

Черемухин Артем Дмитриевич

доцент кафедры «Математика и вычислительная техника»

Cheremuhin A.D.

E-mail: ngieu.cheremuhin@yandex.ru

Пахрин Андрей Алексеевич

студент 4-го курса, Институт информационных
технологий и систем связи

Pakhrin A.A.

E-mail: andrey.pakhrin@gmail.com

Нижегородский государственный инженерно-
экономический университет

Nizhny Novgorod State University of
Engineering and Economics

Нижегородская обл., г. Княгинино,
ул. Октябрьская, д. 22А, Россия, 606340

Тел: +7 (83166) 4-15-50

К ВОПРОСУ ПРОГНОЗИРОВАНИЯ ПРОСТРАНСТВЕННО-ВРЕМЕННОЙ ДИНАМИКИ ПРИЁМА НА ОБУЧЕНИЕ ПО ПРОГРАММАМ ВЫСШЕГО ОБРАЗОВАНИЯ В РОССИЙСКОЙ ФЕДЕРАЦИИ

ON FORECASTING THE SPATIOTEMPORAL DYNAMICS OF ADMISSION TO HIGHER EDUCATION PROGRAMMES IN THE RUSSIAN FEDERATION

Аннотация

Введение: исследование посвящено прогнозированию пространственно-временной динамики приёма на обучение по программам высшего образования в субъектах Российской Федерации с учётом предшествующих состояний региональных образовательных систем и пространственных взаимосвязей между территориями.

Материалы и методы: информационная база сформирована на основе данных федерального статистического наблюдения по общему, среднему профессиональному и высшему образованию. Итоговый массив включает 332 наблюдения по 83 субъектам Российской Федерации за 2021–2024 гг. и 43 переменные. Для прогнозирования использованы линейная и регуляризованные регрессии, PLS, PCA, случайный лес и XGBoost; пространственная зависимость анализировалась с помощью индекса Морана, SAR и SEM. Качество прогнозов оценивалось по MAE, RMSE, sMAPE и коэффициента детерминации.

Результаты: На тестовой выборке за 2024 г. минимальная MAE получена для настроенного XGBoost — 821 человек, минимальный RMSE и максимальный R^2 — для OLS: 1210 человек и 0,992 соответственно, минимальная sMAPE — для базового XGBoost: 14,0 %. Выявлена статистически значимая положительная пространственная

автокорреляция. В XGBoost 83,64 % суммарной важности приходится на среднюю численность студентов высшего образования в прошлом и позапрошлом годах.

Обсуждение: результаты свидетельствуют о выраженной временной инерционности процесса приёма и наличии пространственной зависимости между субъектами Российской Федерации; различные модели имеют преимущества по разным критериям качества.

Заключение: дальнейшая работа должна быть направлена на объединение временной и пространственной составляющих в единой прогнозной модели, расширение временного горизонта и использование более детализированных данных по образовательным организациям и направлениям подготовки.

Ключевые слова: временная инерционность, временные лаги, высшее образование, градиентный бустинг, индекс Морана, машинное обучение, образовательные потоки, приём на обучение, прогнозирование, пространственная автокорреляция, пространственные модели, региональная образовательная система, регрессионные модели.

Annotation

Introduction: The study is devoted to forecasting the spatiotemporal dynamics of admission to higher education programmes across the constituent entities of the Russian Federation, taking into account the preceding states of regional education systems and spatial interrelationships between territories.

Materials and Methods: the information base was formed using federal statistical observation data on general, secondary vocational, and higher education. The final dataset includes 332 observations for 83 constituent entities of the Russian Federation for 2021–2024 and 43 variables. Linear and regularized regression models, PLS, PCA, random forest, and XGBoost were used for forecasting; spatial dependence was analyzed using Moran's I, SAR, and SEM. Forecast quality was assessed using MAE, RMSE, sMAPE, and the coefficient of determination.

Results: on the 2024 test sample, the lowest MAE was obtained for tuned XGBoost and amounted to 821 students. The lowest RMSE and the highest coefficient of determination were obtained for OLS: 1,210 students and 0.992, respectively. The lowest sMAPE, 14.0%, was obtained for the baseline XGBoost model. A statistically significant positive spatial autocorrelation was identified. In XGBoost, 83.64% of the total feature importance was accounted for by the average number of higher education students in the previous and the year before the previous year.

Discussion: the results indicate pronounced temporal inertia in the admission process and the presence of spatial dependence between the constituent entities of the Russian Federation; different models demonstrate advantages according to different quality criteria.

Conclusion: further research should focus on integrating temporal and spatial components within a single forecasting model, extending the observation horizon, and using more detailed data on individual higher education institutions and fields of study.

Keywords: temporal inertia, time lags, higher education, gradient boosting, Moran's I, machine learning, educational flows, admission, forecasting, spatial autocorrelation, spatial models, regional education system, regression models.

Введение

Система высшего образования Российской Федерации функционирует в условиях выраженной территориальной неоднородности, демографических изменений и различий в масштабах региональных образовательных потоков. Численность принятых на обучение по программам высшего образования непосредственно связана с планированием контрольных цифр приёма, распределением бюджетных мест, развитием образовательных программ, кадровым и финансовым обеспечением образовательных организаций. В этих условиях возрастает значение прогнозных оценок, позволяющих учитывать ожидаемые изменения при принятии решений в сфере управления образованием.

Особенность данной задачи состоит в том, что динамика приёма существенно различается между субъектами Российской Федерации и определяется не только изменением собственных характеристик региональных образовательных систем во времени, но и их территориальными взаимосвязями. В результате изменение показателей приёма в отдельном субъекте следует рассматривать не изолированно, а в контексте общей пространственно-временной структуры образовательных процессов. Наличие таких взаимосвязей усложняет прогнозирование, поскольку требует одновременного учёта как предшествующих состояний образовательной системы, так и пространственной неоднородности, и взаимозависимости региональных показателей. Необходимость такого подхода следует и из исходной постановки исследования, в которой отмечаются региональные различия динамики приёма и возможное влияние пространственных факторов.

Тем самым актуальность исследования определяется необходимостью повышения обоснованности прогнозов динамики приёма на обучение по программам высшего образования, используемых при планировании развития системы высшего образования и распределении образовательных ресурсов, а научная проблема связана с необходимостью адекватного отражения в прогнозных моделях одновременно временной и пространственной структуры исследуемого процесса.

Современные исследования, посвящённые прогнозированию параметров приёма и численности обучающихся в системе высшего образования, проходят в двух направлениях: в направлении классического анализа временных рядов и с использованием методов машинного обучения. Одним из ранних примеров применения нелинейных методов является работа [1], в которой прогнозировалось число студентов филиала Университета Пайям-Нур, выбирающих дисциплину «Продвинутое программирование» в следующем семестре. В качестве исходных данных использовались временные ряды за 20 последовательных семестров, включавшие численность первокурсников, студентов базового и продвинутого уровней. Сопоставлялись различные архитектуры искусственных нейронных сетей — сети прямого распространения, сети Элмана и RBF-сети — и алгоритмы их обучения. Качество оценивалось по MAE, MSE и MSEREG с многократным повторением вычислительных экспериментов. Наилучшие результаты были получены для сетей архитектуры 20-п-1, обучаемых методом Левенберга–Марквардта; при этом для RBF-сетей высокая точность интерполяции обучающих значений сочеталась с относительной ошибкой прогноза около 33 %, что указывает на существенную роль

способа проверки прогностических свойств модели, а не только качества описания известных наблюдений [1].

Классический подход к моделированию временной динамики применён в работе [2] при прогнозировании численности иностранных студентов бакалавриата одного из университетов США. Использовались семестровые наблюдения с осени 1999 г. по весну 2014 г. и сезонная модель SARIMA, дополненная внешними факторами: стоимостью обучения и индикаторами изменений визовой политики и структуры международного студенческого потока. Включение дополнительных факторов позволило уменьшить AICс с 737,17 до 711,41. Для последующих семестров формировались не только точечные, но и интервальные прогнозы; относительное отклонение фактических значений от прогнозных для отдельных контрольных периодов составило от 4,86 до 12,68 %.

Наряду с моделированием итоговой численности обучающихся прогнозирование применяется непосредственно в ходе приёмной кампании. В работе [3] рассматривается последовательность прохождения абитуриентами основных стадий приёма — от подачи заявления и принятия решения о зачислении до подтверждения намерения обучаться и фактического начала обучения. Методика сочетает скользящие средние, сглаживание, методы Бокса–Дженкинса, регрессионные и логистические модели с расчётом коэффициентов перехода между отдельными стадиями приёмной воронки. В частности, используется показатель доли потерь между подтверждением намерения обучаться и фактическим началом обучения. Прогноз последовательно уточняется на недельных данных: оценка числа заявлений используется для расчёта ожидаемого числа принятых, после чего с учётом доли перехода на следующую стадию определяется ожидаемая численность фактически приступивших к обучению/

В современных исследованиях заметное место занимает сравнительный анализ статистических методов и методов машинного обучения. В работе [4] сопоставлены ARIMA и LSTM при прогнозировании валового коэффициента охвата высшим образованием в штате Мизорам. Показатель формировался на основе официальной образовательной статистики, а точность моделей оценивалась, в частности, по RMSE. Модель LSTM превзошла ARIMA по RMSE в среднем на 0,1322 % и по использованному авторами показателю точности на 5,6 %; прогнозируемое значение валового коэффициента охвата на 2035 г. составило 34,23 %. В работе [5] для прогнозирования числа вновь принимаемых студентов по образовательным программам Государственного политехнического института Бали применён метод опорных векторов (SVM). В качестве основной меры ошибки использовался RMSE; минимальное значение 7,292 получено для направления электротехники.

Более развернутое сопоставление моделей разных классов проведено в работе [6]. На исторических данных о приёме за 2018–2023 гг. сравнивались простое скользящее среднее, ARIMA и случайный лес для нескольких образовательных программ. Оценивание выполнялось по MAPE, MAD и MSE. Для случайного леса значения MAPE находились в диапазоне 6,17–17,95 %, для ARIMA — 17,95–33,13 %, для скользящего среднего — 14,5–28,25 %. Иной результат получен в работе [7] на годовом ряду численности иностранных студентов одного американского университета за 2000–2022 гг. При разделении данных на обучающую часть 2000–2017

гг. и тестовую 2018–2019 гг. сравнивались ARIMA, Prophet и LSTM; ARIMA(2,1,0) обеспечила MAPE 1,26 %, Prophet — 11,81 %, LSTM — 13,84 %. Полученное соотношение связывается авторами с небольшой длиной ряда и устойчивостью его основной динамики.

В работе [8] использован значительно более длинный временной ряд — 75 лет наблюдений. После разделения данных в отношении 80/20 сопоставлялись случайный лес, ARIMA(2,1,1) и гибридная модель ARIMA–случайный лес. Значения MAPE составили соответственно 15,62, 16,95 и 14,52 %, то есть объединение статистической модели и ансамблевого алгоритма обеспечило наименьшую ошибку среди рассматриваемых вариантов.

Отдельную группу образуют исследования, в которых динамика высшего образования рассматривается на уровне более крупных образовательных систем. В работе [9] анализировалось развитие системы высшего образования Тайваня на временном интервале 1950–2021 гг. В массив включались темпы экономического роста, ВВП на душу населения, численность обучающихся, валовой коэффициент охвата высшим образованием, а также входящие и исходящие международные студенческие потоки. Для сопоставимости абсолютные показатели преобразовывались в индексы и темпы изменения, взаимные временные связи изучались с помощью функции взаимной корреляции, а последующая динамика — посредством ARIMA. Полученные результаты позволили установить взаимосвязи между расширением системы высшего образования и изменениями интенсивности студенческой мобильности. В работе [10] ARIMA использовалась для анализа данных одной организации высшего образования за 2022–2025 гг. и построения прогноза на 2026–2030 гг.; в исходном ряду фиксировалось снижение численности обучающихся в 2023–2024 гг. с последующим восстановлением в 2025 г.

Значительное направление исследований связано с пространственной организацией спроса на высшее образование и студенческих потоков. В работе [11] выбор региона обучения выпускниками школ Нидерландов описывался модифицированной гравитационной моделью, объединяющей характеристики региона происхождения, региона обучения, расстояние между ними и показатели территориальной привлекательности. Расчёты выявили отрицательную связь интенсивности студенческого потока с расстоянием и стоимостью жилья и положительную — с характеристиками городской и региональной инфраструктуры; вклад последних оказался более выраженным, чем влияние ряда показателей качества образования.

На обширном германском материале пространственная организация образовательных потоков исследована в работе [12]. Исходный массив охватывал период 1999–2015 гг., 814 организаций высшего образования и 402 района происхождения студентов; после геокодирования было сформировано 327 228 пар «организация высшего образования — район происхождения». Единицей анализа являлась интенсивность потока студентов из района происхождения в конкретную образовательную организацию. В связи с преобладанием нулевых значений использована пуассоновская модель с избыточным числом нулей, состоящая из бинарной части, характеризующей наличие потока, и счётной части, описывающей его

интенсивность. Модели оценивались отдельно для периодов до, во время и после Болонской реформы, а также для различных направлений подготовки. Полученные результаты свидетельствуют о сравнительно высокой устойчивости пространственного распределения студентов после перехода к двухуровневой системе; характер изменений различался по направлениям подготовки и типам образовательных организаций.

Сетевой подход к пространственным образовательным потокам применён в работе [13] при исследовании мобильности в Европейском пространстве высшего образования. На основе данных Erasmus+ и ЮНЕСКО формировались сети перемещений студентов, а для выделения групп территорий с повышенной интенсивностью внутренних связей использовался показатель модульности. Анализ выявил выраженные субрегиональные структуры и неодинаковую интенсивность направлений мобильности внутри европейского образовательного пространства.

Сочетание пространственного моделирования и методов машинного обучения представлено в работе [14]. Прогнозировались перемещения выпускников школ Германии к местам получения высшего образования на уровне районов; в модель XGBoost были включены показатели гравитационной и радиационной моделей и около 28 региональных факторов. Настройка параметров η , γ , максимальной глубины дерева и минимального веса дочернего узла проводилась по сетке с перекрёстной проверкой; при построении каждого дерева использовалось 70 % признаков. Качество оценивалось индексом Сёренсена, MAE, RMSE, коэффициентом корреляции Пирсона и скорректированным R^2 , а вклад факторов анализировался с помощью SHAP. В наиболее полной спецификации индекс Сёренсена составил 0,75, MAE — 3,26, RMSE — 75,43, скорректированный R^2 — 0,80, коэффициент корреляции Пирсона — 0,87. Наиболее значимым фактором оказалась численность первокурсников в месте назначения; существенное влияние также имели время поездки и число альтернативных образовательных возможностей в окружающем пространстве. Для крупных образовательных центров обнаружен положительный нелинейный эффект существующего масштаба приёма.

Другой класс пространственных методов связан с оцениванием территориальной зависимости образовательных показателей. В работе [15] исследовалась численность мужчин и женщин, обучающихся по STEM-направлениям, во всех 81 провинциях Турции. Методика сочетала OLS с глобальными и локальными процедурами пространственного анализа. В качестве факторов использовались социальные и экономические характеристики территорий. Более высокие значения исследуемых показателей были характерны для прибрежных и промышленно развитых провинций и положительно связаны, в частности, с ВРП на душу населения, численностью преподавателей высшей школы, бюджетными расходами и численностью населения региона.

В работе [16] географически взвешенная регрессия (GWR) применена к данным 34 провинций Индонезии за 2023 г. В качестве объясняющих переменных рассматривались региональный продукт, финансирование образования, отношение численности преподавателей к численности студентов, число студентов, расходы на душу населения, численность населения и индекс глубины бедности. Перед

построением модели глобальная пространственная зависимость оценивалась индексом Морана, а пространственная неоднородность — тестом Бройша–Пагана. Пространственные веса формировались с использованием метода ближайших соседей и гауссова ядра, оптимальная ширина полосы определялась посредством перекрёстной проверки. Модель GWR объяснила 50,57 % вариации валового коэффициента охвата высшим образованием; локальные коэффициенты показали различающийся по провинциям вклад факторов, а на основании комбинаций значимых переменных были выделены пять территориальных групп.

В российской научной литературе пространственная организация системы высшего образования рассматривается прежде всего через межрегиональную образовательную миграцию. В работе [17] исследованы данные о фактическом приёме и перемещениях абитуриентов между субъектами Российской Федерации с использованием официальной статистики и данных о результатах ЕГЭ. На основе соотношения входящих и исходящих образовательных потоков проводилась типологизация территорий по характеру миграционного обмена, что позволило выделить регионы-доноры, регионы-реципиенты и территории с иными вариантами миграционного поведения.

В работе [18] информационная база сформирована на основе форм ВПО-1, ОО-1 и других источников официальной образовательной статистики. Для выявления структуры взаимосвязей региональных показателей применялся факторный анализ, после чего полученные характеристики использовались при группировке субъектов Российской Федерации и исследовании территориальной структуры образовательной миграции. Были выделены группы регионов, различающиеся масштабом и направленностью образовательных потоков, а также крупные территориальные центры притяжения обучающихся.

Территориальная динамика образовательных потоков рассматривается и в работе [19], где сопоставляется положение субъектов Российской Федерации в 2017, 2020 и 2023 гг. Анализ показал изменение состава регионов с различными типами миграционного баланса: в 2017 г. выделялось 12 территорий-акцепторов и 59 доноров, в 2020 г. — 12 акцепторов и 30 доноров при увеличении числа территорий с относительно сбалансированным обменом, а в 2023 г. число акцепторов сохранилось на уровне 12 при изменении состава остальных групп. В работе [20] территориальная структура дополнительно исследуется посредством отдельного расчёта миграционного сальдо для возрастных групп 17–18 и 20–23 года и коэффициента возврата. На этой основе выделены регионы-доноры, акцепторы, транзитные, балансные и замкнутые территории.

Обзор показывает, что результативность прогнозирования показателей приёма определяется тем, насколько модель соответствует структуре конкретных данных. В одних исследованиях преимущество получают методы машинного обучения, в других — классические модели временных рядов, а использование гибридных схем даёт дополнительный эффект лишь при наличии достаточно продолжительной истории наблюдений. Одновременно пространственные исследования устойчиво фиксируют зависимость образовательных потоков от положения территории в системе высшего образования, расстояния, масштаба уже сложившегося приёма, доступности

образовательных центров и других региональных характеристик. В совокупности это позволяет рассматривать динамику приёма не только как последовательность изменений во времени, но и как процесс, формирующийся в системе пространственно связанных территорий.

Однако, в существующих работах временная и пространственная стороны процесса в основном получают самостоятельное развитие: прогнозные модели строятся преимущественно для отдельных университетов, образовательных программ или агрегированных национальных показателей, тогда как исследования территориальной структуры сосредоточены на мобильности студентов, привлекательности регионов, классификации территорий и оценке пространственной неоднородности. Российские исследования подробно описывают межрегиональные образовательные потоки и различия между субъектами Российской Федерации, однако в рассмотренном корпусе не выявлены работы, в которых для страны в целом решалась бы задача количественного прогнозирования численности принятых на обучение по программам высшего образования по субъектам Российской Федерации с одновременным учётом предшествующей динамики образовательных показателей и пространственных взаимосвязей между регионами.

В связи с этим цель исследования состоит в оценке возможностей прогнозирования пространственно-временной динамики приёма на обучение по программам высшего образования в Российской Федерации на основе адаптации комплекса методов анализа данных к региональной образовательной статистике, учитывающего временную инерционность показателей и пространственные взаимосвязи между субъектами Российской Федерации.

Материалы и методы

Для исследования были объединены данные шести форм федерального статистического наблюдения, характеризующих общее, среднее профессиональное и высшее образование. Из форм ВПО-1 и ВПО-2 извлекались показатели высшего образования, из ОО-1 и ОО-2 — показатели общего образования, из СПО-1 и СПО-2 — показатели среднего профессионального образования. Состав использованных показателей представлен в таблице 1.

Таблица 1 – Источники и состав показателей информационной базы исследования

Источник	Извлекаемые показатели	Формирование признаков
ВПО-1	Число принятых на обучение; среднегодовая численность студентов; численность профессорско-преподавательского состава	Число принятых определяется суммированием значений по очной, заочной и очно-заочной формам обучения; рассчитывается число студентов на одного преподавателя; для среднегодовой численности студентов
ВПО-2	Число персональных компьютеров; площадь общежитий; расходы на стипендии	Рассчитываются число компьютеров на одного студента, площадь общежитий на одного студента и расходы на стипендии на одного студента
ОО-1	Число выпускников 9-х и 11-х классов; число продолживших обучение после 9-го класса	Формируются показатели численности выпускников 9-х и 11-х классов за прошлый и позапрошлый периоды; число выпускников 9-

		х классов корректируется с учётом продолживших обучение в школе
ОО-2	Число персональных компьютеров; численность учащихся; площадь школьных помещений; расходы образовательных организаций; число школ с доступом к Интернету не менее 100 Мбит/с	Рассчитываются число компьютеров на одного учащегося, площадь на одного учащегося, расходы на одного учащегося и доля школ с высокоскоростным доступом к Интернету
СПО-1	Число принятых на программы СПО после 9-го класса; число принятых на программы СПО после 11-го класса	Показатели формируются суммированием принятых на бюджетной и платной основе по программам подготовки специалистов среднего звена и квалифицированных рабочих и служащих
СПО-2	Среднегодовая численность студентов СПО; число компьютеров; расходы на стипендии; иная материальная поддержка; площадь общежитий	Рассчитываются число компьютеров на одного студента, расходы на стипендии и иную материальную поддержку на одного студента, площадь общежитий на одного студента

После извлечения исходных показателей данные приводились к единой единице наблюдения «субъект Российской Федерации — год». Абсолютные показатели использовались для характеристики масштаба соответствующих образовательных потоков, а ресурсные характеристики нормировались на численность обучающихся. В частности, рассчитывались показатели обеспеченности персональными компьютерами, площади общежитий и объёма стипендиальных расходов в расчёте на одного студента или учащегося.

Дополнительно на основе первичных данных формировались показатели, характеризующие переходы между уровнями образовательной системы. Рассчитывалось отношение численности выпускников 11-х классов к численности выпускников 9-х классов, а также отношение числа поступивших в организации СПО после 9-го класса к численности соответствующей выпускной когорты. Тем самым в информационную базу включались не только характеристики отдельных уровней образования, но и относительные показатели движения обучающихся между ними.

Поскольку прогноз строился для года t , для большинства характеристик образовательной системы в качестве отдельных признаков брались значения на прошлый и позапрошлый год. При моделировании использовалась только информация, относящаяся к периодам, предшествующим прогнозируемому году, что исключало попадание будущих значений в состав предикторов. Для формирования двухлетних лагов дополнительно привлекались исходные данные за 2017–2018 гг.

В результате объединения данных общего, среднего профессионального и высшего образования был сформирован аналитический массив из 332 наблюдений по 83 субъектам Российской Федерации за 2021–2024 гг. и 43 переменных. Наблюдения за 2021–2023 гг. использовались при построении моделей, а данные за 2024 г. — 83 наблюдения — были выделены в независимую тестовую выборку.

Вычислительный эксперимент будет построен как последовательное сравнение нескольких классов моделей, различающихся по степени интерпретируемости, способу учёта мультиколлинеарности, способности выявлять нелинейные зависимости

и возможности отражать пространственную структуру данных. В анализ включаются следующие модели:

- линейная регрессия OLS — в качестве базовой интерпретируемой модели;
- Ridge-регрессия — для стабилизации оценок при высокой взаимной корреляции признаков;
- Lasso-регрессия — как регуляризованная модель со встроенным отбором признаков;
- Elastic Net — как сочетание двух типов регуляризации;
- случайный лес — для выявления нелинейных зависимостей и взаимодействий между признаками;
- XGBoost — как градиентный ансамблевый метод для моделирования более сложной структуры зависимостей;
- PLS-регрессия — для построения компонент с учётом их связи с прогнозируемым показателем;
- PCA + OLS и PCA + Ridge — для оценки эффективности предварительного снижения размерности признакового пространства;
- OLS на признаках, отобранных Lasso, — как двухэтапная схема отбора признаков и последующего линейного моделирования;
- SAR и SEM — для анализа пространственной зависимости соответственно через пространственный лаг зависимой величины и пространственную автокорреляцию ошибок.

Для моделей с настраиваемыми гиперпараметрами их значения будут определяться только по обучающей части данных. В регуляризованных моделях будет варьироваться сила штрафа, в PLS и PCA — число компонент, в случайном лесе и XGBoost — параметры, определяющие структуру и сложность ансамбля. Независимая тестовая выборка на этапе настройки использоваться не будет.

Валидация будет организована с учётом временного порядка наблюдений. Вместо случайного перемешивания будет использоваться схема *expanding origin*: на каждом шаге модель обучается на доступных предшествующих периодах и проверяется на следующем временном срезе. Такой порядок соответствует реальному сценарию прогноза и исключает попадание информации из будущего периода в обучающие данные. После завершения настройки итоговое качество всех моделей будет оцениваться на полностью отложенной выборке за 2024 г.

Пространственный анализ будет выполнен отдельно: сначала для целевого показателя будет рассчитан индекс Морана, позволяющий проверить наличие глобальной пространственной автокорреляции. При её наличии будут оценены модели SAR и SEM. Их сопоставление будет проводиться с использованием LM-тестов, информационных критериев и проверки пространственной автокорреляции остатков.

Для сопоставления качества прогнозных моделей будет использован единый набор обобщённых метрик, позволяющий оценивать, как абсолютную, так и относительную величину ошибки:

- MAE — средняя абсолютная ошибка прогноза в натуральных единицах; характеризует типичное отклонение прогнозируемого значения от фактического;

- RMSE — среднеквадратичная ошибка, более чувствительная к крупным отклонениям и поэтому позволяющая учитывать влияние значительных ошибок в регионах с большой численностью принятых;
- sMAPE — симметричная средняя абсолютная процентная ошибка, используемая для сопоставления относительной точности прогнозов при различном масштабе региональных показателей;
- R^2 — коэффициент детерминации, характеризующий долю вариации прогнозируемого показателя, объясняемую моделью.

Результаты

После расчетов были получены следующие настройки некоторых моделей:

- Elastic Net была реализована при $\alpha = 0,5$;
- Случайный лес включал 1200 деревьев; минимальный размер конечного узла составлял 5 наблюдений, а число признаков, рассматриваемых при разбиении, задавалось приблизительно как $\sqrt{p}$. Для оценки вклада признаков использовалась перестановочная важность;
- Базовый XGBoost был построен с квадратичной функцией потерь, метрикой RMSE, скоростью обучения $\eta = 0,05$, максимальной глубиной дерева 6, долей наблюдений subsample = 0.8 и долей признаков colsample_bytree = 0.8. Максимальное число итераций составляло 5000, применялась ранняя остановка после 50 итераций без улучшения;
- Настроенный XGBoost был получен после перебора гиперпараметров. В итоговой спецификации использовались $\eta = 0,07$, max_depth = 6, min_child_weight = 1, subsample = 0.9 и colsample_bytree = 0.7;
- PLS-регрессия строилась с подбором числа компонент;
- Для PCA + OLS использовалось число компонент, обеспечивающее сохранение 95 % дисперсии; в итоговой модели оно составило 14;
- Для PCA + Ridge число компонент подбиралось в ходе настройки и в итоговой спецификации составило 26.

Результаты моделей на тестовой выборке представлены в таблице 2.

Таблица 2 – Результаты моделей на тестовой выборке

№	Model	MAE	RMSE	sMAPE	R2
1	XGBoost (tuned)	821	1988	16.9	0.978
2	OLS (all numeric)	850	1210	24.4	0.992
3	Lasso (lambda.1se)	879	1527	19.9	0.987
4	Elastic Net 0.5 (lambda.1se)	945	1610	21	0.986
5	PLS (ncomp tuned)	985	1640	25.9	0.985
6	Random Forest (ranger)	1127	3086	20.1	0.947
7	Ridge (lambda.1se)	1179	1684	35.4	0.984
8	XGBoost (raw)	1212	4709	14	0.877
9	PCA + Ridge tuned (ncomp=26)	1616	2422	39.3	0.967
10	OLS on Lasso-selected	2452	5341	31.1	0.842
11	PCA(95%) + OLS (ncomp=14)	2537	3416	65.1	0.935

Также была проверена глобальная пространственная автокорреляция показателя приёма. Для логарифмированного значения $\log(1 + Y)$ индекс Морана составил $I = 0,161$. Положительная пространственная автокорреляция оказалась статистически

значимой как по асимптотическому критерию ($p = 0,00595$), так и по перестановочному тесту ($p = 0,013$). Полученный результат указывает на наличие пространственной структуры в распределении показателя по субъектам Российской Федерации и послужил основанием для дополнительного оценивания пространственных моделей (таблица 3).

Таблица 3 – Результаты пространственных моделей на тестовой выборке

Model	Parameter	AIC	BIC	logLik
SAR (lag)	0.337	153.0	165.0	-71.5
SEM (error)	0.453	154.0	166.0	-71.8
OLS	—	161.0	171.0	-76.7

Сравнение прогнозных моделей на отложенных данных за 2024 г. показало, что преимущество различных спецификаций зависит от используемого критерия качества. Наименьшая средняя абсолютная ошибка получена для настроенного XGBoost и составила 821 человек. Минимальное значение RMSE, равное 1210 человек, и максимальный коэффициент детерминации $R^2 = 0,992$ получены для OLS на полном наборе количественных признаков. Наименьшее значение относительной ошибки sMAPE — 14,0 % — показал базовый XGBoost. Настроенная версия XGBoost при этом характеризовалась sMAPE 16,9 % и $R^2 = 0,978$. Таким образом, результаты тестирования не выделяют одной спецификации, одновременно доминирующей по всем использованным метрикам.

Для содержательной интерпретации линейной части модели отдельно были рассмотрены ненулевые коэффициенты Lasso. Итоговая спецификация модели имела вид

$$\hat{Y} = 280,8 + 0,1271X_1 + 0,02539X_2 + 0,3047X_3 + 0,008495X_4 \quad (1)$$

где X_1 — средняя численность студентов вузов региона в позапрошлом году, X_2 — число принятых в СПО после 9-го класса в прошлом году, X_3 — число принятых в СПО после 11-го класса в прошлом году, X_4 — средняя численность студентов СПО в позапрошлом году.

Все вошедшие в итоговую модель коэффициенты имеют положительный знак. Наибольший коэффициент получен для числа принятых в организации СПО после 11-го класса в прошлом году: увеличение данного показателя на 1000 человек соответствует росту прогнозируемого приёма в организации высшего образования приблизительно на 305 человек. Для средней численности студентов вузов в позапрошлом году коэффициент составил 0,1271, то есть увеличение показателя на 1000 человек связано с ростом прогнозируемого приёма примерно на 127 человек.

Менее выраженная связь установлена для числа поступивших в СПО после 9-го класса в предыдущем году: увеличение данного показателя на 1000 человек соответствует увеличению прогнозируемого приёма примерно на 25 человек. Для средней численности студентов СПО в позапрошлом году коэффициент составил 0,008495, что соответствует изменению прогноза примерно на 8,5 человека при увеличении исходного показателя на 1000 человек.

Обсуждение

Полученные результаты линейной модели указывают на выраженную зависимость текущего объёма приёма от характеристик образовательной системы, сформировавшихся в предшествующие периоды.

На следующем этапе была проанализирована важность признаков в настроенной модели XGBoost. В качестве показателя использовался Gain, отражающий вклад переменной в уменьшение функции потерь при построении деревьев. Анализ показал, что наиболее значимым признаком стала средняя численность студентов высшего образования в прошлом году — на неё приходится 68,86 % суммарной важности. Второе место занимает средняя численность студентов высшего образования в позапрошлом году с долей 14,78 %. В совокупности эти два признака формируют 83,64 % общей важности модели.

Следующую по значимости группу образуют показатели приёма в систему среднего профессионального образования после 11-го класса. Число поступивших в СПО после 11-го класса в прошлом году обеспечивает 7,80 % Gain, а соответствующий показатель за позапрошлый год — ещё 3,93 %. Таким образом, уже первые четыре признака объясняют 95,37 % суммарной важности XGBoost.

Вклад остальных характеристик существенно ниже. Численность выпускников 11-х классов в позапрошлом году формирует 1,05 % важности, число студентов высшего образования в расчёте на одного преподавателя в прошлом году — 0,92 %, приём в СПО после 9-го класса в прошлом году — 0,83 %, а численность выпускников 11-х классов в прошлом году — 0,53 %. Материально-инфраструктурные характеристики оказались ещё менее значимыми: на число компьютеров в расчёте на одного студента высшего образования в прошлом году приходится 0,30 %, а на площадь общежитий в расчёте на одного студента в прошлом году — 0,13 %.

Полученные результаты подтверждает выраженную временную инерционность исследуемого процесса: текущий объём приёма во многом воспроизводит уже сложившийся масштаб региональной системы высшего образования. Также существенным результатом является наличие пространственной составляющей исследуемого процесса. Значимое положительное значение индекса Морана показывает, что распределение объёма приёма по территории Российской Федерации не является пространственно независимым: регионы со сходными значениями показателя имеют тенденцию к пространственной группировке.

Таким образом, проведённый анализ выявил две взаимосвязанные особенности процесса приёма: выраженную временную инерционность региональных образовательных показателей и статистически значимую пространственную зависимость между субъектами Российской Федерации. Одновременно результаты сравнения моделей показали, что различные методы имеют преимущества по разным критериям качества, что требует рассматривать точность прогноза как многокритериальную характеристику.

Заключение

Проведённое исследование показало возможность прогнозирования региональной динамики приёма на обучение по программам высшего образования на основе объединения показателей общего, среднего профессионального и высшего

образования.

Основным направлением дальнейшей работы является объединение временной и пространственной составляющих в единой прогнозной модели. В проведённом исследовании временные зависимости учитывались в основных прогнозных моделях через лаговые признаки, тогда как пространственные эффекты анализировались отдельно с помощью индекса Морана, SAR и SEM. Следующий этап должен быть связан с построением моделей, одновременно учитывающих динамику показателей во времени и взаимосвязи между регионами.

Кроме того, перспективно расширить временной горизонт наблюдений и перейти от агрегирования на уровне «регион–год» к более детализированным данным по отдельным образовательным организациям и направлениям подготовки. Это позволит учитывать внутригородские и внутрирегиональные различия, образовательную миграцию, изменения структуры сети вузов и другие факторы, которые не отражаются в существующей агрегированной базе. Расширение временного ряда также позволит проводить проверку качества прогноза не на одном, а на нескольких последовательно отложенных периодах.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Pouramini J. Forecasting Number of Students Applicant for Courses by Artificial Neural Networks // *Journal of Mathematics and Computer Science*. 2014. Vol. 12. P. 263–270.
2. Chen Y. A., Li R., Hagedorn L. S. Undergraduate International Student Enrollment Forecasting Model: An Application of Time Series Analysis // *Journal of International Students*. 2019. Vol. 9, No. 1. P. 242–261. DOI: 10.32674/jis.v9i1.266.
3. Langston R., Wyant R., Scheid J. Strategic Enrollment Management for Chief Enrollment Officers: Practical Use of Statistical and Mathematical Data in Forecasting First Year and Transfer College Enrollment // *Strategic Enrollment Management Quarterly*. 2016. Vol. 4, No. 2. P. 74–89. DOI: 10.1002/sem3.2008574.
4. Hussain J., Rosangliana D., Vanlalruata. Student Gross Enrolment Ratio Forecasting: A Comparative Study Using Statistical Method and Machine Learning // *International Journal of Information and Education Technology*. 2023. Vol. 13, No. 3. P. 439–447. DOI: 10.18178/ijiet.2023.13.3.1824.
5. Pradnyana I. P. B. A., Wisnawa I. P. O., Puspita N. N. H. New Student Admission Forecasting Model with Support Vector Machine Method: Case Study of Bali State Polytechnic // *Jurnal Teknologi Informasi dan Pendidikan*. 2025. Vol. 18, No. 1. P. 760–772. DOI: 10.24036/jtip.v18i1.956.
6. Sembiring F., Ayuningsih R. R., Erfina A., Risky. A Multi-Model Forecasting Framework for New Student Admissions: SMA, ARIMA, and Random Forest Approaches // *Scientific Journal of Informatics*. 2026. Vol. 13, No. 2. P. 363–372. DOI: 10.15294/sji.v13i2.42746.
7. Wihardi H., Md Johar M. G. A Comparative Study of Arima, Prophet and LSTM for International Students Enrollment // *Merkurius: Jurnal Riset Sistem Informasi dan Teknik Informatika*. 2026. Vol. 4, No. 1. P. 243–253. DOI: 10.61132/merkurius.v4i1.1480.
8. Tapio R., Tarepe D. Comparative Analysis of Random Forest and Hybrid ARIMA-

Random Forest Models for Student Enrollment Forecasting in Higher Education. *Journal of Advances in Mathematics and Computer Science*, 2025, Vol. 40, No. 3, pp. 124–136. DOI: 10.9734/jamcs/2025/v40i31982.

9. Lo Y.-H., Chang D.-F., Chang A. Exploring Concurrent Relationships between Economic Factors and Student Mobility in Expanding Higher Education Achieving 2030 // *Sustainability*. 2022. Vol. 14. Art. 14612. DOI: 10.3390/su142114612.

10. Cristuta M. D. Forecasting Enrollment Trends in a Higher Education Institution Using Time Series and Autoregressive Integrated Moving Average Models // *Journal of Higher Education Research Disciplines*. 2025. Vol. 10, No. 2. P. 83–95. DOI: 10.65023/jherd.v10i2.254.

11. Sá C., Florax R. J. G. M., Rietveld P. Determinants of the Regional Demand for Higher Education: Tinbergen Institute Discussion Paper TI 2003-013. Amsterdam: Tinbergen Institute, 2003. 36 p.

12. Gareis P., Broekel T. The Spatial Patterns of Student Mobility Before, During and After the Bologna Process in Germany // *Tijdschrift voor Economische en Sociale Geografie*. 2022. Vol. 113, No. 3. P. 290–309. DOI: 10.1111/tesg.12507.

13. Şahin S. Z., Bulut Şahin B., Söylemez E. An Evaluation of the Spatial Repercussions of Student Mobility Policy in European Higher Education Area Using Network Analysis // *Research in Educational Administration & Leadership*. 2024. Vol. 9, No. 1. P. 74–121. DOI: 10.30828/real.1392301.

14. Litmeyer M.-L., Hennemann S. Forecasting First-Year Student Mobility Using Explainable Machine Learning Techniques // *Review of Regional Research*. 2024. Vol. 44. P. 119–140. DOI: 10.1007/s10037-024-00207-x.

15. Özmen Z. Social and Economic Factors and Their Relationship with Gender Gap in STEM in Higher Education: A Spatial Analysis of Turkey // *GeoJournal*. 2024. Vol. 89. Art. 179. DOI: 10.1007/s10708-024-11153-z.

16. Nuramaliyah, Rahkmawati Y., Faulina R., Fitriana I. N. L. Spatial Analysis of Gross Enrollment Ratio in Higher Education Using the Geographically Weighted Regression Approach // *J Statistika*. 2025. Vol. 18, No. 2. P. 1041–1051. DOI: 10.36456/jstat.vol18.no2.a10792.

17. Tapio R., Tarepe D. Comparative Analysis of Random Forest and Hybrid ARIMA-Random Forest Models for Student Enrollment Forecasting in Higher Education // *Journal of Advances in Mathematics and Computer Science*. 2025. Vol. 40, No. 3. P. 124–136. DOI: 10.9734/jamcs/2025/v40i31982.

18. Дождиков А. В., Корнилова Е. В. Образовательная миграция абитуриентов между регионами Российской Федерации как источник данных для планирования развития системы высшего образования // *Высшее образование в России*. 2023. Т. 32, № 3. С. 67–83. DOI: 10.31992/0869-3617-2023-32-3-67-83. EDN: DUSOSM.

19. Дождиков А. В. Межрегиональная миграция молодежи в Российской Федерации // *ДЕМИС. Демографические исследования*. 2024. Т. 4, № 3. С. 119–137. DOI: 10.19181/demis.2024.4.3.8. EDN: ZPVQJZ.

20. Ситковский А. М., Дождиков А. В. Внутрироссийская образовательная миграция молодежи: привлекательность регионов для получения высшего образования // *Образование и наука*. 2025. Т. 27, № 6. С. 149–179. DOI: 10.17853/1994-5639-2025-6-

149-179. EDN: TSLSLP.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Черемухин Артем Дмитриевич – кандидат экономических наук, доцент кафедры «Математика и вычислительная техника», ГБОУ ВО «Нижегородский государственный инженерно-экономический университет», г. Княгинино, Россия, E-mail: ngie.u.cheremuhin@yandex.ru, ORCID: 0000-0003-4076-5916, SPIN код: 3067-9927.

Пахрин Андрей Алексеевич – студент 4 курса, ГБОУ ВО «Нижегородский государственный инженерно-экономический университет», г. Княгинино, Россия, E-mail: andrey.pakhrin@gmail.com, SPIN-code: 3067-9927