: https://rosstat.gov.ru/perepisi_naseleniya (дата обращения: 13.03.2026).

Данная проблематика рассматривается в публикациях ВШЭ[4] и аналитических докладах профильного министерства[5].

Связь между состоянием инфраструктуры и демографическими процессами в сельских территориях исследована в ряде отечественных и зарубежных работ. Авторы [5; 6], анализируя данные Росстата за 2014 и 2022 гг. по 82 субъектам РФ, констатировали значительное отставание сельских домохозяйств от городских по качеству инженерной инфраструктуры, подчеркнув при этом критическую роль ее модернизации для повышения качества жизни.

Развитие социальной инфраструктуры как фактора удержания населения и снижения миграционного оттока рассматривается в работе [7]. Исследователи показали дифференциацию в обеспеченности объектами социальной инфраструктуры сельских поселений в зависимости от их размера и пространственного положения, указав на то, что приоритетное инфраструктурное развитие опорных населенных пунктов повышает обеспеченность услугами, однако может углублять дифференциацию между поселениями. В аналогичном направлении работают авторы [8], исследования которых по пространственному неравенству и типологии сельских регионов России задают рамочную концепцию для анализа демографических различий.

В зарубежной литературе влияние инфраструктуры на устойчивость сельских домохозяйств изучено на панельных данных Таиланда и Вьетнама [9]: транспортная инфраструктура и ИКТ улучшают абсорбционную способность домохозяйств к экономическим шокам и снижают риск бедности. Дороги, электроснабжение, системы водоснабжения, медицинские учреждения и школы – ключевые предпосылки долгосрочного экономического роста сельских территорий. Параллельно накоплена обширная доказательная база, свидетельствующая о том, что жилищная обеспеченность в сельской местности непосредственно связана с демографической динамикой. Так, в работах по ЕАЭС [10] и Киргизской Республике [11] отражена сильная положительная взаимосвязь между объемом жилищного фонда и численностью сельского населения.

Относительно транспортной инфраструктуры существует академически зафиксированный «двойной эффект»: с одной стороны, дороги снижают транспортные барьеры и расширяют доступ к услугам, что потенциально удерживает население; с другой – облегчают выезд за счет снижения стоимости миграции [12]. Это противоречие отражено в концепции «транспортного парадокса» и фигурирует в гипотезе 4 настоящего исследования.

Важную роль в изучении сельской демографии играет подход, основанный на типологизации регионов. На уровне российских регионов [13] выявлены два ме-

[4] Демография и миграция. Информационный бюллетень «Тренды. События. Цифры». Выпуск 2. Июнь-август 2024 // Стратегические технологические проекты НИУ ВШЭ : [сайт]. URL: https://stratpro.hse.ru/mirror/pubs/share/1015020046.pdf (дата обращения: 13.03.2026).

[5] Доклад о результатах проведенного мониторинга состояния социально-экономического развития сельских территорий Российской Федерации // Министерство сельского хозяйства Российской Федерации : [сайт]. URL: https://mcx.gov.ru/upload/iblock/aa9/882w7zzcx3bn8qk6uu348iutyi8zdsue.pdf (дата обращения: 13.03.2026).

гаполиса: аграрно-традиционные республики Юга и Северного Кавказа, промышленно-урбанизированный пояс и Север/Дальний Восток – с различающимися связями между плотностью населения и социально-экономическими индикаторами.

В международной практике сочетание PCA [14] и кластеризации K-means признано надежной стратегией многомерного анализа. К примеру, для Турции аналогичная методология PCA + K-means [15] позволила выявить существенные различия по миграционным паттернам, темпам роста населения и социально-экономическим характеристикам между провинциями. GRANULAR-исследование[6] по типологии сельских территорий ЕС показывает, что регионы делятся на преимущественно сельские (>50% населения в сельских муниципалитетах), промежуточные и преимущественно городские, с различными механизмами демографического воспроизводства.

Пандемия COVID-19 вызвала значительный интерес к потенциальному «контрурбанизационному» сдвигу в миграционных паттернах. Так, в отношении США [16] было установлено, что начиная с 2020 г. отток из сельских территорий снизился и оставался низким на протяжении трех последующих лет, что привело к стабилизации и частичному росту населения. В Испании [17] был зафиксирован рост миграционного притока в сельские районы, расположенные вблизи городов, и территории с высокой долей вторичного жилья в 2020 г. по сравнению с периодом 2016–2019 гг. В Сербии [18] анализ показал, что после первоначального снижения в 2020 г. интенсивность переселений в 2021 г. превысила допандемийный уровень, а неблагоприятные экономические характеристики сельских мигрантов несколько улучшились в пандемийные годы.

В то же время результаты международных исследований [19] не свидетельствуют о «массовом исходе» из городов. Скорее речь идет о перераспределении паттернов внутри сложившихся систем расселения. Аналогичный вывод применим и к России: основной мотив переезда в город по-прежнему носит экономический характер.

Методология и методы исследования, источники информации

Использование методов ML в миграционных и демографических исследованиях активно реализуется начиная с 2010-х гг. Модели случайного леса превосходят логистическую регрессию и метод опорных векторов в предсказании миграционных решений домохозяйств, выступая как перспективный инструмент для выявления нюансированных паттернов в сложных социальных наборах данных. Применительно к городскому планированию и распределению мигрантов Random Forest [20] позволил выделить демографическую и этническую композицию района как ключевой фактор в пространственном распределении мигрантов [21]. Метод SHAP (SHapley Additive exPlanations) [22] в сочетании с Random Forest показал высокую эффективность при анализе пространственной справедливости и экономической мобильности.

[6] Stjernberg, M., Norlén, G., Vasilevskaya, A., et al. Scoping Report on European Rural Typologies // Zenodo : [site]. 31.05.2023. URL: https://zenodo.org/records/13767183 (accessed on 13.03.2026).

В российском контексте ансамблевые ML-методы применялись для определения детерминант ВРП регионов: Light GBM показал R^2=0,7345, притом ведущую роль сыграли доходы населения и иностранные инвестиции с лаговым эффектом [23].

Регрессионный анализ [14] изучавших миграционные потоки между городскими и сельскими районами России в 2011–2020 гг. с использованием цепей Маркова и модели пространственного взаимодействия, показал, что высока вероятность миграции из сельской в городскую местность, а среди детерминантов значимую роль играют численность населения, заработная плата, занятость, жилищная доступность и осадки, при этом эффекты варьируются в зависимости от типа отправного пункта (сельская или городская местность).

Исследование опирается на *количественную методологию* и реализует пайплайн последовательных аналитических операций: разведочный анализ данных (EDA) – формирование синтетических панельных данных – многомерная кластеризация – регрессионное и ML-моделирование – первичная проверка гипотез. С учетом ограничений объема информации выводы носят предварительный характер.

Первичную информационную базу исследования составляют официальные статистические данные Федеральной службы государственной статистики (Росстат), включая:

– бюллетени Росстата[7]: численность прибывших и выбывших сельских жителей по субъектам РФ, расчет миграционного сальдо на 1000 человек сельского населения;

– данные текущего демографического учета: естественное движение сельского населения (рождаемость, смертность, естественный прирост), изменение численности;

– инфраструктурно-экономические показатели сельских территорий: инвестиции в основной капитал на одного сельского жителя, ввод жилья, плотность объектов торговли и бытовых услуг, протяженность дорог с твердым покрытием, показатели ЖКХ, численность объектов социальной инфраструктуры (здравоохранение, спорт, туризм);

– показатели экономической активности: индексы, отражающие занятость, валовую добавленную стоимость и интенсивность природопользования в сельских территориях регионов.

Показатели за годы объединены в интегрированную панель «регион–год» с применением процедур очистки, нормировки и синтетического дополнения пропусков. Для устранения влияния выбросов применялась винзоризация переменных[8]. Воспроизводимый код и полная документация исследования доступны в открытом фиксированном репозитории[9]: данные конструировались из источников

[7] Численность и миграция населения Российской Федерации. URL: https://www.rosstat.gov.ru/compendium/document/13283 (дата обращения: 13.04.2026).

[8] Винзоризация (winsorization) – это статистический метод преобразования данных, который заключается в замене экстремальных значений в наборе данных менее экстремальными значениями с целью снижения влияния выбросов на анализ. Метод назван в честь статистика Чарльза П. Винзора.

[9] Машинный код для тестирования гипотез и результаты тестирования: Rural_9Hyp.ipynb // GitHub : [site]. URL: https://github.com/AntonDozhdikov/Demography_migration/blob/main/Rural_9Hyp.ipynb (дата обращения: 13.03.2026).

по экономике и инфраструктуре[10] и демографическим показателям[11] на основе актуальной статистики, но при этом недостаточно полной и требующей уточнения.

Перспективное направление развития исследования – это повтор эксперимента на данных с 2014 по 2025 г. или более широкого периода, что требует дополнительных исследований и сбора информации.

Для операционализации сложных многомерных понятий (доступность услуг, состояние транспортно-жилищной и социальной инфраструктуры, экономическая активность) сконструированы синтетические индексы путем агрегирования нормированных показателей. Формально, для каждого индекса I_k по региону i и году t применялась z-стандартизация:

$$I_{k,i,t} = \frac{1}{|V_k|} \sum_{j \in V_k} \frac{x_{j,i,t} - \bar{x}_j}{\sigma_{x_j}}$$

где V_k – множество входящих переменных индекса k, $\bar{x}_j$ – выборочное среднее, σ_{x_j} – стандартное отклонение переменной j (приведена общая формула).

Для минимизации влияния экстремальных наблюдений переменные перед включением в модели были винзоризованы на уровне 1-го и 99-го процентилей. Лаговые версии ключевых переменных (например, invest_per_rural_capita_lag1) формировались путем сдвига временных рядов на один период вперед с сортировкой по региону и году.

Регрессионный анализ (OLS). Для проверки каждой из гипотез 1–5 и 8–9 строились линейные регрессионные модели вида:

$$y_{it} = \alpha + \beta_1 x_{1,it} + \beta_2 D_{2023} + \varepsilon_{it}$$

где y_{it} – зависимая демографо-миграционная переменная (миграционное сальдо или изменение численности населения на 1000 жителей), $x_{1,it}$ – целевой предиктор (инфраструктурный или экономический индикатор), D_{2023} – временная дамми-переменная для 2023 г., учитывающая общий временной сдвиг. Значимость коэффициентов оценивалась по двустороннему критерию при уровне значимости α=0,05 (пограничено – α=0,1). Для взаимодействия факторов (гипотезы 4, 9) использовались модели с interaction-термином:

$$y_{it} = \alpha + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \beta_3 (x_1 \cdot x_2)_{it} + \gamma D_{2023} + \varepsilon_{it}$$

Для гипотезы 8 строилась лаговая регрессия, включающая $x_{1,i,t-1}$ вместо $x_{1,i,t}$.

Квартильный анализ. В дополнение к регрессиям для каждого предиктора рассчитывались описательные статистики зависимой переменной внутри квартильных групп распределения предиктора (Q1–Q4), что позволяло выявить нелинейные и пороговые эффекты, не улавливаемые линейными моделями (8 гипотеза).

Кластерный анализ (6 гипотеза). Для формирования типологии регионов применялась связка методов: 1) стандартизация признаков; 2) снижение размерности с помощью метода главных компонент (PCA) – выделение двух ведущих компонент

[10] Источник данных для пилотного исследования по инфраструктуре: rural_econ_infra_othes_panel_2014_2024.csv // GitHub : [site]. URL: https://github.com/AntonDozhdikov/Demography_migration/blob/main/rural_econ_infra_othes_panel_2014_2024.csv (дата обращения: 13.03.2026).

[11] Источник данных для пилотного исследования по миграции: rosstat_rural_panel_2014_2024.csv // GitHub : [site]. URL: https://github.com/AntonDozhdikov/Demography_migration/blob/main/ rosstat_rural_panel_2014_2024.csv (дата обращения: 13.03.2026).

PC1 и PC2; 3) кластеризация K-means[12] с числом кластеров K=4. Значимость различий демографических исходов между кластерами проверялась непараметрическим критерием Краскала-Уоллиса[13]:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

где R_j – сумма рангов в кластере j, n_j – его размер, n – общее число наблюдений.

ML-блок: регуляризованные регрессии. Для оценки совокупного вклада инфраструктурно-экономических факторов в объяснение дисперсии миграционного сальдо применялись три регуляризованных линейных метода. *Гребневая (Ridge) регрессия* минимизирует целевую функцию с L2-штрафом:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Lasso-регрессия использует L1-штраф, осуществляющий отбор переменных:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

ElasticNet объединяет оба типа регуляризации с параметром смешения $\alpha \in [0,1]$

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right] \right\}$$

Оптимальные значения λ и α подбирались методом кросс-валидации. При применении жесткой регуляризации Lasso и ElasticNet занулили все коэффициенты предикторов.

ML-блок: Random Forest (нелинейный ансамбль). Для выявления нелинейных и взаимодействующих эффектов строилась модель *регрессии случайного леса* (Random Forest). Оценки важности признаков (feature importance) рассчитывались как средняя примесь (mean impurity decrease) по всем деревьям ансамбля. Random Forest не требует предположения о линейности связей и устойчив к мультиколлинеарности, что делает его адекватным инструментом при наличии коррелированных инфраструктурных индексов.

ML-блок: логистическая регрессия-классификатор. Для оценки предсказуемости знака миграционного сальдо (положительное/неположительное) решена задача

[12] На большем объеме данных целесообразно использовать и DBSCAN и нефиксированное число кластеров, определяемых при помощи «метода локтя». DBSCAN (Density-Based Spatial Clustering of Applications with Noise) – алгоритм кластеризации, основанный на плотности данных. Он группирует точки, которые тесно расположены в пространстве, и помечает как выбросы точки, находящиеся в областях с малой плотностью. Метод локтя (elbow method) – инструмент для определения оптимального числа кластеров в алгоритмах кластеризации.

[13] Критерий Краскела – Уоллиса (H-критерий) – это непараметрический метод в математической статистике, который используется для сравнения медиан трех и более независимых выборок. Он позволяет определить, существуют ли статистически значимые различия между группами по уровню какого-либо признака.

с балансировкой классов (class_weight='balanced') и групповой кросс-валидацией по регионам (GroupKFold, 5 фолдов). Качество оценивалось по метрикам Accuracy, ROC AUC, Precision, Recall и F1-score для каждого класса. Применение GroupKFold предотвращает «утечку» данных между временными наблюдениями одного региона в обучающую и тестовую выборку, что обеспечивает несмещенную оценку обобщающей способности модели.

Ключевые *ограничения исследования:*

1. Недостаток данных: короткий временной ряд снижает мощность тестов и не позволяет оценить лаговые эффекты. Требуется расширение выборки.

2. Линейные модели: не учитывают индивидуальные региональные эффекты, что может приводить к смещению оценок. Для точности нужны ансамблевые модели (к примеру, CatBoost), но они усложняют интерпретацию, требуют оценки признаков («весов модели») через feature impotence [14] и SHAP[15].

3. Синтетические индексы: агрегируют разнородные показатели, маскируя противоположные эффекты отдельных компонентов.

4. Геополитический контекст (2022–2023 гг.): миграционные шоки трудно отделить от структурных изменений; период менее показателен, чем 2020 г.

Исследование – пилотное, предварительный анализ для дальнейшей проверки на большей выборке и тонкой настройки моделей.

Результаты эксперимента

Гипотеза 1 не подтвердилась: инвестиции в сельского жителя не связаны с миграционным сальдо (рис. 1).

В линейной регрессии (174 наблюдения) коэффициент при логарифме инвестиций положителен (1,15), но статистически незначим ($p=0,75$; $R^2=0,017$).

Единственная погранично значимая переменная – индикатор 2023 г. (коэффициент ≈9,0; $p\approx0,075$), что отражает общий сдвиг режима, а не эффект инвестиций.

Квартильный анализ не выявил устойчивой связи: среднее сальдо варьируется от ≈−5,6 до ≈4,3, при высоких стандартных отклонениях (до 66,6), что говорит о сильной внутригрупповой вариативности и отсутствии монотонной зависимости.

Вывод: по данным 2022–2023 гг. более высокий уровень инвестиций в сельского жителя не демонстрирует статистически значимой и устойчивой связи с миграционным сальдо.

Гипотеза 2 оказалась верной: рост ввода жилья в сельской местности статистически значимо связан с замедлением убыли населения (рис. 2).

В модели по 174 наблюдениям коэффициент при винзоризованном вводе жилья положителен и значим (0,0068; $p\approx0,0002$; $R^2=0,122$), что указывает на связь между строительством жилья и менее негативной динамикой численности.

[14] Мера того, насколько каждый признак (переменная) в целом влияет на предсказания модели машинного обучения. Она позволяет определить, какие признаки являются наиболее значимыми для модели, и ранжировать их по степени влияния на итоговый результат.

[15] SHapley Additive exPlanations – это метод интерпретации прогнозов моделей машинного обучения, основанный на теории игр (значения Шепли). Он объясняет, какой вклад каждый признак внес в конкретное предсказание модели (в отличие от общего feature impotence).

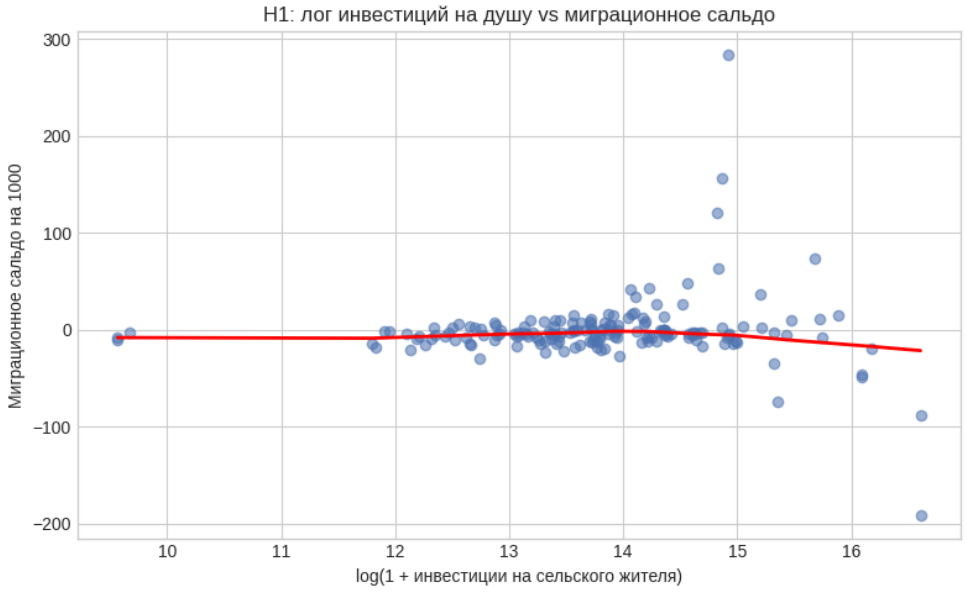

Рис. 1. **Отображение зависимости инвестиций и миграционного сальдо**
Fig. 1. **Display of the dependence of investments and migration balance**
Источник: ячейка репозитория 22. H1 – инвестиции и миграционное сальдо

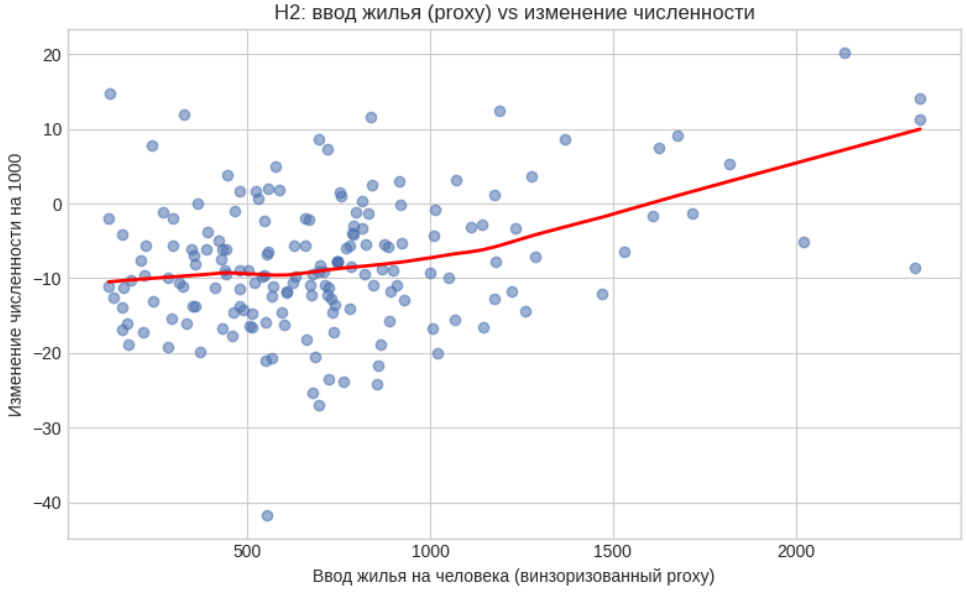

Рис. 2. **Отображение зависимости ввода жилья и изменения численности населения**
Fig. 2. **Display of the relationship between housing input and population change**
Источник: ячейка репозитория 23. H2 – жилищное строительство и изменение численности

Квартильный анализ: в нижнем квартиле убыль составляет ≈–8,55 на 1 000, в верхнем ≈–4,01 на 1 000. Жилищное строительство смягчает, однако не компенсирует полностью убыль.

Вывод: более высокий ввод жилья на селе, предположительно, связан с замедлением убыли населения. Но требуется уточнение и перепроверка на большем объеме данных.

Гипотеза 3 получила частичное подтверждение на ограниченном наборе данных: более высокая обеспеченность услугами связана с благоприятным миграционным сальдо, но эффект слабый и нестабильный (рис. 3).

В регрессии коэффициент при сервисном индексе положителен (3,77), p≈0,097, R^2≈0,029 – связь неустойчива.

Квартильный анализ: в Q1 среднее сальдо ≈1,63 (высокая дисперсия), в Q2–Q3 – отрицательные значения, в Q4 – положительное (≈3,12). Медиана смещается к менее негативным значениям. Вывод: сервисная доступность в среднем связана с лучшим сальдо, однако эффект слаб, нестабилен и требует дальнейшего изучения.

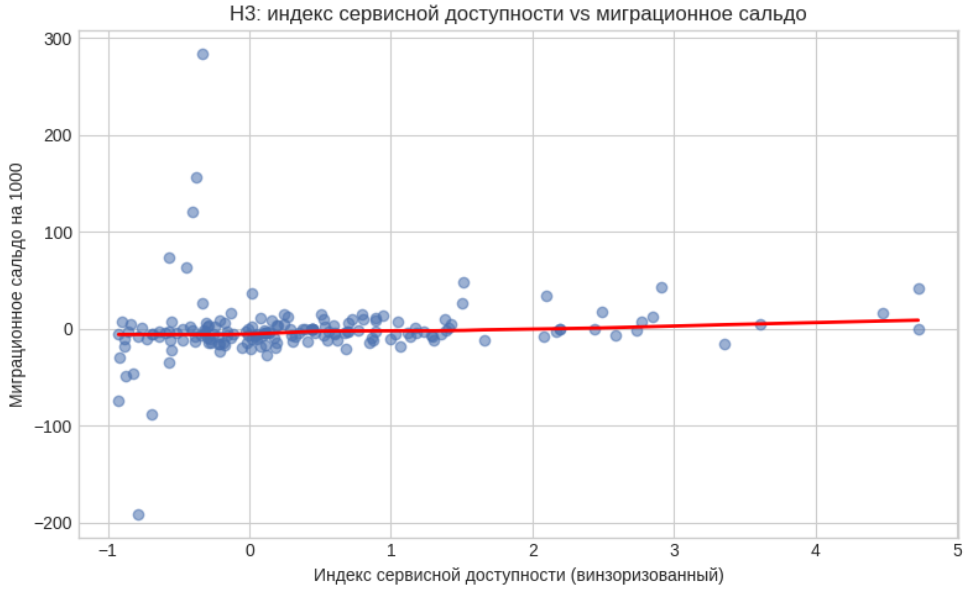

***Рис. 3.* Связь сервисных параметров и качества услуг с миграционным сальдо**
***Fig. 3.* The relationship between service parameters and service quality with the migration balance**

Источник: ячейка репозитория 24 H3 – доступность услуг и миграция

Гипотеза 4 на ограниченном наборе не подтвердилась: совместный эффект транспортно-жилищной инфраструктуры и экономической активности на миграционное сальдо статистически незначим (рис. 4).

В регрессии коэффициент при interaction положителен (≈0,97; p≈0,34; R^2≈0,049), что не фиксирует усиления ни удержания, ни оттока населения.

Отдельные индексы инфраструктуры и экономики также незначимы, их эффекты разнонаправлены и «зашумлены».

Квартильный анализ показывает высокую вариативность и отсутствие устойчивой зависимости: в Q1 сальдо положительное, в Q2–Q3 – отрицательное, в Q4 – близко к нулю.

Вывод: на ограниченной выборке не выявлено устойчивого мультипликативного эффекта. Влияние инфраструктуры и экономики на миграцию, скорее всего, нелинейно и требует более длительного периода наблюдений.

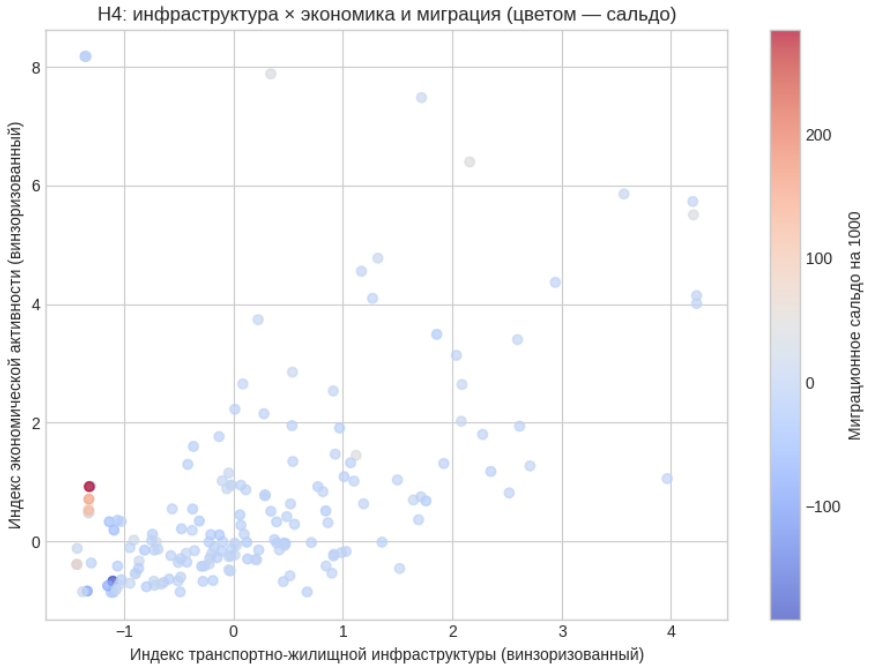

***Рис. 4.* Связь транспортно-жилищной инфраструктуры и экономической активности с миграционными показателями**

***Fig. 4.* The relationship between transport and housing infrastructure and economic activity with migration indicators**

Источник: ячейка репозитория 25. H4 – совместный эффект инфраструктуры и экономики

Гипотеза 5 строгого подтверждения на малой выборке не получила: линейной связи между социальной инфраструктурой и миграционным сальдо не выявлено, но описательные различия квартилей указывают на возможную слабую положительную ассоциацию (рис. 5).

В регрессии коэффициент при social_infra_index положителен (≈1,65), но статистически незначим ($p \approx 0{,}32$; $R^2 \approx 0{,}02$).

Квартильный анализ: в Q1 среднее сальдо ≈0,93 (высокая дисперсия), в Q2–Q3 – отрицательное (≈–5,04 и –2,97), в Q4 – положительное (≈3,77), медиана смещается к менее негативным значениям.

Переменная 2023 г. отражает общий временной сдвиг, а не специфический эффект инфраструктуры.

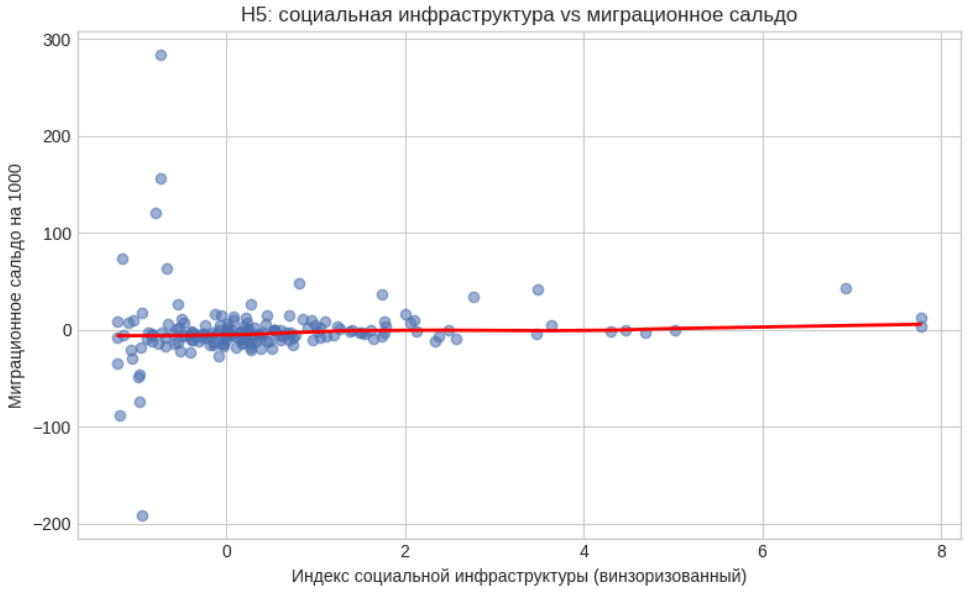

***Рис. 5.* Иллюстрация связи между индексом социальной инфраструктуры и миграционным сальдо**

***Fig. 5.* Illustration of the relationship between the social infrastructure index and the migration balance**

Источник: ячейка репозитория 26. H5 – социальная инфраструктура и миграция

Вывод: статистически значимой линейной связи нет, вместе с тем описательные данные допускают слабую положительную ассоциацию. Для проверки влияния на смертность и естественный прирост требуется отдельный анализ на большем объеме данных.

Гипотеза 6 подтвердилась: выявлены устойчивые кластеры сельских территорий с разными моделями демографической динамики и чувствительностью к инфраструктурно-экономическим факторам (рис. 6).

На основе индикаторов построены главные компоненты (PC1, PC2), выполнена кластеризация K-means.

Тест Краскела-Уоллиса ($H \approx 18{,}95$; $p < 0{,}001$) показал значимые различия демографических исходов между кластерами.

Это свидетельствует о кластерной неоднородности и содержательной интерпретируемости структуры, при этом детальная характеристика кластеров требует дополнительного анализа. Дальнейшее развитие гипотезы может быть связано с выделением условных регионов с сильным притяжением крупных городов, ресурсодобывающих регионов, традиционных аграрных регионов и депрессивных периферийных регионов, инфраструктурно-экономические факторы внутри этих кластеров, вероятно, различаются по знаку и величине.

Гипотеза 7 подтверждена частично: эффект прироста социальной инфраструктуры нелинеен, но максимальная отдача в слабо обеспеченных регионах не доказана (рис. 7).

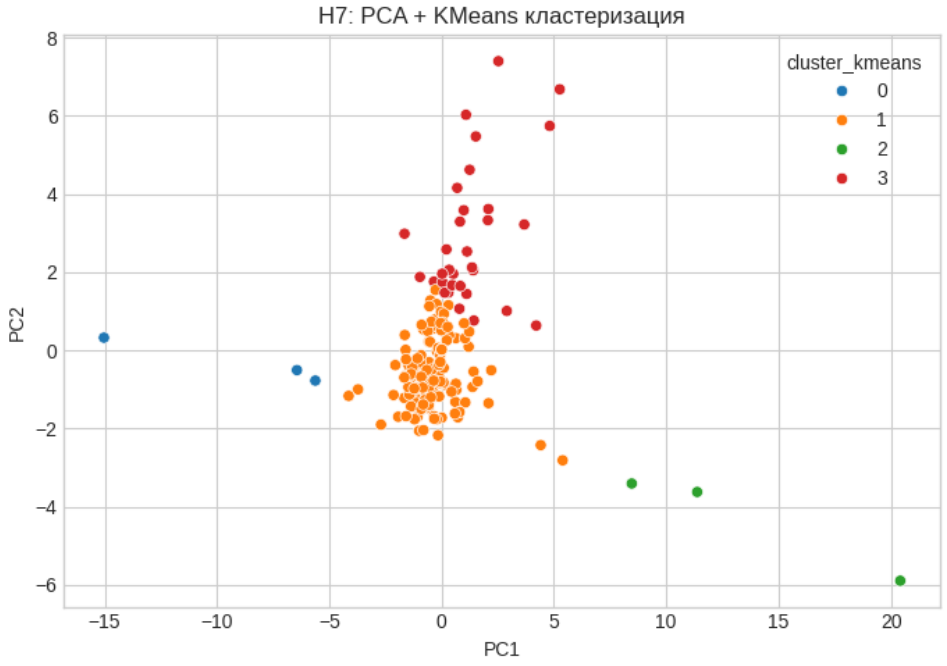

Рис. 6. **Кластеризация регионов Российской Федерации**
Fig. 6. **Clustering of regions of the Russian Federation**
Источник: ячейка депозитария 28. H7 — кластеризация региональных профилей

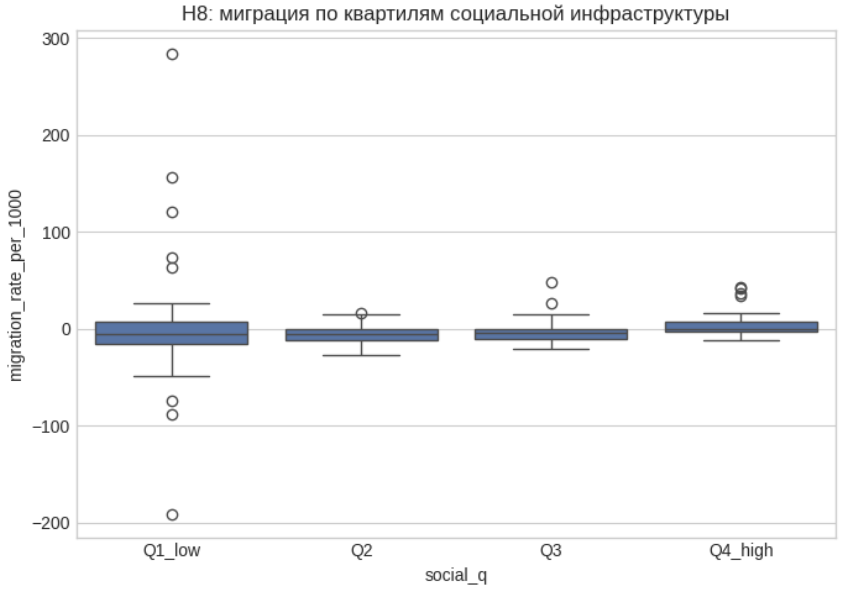

Рис. 7. **Квартильное распределение по показателям социальной инфраструктуры**
Fig. 7. **Quartile distribution of social infrastructure indicators**
Источник: ячейка репозитория 29. H8 – квартили социальной инфраструктуры

Квартильный анализ: в Q1 среднее сальдо ≈0,93 (высокая дисперсия), в Q2–Q3 – устойчиво отрицательное (≈–5,0 и –3,0), в Q4 – положительное (≈3,77). Медиана близка к нулю или слабо отрицательна.

Это указывает на ступенчатый переход: более развитая инфраструктура снижает отток, но особого эффекта малых приростов в самых бедных регионах не выявлено.

Вывод: гипотеза о максимальной отдаче в наименее обеспеченных территориях требует проверки на большем объеме данных.

Гипотеза 8 не подтвердилась[16]: однолетний лаг инвестиций не влияет на миграционное сальдо. В регрессии по 87 наблюдениям коэффициент при лаговом показателе близок к нулю (≈$-1{,}36\cdot10^{-6}$), p≈0,71, R^2≈0,073. Если отложенный эффект и существует, он проявляется на более длительном горизонте (3+ лет) или имеет нелинейный, кластер-специфический характер (рис. 8).

Гипотеза 9 также не подтверждена[17]: структурного сдвига связи социальной инфраструктуры и миграции после 2020 г. не выявлено. Interaction-термин «social_infra_index × post_2020» статистически незначим (p≈0,32), как и базовый индекс социальной инфраструктуры (p≈0,32), и сам индикатор post_2020 (p≈0,25); R^2≈0,020. Изменения миграционного сальдо, возможно, связаны с другими факторами, а не с перестройкой инфраструктурного эффекта.

Линейные и нелинейные модели. Совокупный вклад инфраструктурно-экономических факторов в миграционное сальдо ограничен[18]. В Ridge-регрессии наибольшие веса у индекса экономической активности (≈1,78) и социальной инфраструктуры (≈0,58), но все коэффициенты малы – ни один фактор не является определяющим.

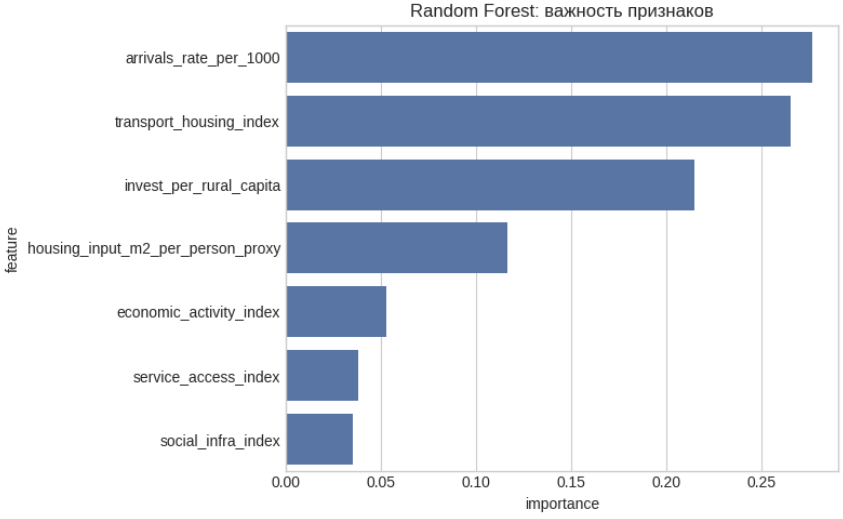

***Рис.* 8. Важность признаков модели**
***Fig.* 8. Importance of model features**
Источник: ячейка 33. ML-блок – random forest importance

[16] Ячейка репозитория 30. H9 – лаговый эффект инвестиций.
[17] Ячейка репозитория 31. H10 – условный структурный сдвиг после 2020 г.
[18] Ячейка репозитория 32. ML-БЛОК – регрессионные модели.

В Lasso и ElasticNet все коэффициенты зануляются: добавление инфраструктурных и экономических показателей не дает устойчивого прироста объясненной дисперсии, что говорит о слабом и нестабильном линейном сигнале.

В Random Forest[19] наибольшую важность имеют текущие потоки прибывающей миграции (≈0,28) и транспортно-жилищная инфраструктура (≈0,27), далее – инвестиции (≈0,21) и ввод жилья. Экономическая и социальная инфраструктура дают меньший, но ненулевой вклад (рис. 8).

Вывод: на ограниченных данных краткосрочные различия миграционного сальдо между сельскими территориями в основном объясняются инерцией миграционных потоков и состоянием базовой инфраструктуры.

Классификация положительного миграционного сальдо. Для оценки предсказуемости факта положительного миграционного сальдо была построена логистическая регрессия с балансировкой классов и групповым кросс-валидированием по регионам (GroupKFold, 5 фолдов). В качестве признаков использовались те же инфраструктурно-экономические индикаторы, что и в регрессионном ML-блоке, целевой переменной выступал бинарный признак «миграционное сальдо >0» (табл. 1).

Таблица 1

Таблица ключевых метрик модели

Table 1

Table of key model metrics

Метрика	Класс 0 (≤ 0)	Класс 1 (> 0)	В целом
Precision	0,79	0,43	–
Recall	0,57	0,68	–
F1-score	0,66	0,52	–
Accuracy	–	–	0,60
ROC AUC	–	–	0,65

Источник: ячейка 34. ML-блок – классификация положительной миграции

Модель показывает умеренное качество: точность составляет около 0,60, а ROC AUC ≈0,65, то есть предсказательная способность заметно выше случайного угадывания, но далека от уровня, достаточного для надежного прогнозирования статуса территорий. При этом поведение классов асимметрично: для территорий с неположительным миграционным сальдо (класс 0) точность высока (precision ≈0,79), однако полнота ограничена (recall ≈0,57), тогда как для положительного сальдо (класс 1) точность ниже (≈0,43), зато полнота выше (≈0,68). Матрица ошибок отражает этот дисбаланс: из 118 наблюдений класса 0 корректно классифицированы 67, 51 – отнесены к положительному классу, а из 56 наблюдений класса 1 правильно предсказаны 38, 18 ошибочно – отнесены к классу 0.

Такая конфигурация модели показывает, что по доступным инфраструктурно-экономическим признакам различение территорий с положительным и неположительным миграционным балансом ограничено. Отрицательное сальдо предсказывается относительно надежно, а положительное – с заметным числом ложных сра-

[19] Ячейка репозитория 33. ML-блок – random forest importance.

батываний и пропусков. Это подчеркивает высокую роль неучтенных факторов (институциональных, поведенческих, культурно-этнических и иных) и одновременно недостаток данных для обучения модели.

Обсуждение результатов

Прямое влияние инвестиций на сельское миграционное сальдо не подтвердилось (коэффициент ≈1,15; p≈0,75; R^2≈0,02). Эффект инвестиций проявляется опосредованно и нелинейно [14], а для его выявления нужны отраслевые, качественные данные и длинные временные ряды. В текущей спецификации смешаны разные типы вложений, что затрудняет анализ.

Единственная значимая связь выявлена для ввода жилья: его рост на сельского жителя снижает миграционную убыль (коэффициент ≈0,0068; p≈0,0002; R^2≈0,12; разница между квартилями: с −8,55 до −4,01). Что соответствует международным [25] и российским данным[20], где жилье и льготная ипотека – ключевые инструменты удержания населения.

Индекс сервисной доступности показал положительную, но погранично значимую связь с миграционным сальдо (p≈0,097; R^2≈0,03). Это говорит об ограниченности линейной модели для оценки нелинейного и гетерогенного эффекта услуг, что согласуется с высокой дифференциацией их обеспеченности между сельскими поселениями [7].

Мультипликативная гипотеза о совместном влиянии инфраструктуры и экономической активности не подтвердилась (p≈0,34; R^2≈0,05). Негативный коэффициент транспортного индекса может отражать «транспортный парадокс»: развитая доступность облегчает и удержание, и отток. Требуется дезагрегированный анализ с разделением территорий и учетом маятниковой миграции.

Возможно изменение формулировки четвертой гипотезы для проверки на большом наборе данных в следующем виде: качество транспортной доступности (протяженность дорог с твердым покрытием) и развитость почтово-коммуникационной инфраструктуры одновременно могут способствовать как удержанию населения (через доступ к услугам), так и усилению маятниковой и окончательной миграции (через снижение транспортных барьеров); знак эффекта зависит от уровня развития экономики региона.

Индекс социальной инфраструктуры имеет положительный, но статистически незначимый коэффициент (p≈0,32; R^2≈0,02). Квартильный анализ подтверждает пороговый и нелинейный характер ее влияния: в регионах с неразвитой инженерной инфраструктурой базовый уровень социальных услуг не влияет на миграцию, пока не будет достигнут определенный «порог». Исследователи из Японии [25] отмечают, что переезд в город повышает удовлетворенность жизнью.

Критерий Краскела-Уоллиса (H≈18,95; p<0,001) указывает на значимые различия демографических исходов между кластерами. Выделен один крупный и несколько малочисленных кластеров, что отражает типологическое разнообразие российских сельских территорий: доминируют «типичные» регионы, а особые

[20] Сельская ипотека в 2026 году: ответы на популярные вопросы // Домклик : [сайт]. 15.12.2025. URL: https://blog.domclick.ru/ipoteka/post/selskaya-ipoteka-otvety-na-populyarnye-voprosy (дата обращения: 13.03.2026).

группы (условно ресурсные, аграрные, нестандартные) – редкость. Это согласуется с существующими типологиями, но требует уточнения границ и методов кластеризации.

Квартильный анализ выявил пороговый эффект: в верхней группе по индексу социальной инфраструктуры миграционное сальдо существенно лучше (≈3,77 против ≈–5,04 во втором квартиле), а в зоне минимальной обеспеченности (Q1) маржинальный эффект отсутствует из-за высокой дисперсии (std≈65,0). Это доказывает, что улучшения ниже минимального уровня не влияют на миграцию, а эффект накапливается при уже относительно высокой базовой обеспеченности. Аналогичные результаты отмечены в период COVID-19: приток наблюдался только в районах с развитой инфраструктурой и близостью к городам, а не в депрессивных территориях [17; 26].

Однолетний лаг инвестиций не оказал значимого влияния на миграционное сальдо (p≈0,71), что соответствует представлению о долгосрочном характере инфраструктурных изменений. Сдвиг влияния инфраструктуры после 2020 г. не обнаружен, что расходится с международными данными о пандемии. В России это может объясняться слабой контрурбанизацией и геополитическими шоками, не отраженными стандартными индексами.

Модели машинного обучения показали умеренную предсказательную силу. Random Forest (R^2 выше, чем у OLS) выделяет ключевую роль текущих потоков миграции, транспортно-жилищной инфраструктуры и инвестиций; вклад социальной инфраструктуры и экономической активности меньше. Lasso и ElasticNet обнуляют коэффициенты, что говорит о слабом линейном сигнале и высокой мультиколлинеарности. Классификатор демонстрирует умеренную точность (accuracy ≈0,60; ROC AUC ≈0,65), что указывает на значительную долю неучтенных факторов.

Совокупность результатов показывает, что инвестиции и агрегированные инфраструктурные индексы не являются надежными краткосрочными предикторами миграционного сальдо сельского населения. Исключение – жилищное строительство, которое устойчиво связано с замедлением убыли. Это ставит под сомнение «инфраструктурный детерминизм» и подчеркивает важность комплексных мер: развития инфраструктуры, экономических стимулов и точечных реформ. В России акцент должен сместиться с валовых объемов на дифференцированную политику, учитывающую кластерную специфику (аграрное ядро, ресурсная периферия, субурбанизированные и депрессивные территории – типология требует содержательного уточнения).

Заключение

Пилотное исследование показало: инвестиции не влияют на миграционное сальдо, а жилищное строительство – значимый фактор замедления убыли. Эффект сервисной и социальной инфраструктуры слабый и нелинейный, мультипликативный эффект связи с экономической активностью не подтвердился. Кластеризация выявила неоднородность сельских территорий, а пороговый эффект социальной инфраструктуры проявляется только при высокой обеспеченности. Лаг инвестиций и структурный сдвиг после 2020 г. на малой выборке не выявлены.

Модели машинного обучения подтверждают ограниченную предсказательную силу инфраструктурно-экономических факторов, определенную роль транспортно-жилищной инфраструктуры и существенный вклад «истории», демографических состояний предыдущих периодов и определенной инерции за счет прошлых событий и состояний миграционных потоков. Промежуточное исследование в рамках настоящей публикации предполагает необходимость расширения выборки, уточнения методов и их параметров для более надежной оценки демографических миграционных процессов. Так же необходимо скорректировать гипотезы с учетом нелинейности, содержательно наполнить результаты кластеризации, и на большей выборке применить другие ML-модели (CatBoost, HGBoost и аналоги) для построения классификационных выводов и регрессионных прогнозов и интерпретацией содержания моделей и прогнозов по feature importance (или другого метода в зависимости от модели) и SHAP.

Список литературы

1. *Manggat, I.* The Impact of Infrastructure Development on Rural Communities: A Literature Review / I. Manggat, R. Zain, Z. Jamaluddin // International Journal of Academic Research in Business and Social Sciences. 2018. Vol. 8, № 1. Pp. 637–648. DOI 10.6007/IJARBSS/v8-i1/3837.

2. *Hunter, L. M.* The Association Between Natural Amenities, Rural Population Growth, and Long-Term Residents' Economic Well-Being / L. M. Hunter, J. D. Boardman, J. M. Saint Onge // Rural Sociology. 2005. Vol. 70, № 4. Pp. 452–469. DOI 10.1526/003601105775012714.

3. *Mkrtchyan, N. V.* Migration in Rural Areas of Russia: Territorial Differences // Population and Economics. 2019. Vol. 3, № 1. Pp. 39–51. DOI 10.3897/popecon.3.e34780.

4. *Карцева, М. А.* Сельско-городская миграция в современной России через призму количественного и качественного анализа / М. А. Карцева, Н. В. Мкртчян, Ю. Ф. Флоринская // Крестьяноведение. 2024. Т. 9, № 2. С. 153–179. DOI 10.22394/2500-1809-2024-9-2-153-179. EDN QIKHOV.

5. *Пилипенко, И. В.* Региональные приоритеты в модернизации инженерной инфраструктуры в сельской местности для повышения качества жизни населения (часть первая) / И. В. Пилипенко, И. М. Шнейдерман // Народонаселение. 2024. Т. 27, № 1. С. 20–32. DOI 10.24412/1561-7785-2024-1-20-32. EDN MMZEPF.

6. *Пилипенко, И. В.* Региональные приоритеты в модернизации инженерной инфраструктуры в сельской местности для повышения качества жизни населения (часть вторая) / И. В. Пилипенко, И. М. Шнейдерман // Народонаселение. 2024. Т. 27, № 2. С. 26–40. DOI 10.24412/1561-7785-2024-2-26-40. EDN EOQSCH.

7. *Семенова, Е. И.* Планирование развития социальной инфраструктуры сельских территорий / Е. И. Семенова, С. Ю. Симонов, А. В. Семенов // АПК: экономика, управление. 2022. № 12. С. 84–89. DOI 10.33305/2212-84. EDN LZINEK.

8. *Нефедова, Т. Г.* Полимасштабный подход к выявлению пространственного неравенства в России как стимула и тормоза развития / Т. Г. Нефедова, А. И. Трейвиш, А. В. Шелудков // Известия Российской академии наук. Серия географическая. 2022. Т. 86, № 3. С. 289–309. DOI 10.31857/S2587556622030128. EDN FCOHMS.

9. *Hartwig, T.* Local Infrastructure, Rural Households' Resilience Capacity and Poverty: Evidence from Panel Data for Southeast Asia / T. Hartwig, T. T. Nguyen // Journal of Economics and Development. 2023. Vol. 25, № 1. Pp. 2–21. DOI 10.1108/JED-10-2022-0199.

10. *Kerimkhulle, S.* Applying a Housing Construction Model to Improve a Small Town Demographic Dynamics / S. Kerimkhulle, A. Mukhanova, M. Kantureyeva, et al. // AIP Conference Proceedings. 2023. Vol. 2700. Art. 040047. DOI 10.1063/5.0125066.

11. *Assylbayev, A.* Interdependence of the Rural Housing Stock and the Rural Population: A Scientific Perspective / A. Assylbayev, K. Niiazalieva, L. Kryzhanova, E. Ploskikh // BIO Web of Conferences. 2024. Vol. 83. Art. 07007. DOI 10.1051/bioconf/20248307007.

12. *Chapliatakaya, A.* Rural-Urban Migration within Russia: Prospects and Drivers / A. Chapliatakaya, G. Tassinari, W. J. M. Heijman, J. Van Ophem // Regional Science Policy & Practice. 2024. Vol. 16, № 9. Art. 100053. DOI 10.1016/j.rspp.2024.100053.

13. *Bezverbny, V.* The Impact of Population Density on the Socio-Economic Development of Russian Regions: From Correlation Portraits to the Cluster-Differentiated Density Governance / V. Bezverbny, T. Rostovskaya, A. Sitkovsky, S. Roslavtsev // Frontiers in Political Science. 2025. Vol. 7. P. 1715504. DOI 10.3389/fpos.2025.1715504. EDN KFFAYE.

14. *Celepcikay, O. U.* Regional Pattern Discovery in Geo-Referenced Datasets Using PCA / O. U. Celepcikay, C. F. Eick, C. Ordonez // Machine Learning and Data Mining in Pattern Recognition. MLDM 2009. Lecture Notes in Computer Science / P. Perner (ed.). Berlin : Springer, 2009. Pp. 433–447. DOI 10.1007/978-3-642-03070-3_54.

15. *Ayberkin, D.* Analysis of Turkey's Demographic and Social Structure Using PCA and K-Means Clustering / D. Ayberkin, O. Sebetci // MANAS Sosyal Araştırmalar Dergisi. 2026. Vol. 15, № 1. Pp. 146–166. DOI 10.33206/mjss.1481386.

16. *Petersen, J.* Changes to Rural Migration in the COVID-19 Pandemic / J. Petersen, R. Winkler, M. Mockrin // Rural Sociology. 2024. Vol. 89, № 1. Pp. 130–155. DOI 10.1111/ruso.12530.

17. *González-Leonardo, M.* Rural Revival? The Rise in Internal Migration to Rural Areas during the COVID-19 Pandemic. Who Moved and Where? / M. González-Leonardo, F. Rowe, A. Fresolone-Caparrós // Journal of Rural Studies. 2022. Vol. 96. Pp. 332–342. DOI 10.1016/j.jrurstud.2022.11.006.

18. *Lukić, V.* Did the COVID-19 Pandemic Change Internal Rural Migration Patterns in Serbia? / V. Lukić, S. Lović, J. Stojilković Gnjatović // Erdkunde. 2023. Vol. 77, № 3. Pp. 233–249. DOI 10.3112/erdkunde.2023.03.04.

19. *Incaltarau, C.* Exploring the Urban-Rural Dichotomies in Post-Pandemic Migration Intention: Empirical Evidence from Europe / C. Incaltarau, K. Kourtit, G. C. Pascariu // Journal of Rural Studies. 2024. Vol. 111. Pp. 1–19. DOI 10.1016/j.jrurstud.2024.103428.

20. *Belgiu, M.* Random Forest in Remote Sensing: A Review of Applications and Future Directions / M. Belgiu, L. Drăguț // ISPRS Journal of Photogrammetry and Remote Sensing. 2016. Vol. 114. Pp. 24–31. DOI 10.1016/j.isprsjprs.2016.01.011.

21. *Best, K. B.* Random Forest Analysis of Two Household Surveys Can Identify Important Predictors of Migration in Bangladesh / K. B. Best, J. M. Gilligan, H. Baroud, et al. // Journal of Computational Social Science. 2021. Vol. 4. Pp. 77–100. DOI 10.1007/s42001-020-00066-9.

22. *Deb, D.* Application of Random Forest and SHAP Tree Explainer in Exploring Spatial (In)Justice to Aid Urban Planning / D. Deb, R. M. Smith // ISPRS International Journal of Geo-Information. 2021. Vol. 10, № 9. Art. 629. DOI 10.3390/ijgi10090629.

23. *Бадыкова, И. Р.* Детерминанты валового регионального продукта в России: подход машинного обучения // Экономика региона. 2026. Т. 22, № 1. С. 97–108. DOI 10.17059/ekon.reg.2026-1-8.

24. *White, M.* Rural Growth Requires More Housing // Farmdoc Daily. 2024. Vol. 14, № 97. DOI 10.22004/ag.econ.358509.

25. *Kumagai, J.* Impacts of Urban – Rural Migration on Domain-Specific Life Satisfaction / Kumagai J., Yoo S., Managi S., et al. // Cities. 2025. Vol. 139. Art. 106056. DOI 10.1016/j.cities.2025.106056.

26. *Loras-Gimeno, D.* Rural Depopulation in the 21st Century: A Systematic Review of Policy Assessments / D. Loras-Gimeno, J. Díaz-Lanchas, G. Gómez-Bengoechea // Regional Science Policy & Practice. 2025. Vol. 17, № 5. Art. 100176. DOI 10.1016/j.rspp.2025.100176.

Сведения об авторе

Дождиков Антон Валентинович, *кандидат политических наук, старший научный сотрудник, Институт социально-политических исследований ФНИСЦ РАН, Москва, Россия.*

Контактная информация: *e-mail: antondnn@yandex.ru; ORCID ID: 0000-0002-1069-1648; РИНЦ SPIN-код 2208-1891; Web of Science Researcher ID: KYP-9166-2024; Scopus Author ID: 57221684847.*

Статья поступила в редакцию 14.03.2026; принята в печать 15.06.2026.
Автор прочитал и одобрил окончательный вариант рукописи.

FACTORS OF DEMOGRAPHIC WELL-BEING IN RUSSIAN VILLAGES: PRELIMINARY HYPOTHESES FOR ML MODELING

Anton V. Dozhdikov
Institute of Socio-Political Research FCTAS RAS, Moscow, Russia
E-mail: antondnn@yandex.ru

For citation: Dozhdikov, A. V. Factors of Demographic Well-Being in Russian Villages: Preliminary Hypotheses for ML Modeling. *DEMIS. Demographic Research.* 2026. Vol. 6, No. 2. Pp. 240–262. DOI 10.19181/demis.2026.6.2.14. (In Russ.)

Abstract. *The study is devoted to formulating preliminary hypotheses and selecting machine learning methods to quantitatively analyze the demographic and migratory dynamics of rural populations in Russian regions, and their connection with infrastructure and economic characteristics The relationships between migration, natural population movement, and the level of infrastructure, investment, housing construction, access to services, and economic activity are all considered. The goal is to create a model that identifies stable relationships between factors and a cluster typology for regions, as well as a preliminary test of primary hypotheses using machine learning techniques. A limited set of data from Rosstat is used, which is aggregated into a regional-year panel after cleaning, normalizing and constructing synthetic indices. The methods used are EDA, PCA+K-means, regression, decision trees, and Random Forest. To continue the research and further adjust the hypotheses, we found that: there is no connection yet identified between investment and an increase in net migration, but there are clear signs of a relationship between housing construction and slowing population decline. The effects of infrastructure were limited, and it takes longer to identify relationships between managerial decisions and outcomes, as well as cluster heterogeneity across regions, and the predictive power of models for migration is limited. Limitations of the study include a short panel, data aggregation, and shocks in 2020-2021. The study has prospects for testing hypotheses using methods applied to a large region-year dataset.*

Keywords: *rural population, internal migration, demographic processes, rural areas, infrastructure, housing construction, social infrastructure, panel data, cluster analysis, machine learning*

References

1. Manggat, I., Zain, R., Jamaluddin, Z. The Impact of Infrastructure Development on Rural Communities: A Literature Review. *International Journal of Academic Research in Business and Social Sciences.* 2018. Vol. 8, No. 1. Pp. 637–648. DOI 10.6007/IJARBSS/v8-i1/3837.

2. Hunter, L. M., Boardman, J. D., Saint Onge, J. M. The Association between Natural Amenities, Rural Population Growth, and Long-Term Residents' Economic Well-Being. *Rural Sociology.* 2005. Vol. 70, No. 4. Pp. 452–469. DOI 10.1526/003601105775012714.

3. Mkrtchyan, N. V. Migration in Rural Areas of Russia: Territorial Differences. *Population and Economics.* 2019. Vol. 3, No. 1. Pp. 39–51. DOI 10.3897/popecon.3.e34780.

4. Kartseva, M. A., Mkrtchyan, N. V., Florinskaya, Yu. F. Rural-Urban Migration in Contemporary Russia Through the Prism of Quantitative and Qualitative Analysis. *Russian Peasant Studies.* 2024. Vol. 9, No. 2. Pp. 153–179. DOI 10.22394/2500-1809-2024-9-2-153-179. (In Russ.).

5. Pilipenko, I. V., Shneiderman, I. M. Regional Priorities in the Modernization of Engineering Infrastructure in Rural Areas to Improve the Quality of Life (Part One). *Population.* 2024. No. 1. Pp. 20–32. DOI 10.24412/1561-7785-2024-1-20-32. (In Russ.).

6. Pilipenko, I. V., Shneiderman, I. M. Regional Priorities in the Modernization of Engineering Infrastructure in Rural Areas to Improve the Quality of Life (Part Two). *Population.* 2024. No. 2. Pp. 26–40. DOI 10.24412/1561-7785-2024-2-26-40. (In Russ.).

7. Semenova, E. I., Simonov, S. A., Semenov, A. V. Planning the Development of Social Infrastructure in Rural Areas. *AIC: Economics, Management.* 2022. No. 12. Pp. 84–89. DOI 10.33305/2212-84. (In Russ.).

8. Nefedova, T. G., Treivish, A. I., Sheludkov, A. V. A Multi-Scale Approach to Identifying Spatial Inequality in Russia as a Driver and Barrier to Development. *Izvestiya RAN (Akad. Nauk SSSR). Seriya Geograficheskaya.* 2022. Vol. 86, No. 3. Pp. 289–309. DOI 10.31857/S2587556622030128. (In Russ.).

9. Hartwig, T., Nguyen, T. T. Local Infrastructure, Rural Households' Resilience Capacity and Poverty: Evidence from Panel Data for Southeast Asia. *Journal of Economics and Development.* 2023. Vol. 25, No. 1. Pp. 2–21. DOI 10.1108/JED-10-2022-0199.

10. Kerimkhulle, S., Mukhanova, A., Kantureyeva, M., Koishybaeva, M., Azieva, G. Applying a Housing Construction Model to Improve a Small Town Demographic Dynamics. *AIP Conference Proceedings.* 2023. Vol. 2700. Art. 040047. DOI 10.1063/5.0125066.

11. Assylbayev, A., Niiazalieva, K., Kryzhanova, L., Ploskikh, E. Interdependence of the Rural Housing Stock and the Rural Population: A Scientific Perspective. *BIO Web of Conferences.* 2024. Vol. 83. Art. 07007. DOI 10.1051/bioconf/20248307007.

12. Chapliatakaya, A., Tassinari, G., Heijman, W. J. M., Van Ophem, J. Rural-Urban Migration within Russia: Prospects and Drivers. *Regional Science Policy & Practice.* 2024. Vol. 16, No. 9. Art. 100053. DOI 10.1016/j.rspp.2024.100053.

13. Bezverbny, V. A., Rostovskaya, T. K., Sitkovsky, A. S., Roslavtsev, S. N. The Impact of Population Density on the Socio-Economic Development of Russian Regions: From Correlation Portraits to the Cluster-Differentiated Density Governance. *Frontiers in Political Science.* 2025. Vol. 7. P. 1715504. DOI 10.3389/fpos.2025.1715504.

14. Celepcikay, O. U., Eick, C. F., Ordonez, C. Regional Pattern Discovery in Geo-Referenced Datasets Using PCA. In: *Machine Learning and Data Mining in Pattern Recognition.* MLDM 2009. Lecture Notes in Computer Science, Vol. 5632. P. Perner (ed.). Berlin: Springer, 2009. Pp. 433–447. DOI 10.1007/978-3-642-03070-3_54.

15. Ayberkin, D., Sebetci, Ö. Analysis of Turkey's Demographic and Social Structure Using PCA and K-Means Clustering. *MANAS Journal of Social Studies.* 2026. Vol. 15, No. 1. Pp. 146–166. DOI 10.33206/mjss.1481386.

16. Petersen, J., Winkler, R., Mockrin, M. Changes to Rural Migration in the COVID-19 Pandemic. *Rural Sociology.* 2024. Vol. 89, No. 1. Pp. 130–155. DOI 10.1111/ruso.12530.

17. González-Leonardo, M., Rowe, F., Fresolone-Caparrós, A. Rural revival? The Rise in Internal Migration to Rural Areas during the COVID-19 Pandemic. Who Moved and Where? *Journal of Rural Studies.* 2022. Vol. 96. Pp. 332–342. DOI 10.1016/j.jrurstud.2022.11.006.

18. Lukić, V., Lović, S., Stojilković Gnjatović, J. Did the COVID-19 Pandemic Change Internal Rural Migration Patterns in Serbia? *Erdkunde.* 2023. Vol. 77, No. 3. Pp. 233–249. DOI 10.3112/erdkunde.2023.03.04.

19. Incaltarau, C., Kourtit, K., Pascariu, G. C. Exploring the Urban-Rural Dichotomies in Post-Pandemic Migration Intention: Empirical Evidence from Europe. *Journal of Rural Studies.* 2024. Vol. 111. Pp. 1–19. DOI 10.1016/j.jrurstud.2024.103428.

20. Belgiu, M., Drăguț, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing.* 2016. Vol. 114. Pp. 24–31. DOI 10.1016/j.isprsjprs.2016.01.011.

21. Best, K. B., Gilligan, J. M., Baroud, H. et al. Random Forest Analysis of Two Household Surveys Can Identify Important Predictors of Migration in Bangladesh. *Journal of Computational Social Science.* 2021. Vol. 4. Pp. 77–100. DOI 10.1007/s42001-020-00066-9.

22. Deb, D., Smith, R. M. Application of Random Forest and SHAP Tree Explainer in Exploring Spatial (In)Justice to Aid Urban Planning. *ISPRS International Journal of Geo-Information.* 2021. Vol. 10, No. 9. Art. 629. DOI 10.3390/ijgi10090629.

23. Badykova, I. R. Determinants of Gross Regional Product in Russia: A Machine Learning Approach. *Economy of Regions.* 2026. Vol. 22, No. 1. Pp. 97–108. DOI 10.17059/ekon.reg.2026-1-8.

24. White, M. Rural Growth Requires More Housing. *Farmdoc Daily.* 2024. Vol. 14, № 97. DOI 10.22004/ag.econ.358509.

25. J. Kumagai, S. Yoo, S. Managi, et al. Impacts of Urban – Rural Migration on Domain-Specific Life Satisfaction. *Cities.* 2025. Vol. 139. Art. 106056. DOI 10.1016/j.cities.2025.106056.

26. Loras-Gimeno, D., Díaz-Lanchas, J., Gómez-Bengoechea, G. Rural Depopulation in the 21st Century: A Systematic Review of Policy Assessments. *Regional Science Policy & Practice.* 2025. Vol. 17, No. 5. Art. 100176. DOI 10.1016/j.rspp.2025.100176.

Bio note

Anton V. Dozhdikov, *Candidate of Political Sciences, Senior Researcher, Institute of Socio-Political Research FCTAS RAS, Moscow, Russia.*

Contact information: *e-mail: antondnn@yandex.ru; ORCID ID: 0000-0002-1069-1648; RSCI SPIN-code 2208-1891; Web of Science Researcher ID: KYP-9166-2024; Scopus Author ID: 57221684847.*

Received on 14.03.2026; accepted for publication on 15.06.2026.
The author has read and approved the final manuscript.