СУРАЙ А. А., РОЧЕВ К. В. АВТОМАТИЗИРОВАННОЕ СРАВНЕНИЕ ВУЗОВ НА ОСНОВЕ ОТКРЫТЫХ ГОСУДАРСТВЕННЫХ ДАННЫХ: МЕТОДИКА И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

УДК 004.9:378, ГРНТИ 28.29.15
Статья поступила в редакцию 10.06.2026

Автоматизированное сравнение ВУЗов на основе открытых государственных данных: методика и программная реализация

А. А. Сурай[1], К. В. Рочев[2]

[1]ООО «Газпром трансгаз Ухта», г. Ухта;
[2]Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, г. Москва

Automated comparison of universities based on open government data: methodology and software implementation

A. A. Suray[1], K. V. Rochev[2]

[1]Gazprom Transgaz Ukhta LLC, Ukhta
[2]Russian Presidential Academy of National Economy and Public Administration, Moscow

В работе рассматривается задача автоматизированного многомерного сравнения образовательных организаций высшего образования на основе открытых государственных данных. Проведён анализ существующих подходов к ранжированию вузов, выявлены их ограничения. Предложена методика, охватывающая полный цикл обработки данных: от извлечения из открытых источников до генерации текстовых аналитических заключений. Методика включает робастную фильтрацию статистических выбросов на основе медианного абсолютного отклонения (MAD), комбинированное нормирование показателей и три режима многомерного сравнения (табличный, линейный график, лепестковая диаграмма). Разработан и апробирован программный комплекс UNI-METRICS, реализованный по

The paper addresses the problem of automated multidimensional comparison of higher education institutions based on open government data. An analysis of existing approaches to university ranking is conducted, and their limitations are identified. A methodology is proposed that covers the full data processing cycle: from extraction from open sources to generation of textual analytical reports. The methodology includes robust statistical outlier filtering based on median absolute deviation (MAD), combined indicator normalization, and three modes of multidimensional comparison (tabular, line chart, radar chart). The UNI-METRICS software system, implemented with a three-tier architecture (MongoDB, ASP.NET

трехуровневой архитектуре (MongoDB, ASP.NET Core 8, Vanilla JS) с подсистемой безопасности на основе JWT и RBAC. Экспериментальная апробация на выборке из пяти вузов нефтегазового профиля за период 2021–2024 гг. подтвердила способность методики выявлять содержательные закономерности при сокращении временных затрат на сравнительный анализ до 490 раз по сравнению с ручным методом.

Core 8, Vanilla JS) with a security subsystem based on JWT and RBAC, has been developed and tested. Experimental validation on a sample of five oil and gas universities for the period 2021–2024 confirmed the methodology's ability to reveal meaningful patterns while reducing the time required for comparative analysis by a factor of 490 compared to the manual method.

Ключевые слова: *мониторинг деятельности вузов, открытые данные, автоматизированное сравнение, робастная фильтрация, медианное абсолютное отклонение (MAD)*

Keywords: *university performance monitoring, open data, automated comparison, robust filtering, median absolute deviation (MAD), multidimensional visualization, ETL process, large language models*

Введение

Цифровая трансформация системы высшего образования Российской Федерации привела к формированию беспрецедентных по объему массивов открытых данных о деятельности образовательных организаций [1]. Ежегодно на официальном портале «Мониторинг деятельности образовательных организаций высшего образования» публикуются сведения более чем по 50 показателям для свыше 1000 вузов и филиалов, охватывающие образовательную, научно-исследовательскую, международную, финансово-экономическую, инфраструктурную и кадровую сферы деятельности [2].

Однако между наличием данных и возможностью их практического использования существует фундаментальный разрыв. Во-первых, данные публикуются в «сыром» виде и требуют трудоемкой предобработки: удаления дубликатов, нормализации наименований вузов, фильтрации статистических выбросов. Во-вторых, существующие инструменты сравнения — академические рейтинги (RAEX, QS, THE) и штатный интерфейс портала мониторинга — либо методологически непрозрачны и статичны, либо ограничены табличной парадигмой представления, не позволяющей выполнять многомерный визуальный анализ. Ни один из существующих инструментов не предоставляет конечному пользователю возможности управлять критериями сравнения.

Таким образом, актуальной научно-практической задачей является разработка методики, обеспечивающей полный цикл автоматизированного

сравнения вузов: от экстракции и очистки открытых данных до формирования многомерных визуализаций и текстовых аналитических заключений.

Анализ существующих подходов

Анализ глобальных и национальных академических рейтингов (ARWU, QS, THE, RAEX) позволяет выделить ряд фундаментальных ограничений. Основные из них — отсутствие прозрачности и гибкости весовых коэффициентов, значительная доля субъективных репутационных компонентов, а также статичность публикации и отсутствие интерактивного анализа. Государственный портал, являющийся источником данных для настоящего исследования, предоставляет широкие возможности для просмотра отдельных показателей, однако его функциональность ограничена табличной парадигмой, отсутствием средств предобработки данных (в частности, фильтрации выбросов) и невозможностью пользовательской композиции показателей [3].

Методика автоматизированного сравнения

Разработанная методика представляет собой многостадийный конвейер обработки данных, концептуально разделенный на четыре последовательных этапа: ETL-процесс сбора данных, очистка и нормализация, интеграция и расчет интегральных показателей, сравнительный анализ и визуализация.

ETL-процесс сбора данных. Извлечение данных реализуется по иерархической схеме, последовательно обходящей три уровня вложенности: год мониторинга, субъект РФ и конкретный вуз. Для минимизации сетевого трафика применяется механизм файлового кэширования. Парсинг загруженной HTML-страницы включает извлечение метаданных и разбор таблиц показателей, поддерживающих три формата представления [4].

Очистка и нормализация данных. Ключевой проблемой является неединообразие полных официальных наименований образовательных организаций. Методика включает процедуру приведения наименований к каноническому краткому виду с использованием регулярных выражений для удаления организационно-правовых форм и статусных префиксов.

Для фильтрации статистических выбросов применяется метод медианного абсолютного отклонения (MAD), относящийся к классу робастных статистических методов. MAD определяется как:

$$MAD=median(|Xi-median(X)|)$$

Наблюдение классифицируется как выброс при выполнении условия:

$$|Xi-median(X)|>t\times k\times MAD$$

где k ≈ 1.4826 — константа, связывающая MAD со стандартным отклонением для нормального распределения, t — пороговый множитель.

Ключевое преимущество MAD перед классическим правилом «трех сигм» — устойчивость к наличию выбросов в самой выборке.

Расчет интегральных показателей. Для приведения гетерогенных показателей к единой безразмерной шкале применяются два взаимодополняющих алгоритма. Минимаксное нормирование:

pi′=pi−min(P)max(P)−min(P)*pi′*=max(*P*)−min(*P*)*pi*−min(*P*)

Центильное нормирование определяет положение вуза относительно эмпирической функции распределения. Интегральная оценка вычисляется как аддитивная свертка:

I(vj)=∑iwi×pij′*I*(*vj*)=∑*iwi*×*pij*′

где w_i — вес i-го показателя, задаваемый пользователем, Σ w_i = 1. Принципиальным отличием от рейтинговых систем является возможность динамического профилирования с пользовательскими весами.

Для кластеризации вузов и выделения лиг применяется алгоритм k-средних с автоматическим определением оптимального числа кластеров на основе отношения размаха вариации к стандартному отклонению [5].

Интеллектуальная интерпретация результатов. Завершающим компонентом методики является блок интеллектуальной интерпретации с использованием больших языковых моделей (LLM). Модель получает на вход структурированный контекст: наименования сравниваемых вузов, значения ключевых показателей, позиции в рейтингах, выявленные отклонения. Важно подчеркнуть, что LLM не производит вычислений — она выполняет функцию семантического интерпретатора над результатами, полученными строгими статистическими методами.

Программная реализация

Разработанный программный комплекс UNI-METRICS спроектирован по многослойной архитектуре и состоит из трех логических уровней. Уровень данных представлен документно-ориентированной СУБД MongoDB, выбор которой обоснован гетерогенностью схемы данных: перечень показателей мониторинга изменяется от года к году. Уровень бизнес-логики реализован на платформе .NET 8 с использованием языка C# и фреймворка ASP NET Core 8 и разделен на библиотеку сбора и очистки данных, библиотеку аналитики и веб-сервис с REST API. Уровень представления реализован в виде набора статических HTML-страниц с использованием чистого JavaScript.

Подсистема ETL организована по паттерну «Фасад». Центральным компонентом выступает класс VuzMonitorFacade, инкапсулирующий взаимодействие специализированных компонентов для загрузки контента,

парсинга HTML, очистки и сохранения данных. Ключевой алгоритмической особенностью является интеллектуальное детектирование изменений при сохранении в MongoDB: метод HasUniversityChanged() выполняет полное сравнение всех полей существующего документа с обновленным объектом.

Клиентская часть реализует семь типов визуализации: диаграмма рассеяния, тепловая карта, боксплот, гистограмма, тренд по годам, топ-N вузов и пузырьковая диаграмма. Фильтрация выбросов методом MAD выполняется на стороне клиента непосредственно перед визуализацией, что позволяет применять фильтрацию к данным, уже агрегированным и отфильтрованным пользователем по региону, году и кластеру.

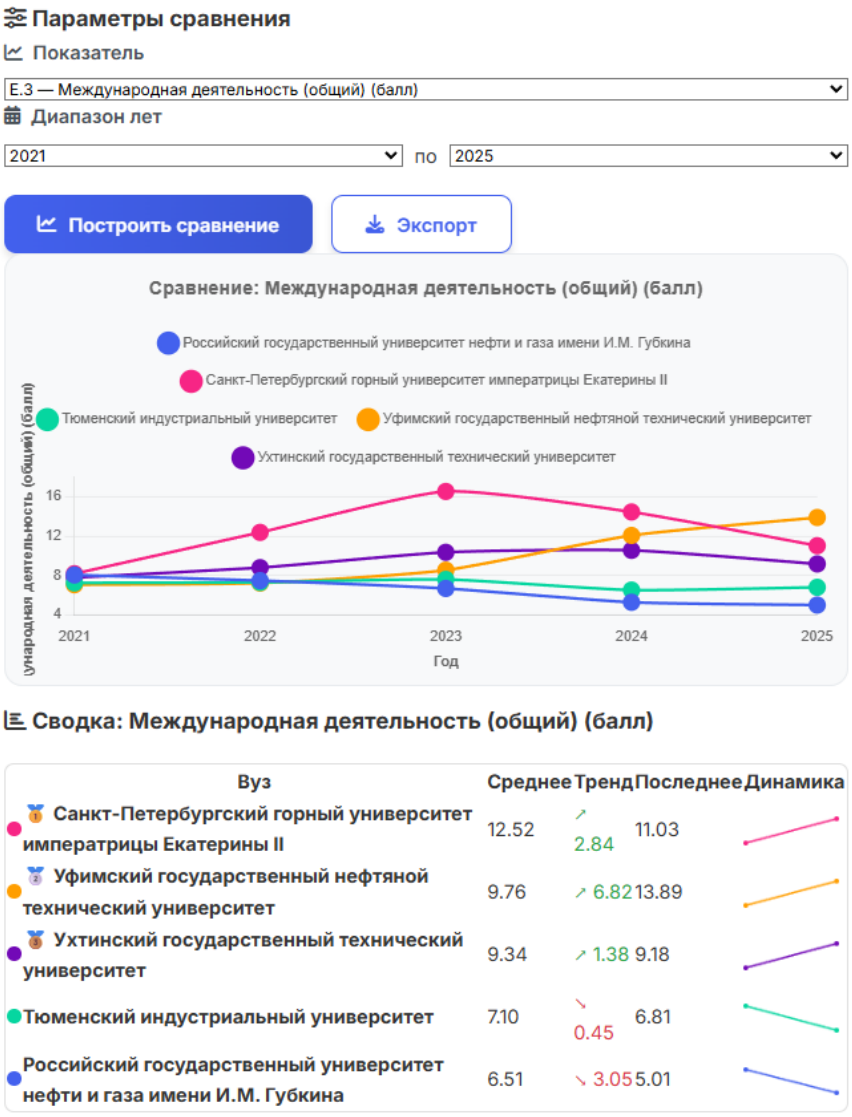

Рисунок 1. Динамика по годам

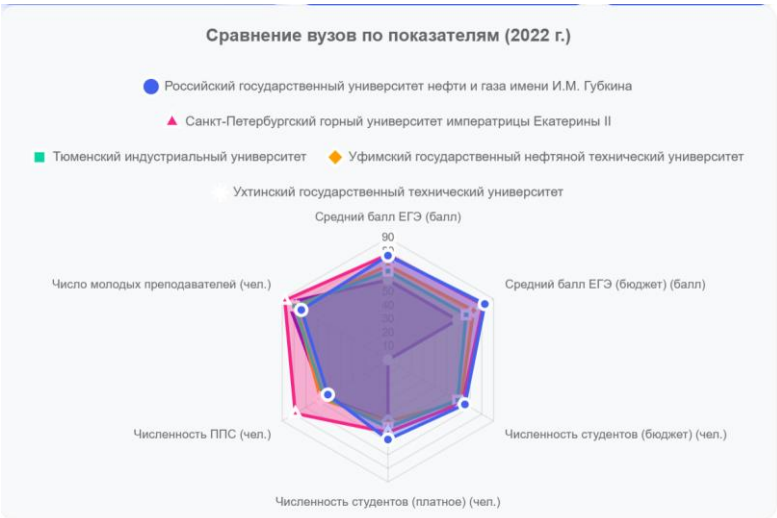

Рисунок 2. Лепестковая диаграмма (3–8 показателей)

Экспериментальная апробация

Экспериментальная база данных UNI-METRICS содержит 8600 записей, охватывающих все образовательные организации высшего образования Российской Федерации за период 2021–2025 гг. Для проведения эксперимента сформирована целевая выборка из пяти вузов нефтегазового профиля: РГУ нефти и газа им. И.М. Губкина, Санкт-Петербургский горный университет, УГНТУ, ТИУ и УГТУ.

Сравнение выполнено за 2024 год по шести показателям. Результаты (Таблица 1) позволили выявить многократный разрыв по научной продуктивности (E.2): показатель Горного университета (6029.97 тыс. руб. на НПР) превышает показатель Ухтинского ГТУ (165.19) в 36.5 раз. По международной деятельности (E.3) неожиданным является лидерство Уфимского нефтяного технического университета (13.89 балла), опережающего столичные вузы.

Таблица 1. Сравнение вузов нефтегазового профиля по ключевым показателям, 2024 г.

Показатель	Губкина	Горный СПб	УГНТУ	ТИУ	УГТУ
E.1 – Средний балл ЕГЭ, балл	74,90	78,52	69,36	65,02	62,84
E.2 – Объем НИОКР на НПР, тыс. руб.	1418,13	6029,97	802,57	246,69	165,19
E.3 – Международная деятельность, балл	5,01	11,03	13,89	6,81	9,18

Анализ динамики научной продуктивности за период 2021–2024 гг. показал, что все пять вузов демонстрируют положительную динамику, однако наибольший абсолютный прирост приходится на Горный университет (+3957.21 тыс. руб.). Интегральное профилирование позволило классифицировать вузы по типу развития: «сбалансированный лидер» (Горный университет), «столичный середняк» (Губкинский университет), «интернационализация без науки» (УГНТУ), «региональный отраслевой» (ТИУ, УГТУ).

Сокращение временных затрат на формирование сравнительного анализа достигает до 490 раз по сравнению с ручным методом: с ~180 минут до ~22 секунд.

Заключение

Разработанная методика автоматизированного сравнения вузов представляет собой полный, формально описанный и воспроизводимый процесс преобразования открытых государственных данных в многомерные сравнения и текстовые аналитические заключения. В отличие от существующих рейтинговых систем, методика обеспечивает прозрачность алгоритмов, управляемость критериев и автоматизацию всех этапов анализа — от экстракции данных до формирования текстового заключения. Экспериментальная апробация подтвердила способность методики выявлять содержательные закономерности при сокращении временных затрат на анализ до 490 раз по сравнению с ручным методом.

Список использованных источников и литературы

1. О проведении мониторинга эффективности образовательных организаций высшего образования : приказ Министерства науки и высшего образования РФ от 18.01.2023 № 33. — URL: https://monitoring.miccedu.ru/ (дата обращения: 10.04.2026).

2. Об образовании в Российской Федерации : Федеральный закон от 29.12.2012 № 273-ФЗ (ред. от 25.12.2024). — Москва, 2012. — 404 с.

3. Сурай А. А., Рочев К. В., Маринина А. А. Сравнительный анализ научно-образовательного потенциала нефтегазовых университетов Российской Федерации с помощью информационной системы «UNI-METRICS» // Материалы XXVI Международной молодёжной научной конференции «СЕВЕРГЕОЭКОТЕХ-2025». — 2025. — Т. 26. — С. 254–257.

4. Саати, Т. Л. Принятие решений. Метод анализа иерархий / Т. Л. Саати; пер. с англ. — Москва : Радио и связь, 1993. — 278 с.

5. Лемешко, Б. Ю. Робастные статистические методы: анализ и применение / Б. Ю. Лемешко, С. Б. Лемешко. — Новосибирск : Изд-во НГТУ, 2019. — 320 с.

List of references

1. On Conducting Monitoring of the Effectiveness of Higher Education Institutions: Order of the Ministry of Science and Higher Education of the Russian

Federation No. 33 dated January 18, 2023. – URL: https://monitoring.miccedu.ru/ (accessed: 10.04.2026).

2. On Education in the Russian Federation: Federal Law No. 273-FZ dated December 29, 2012 (as amended on December 25, 2024). – Moscow, 2012. – 404 p.

3. Suraj, A.A., Rochev, K.V., Marinina, A.A. Comparative Analysis of the Scientific and Educational Potential of Oil and Gas Universities of the Russian Federation Using the Information System "UNI-METRICS" // *Proceedings of the XXVI International Youth Scientific Conference "SEVERGEOEKOTECH-2025"*. – 2025. – Vol. 26. – P. 254–257.

4. Saaty, T.L. *Decision Making. The Analytic Hierarchy Process*. – Moscow: Radio i Svyaz, 1993. – 278 p.

5. Lemeshko, B.Yu., Lemeshko, S.B. *Robust Statistical Methods: Analysis and Application*. – Novosibirsk: NSTU Publishing House, 2019. – 320 p.