УДК 330.46:004.6

КОМПЛЕКСНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ БОЛЬШИХ ДАННЫХ: ИНТЕГРАЦИЯ МЕТОДОВ СНИЖЕНИЯ РАЗМЕРНОСТИ, КЛАСТЕРИЗАЦИИ И РЕГРЕССИИ

Е.И. Полоско
Старший преподаватель кафедры экономической информатики БГУИР
e.i.polosko@gmail.com

Е.И. Полоско

Окончила Белорусский государственный университет. Область научных интересов связана с анализом данных и методами машинного обучения, экономико-математическим моделированием, технологиями Big Data и решениями на основе интернета вещей (IoT).

Аннотация. В статье рассматривается методология комплексного статистического анализа, объединяющая методы описательной статистики, корреляционного анализа, снижения размерности (метод главных компонент), кластеризации (K-Means) и регрессионного моделирования для обработки больших данных. Используется пошаговый подход к анализу данных, реализованный в среде Python и апробированный на имитированных производственных и энергетических данных ОАО «МАЗ» за 2021–2025 гг. Проведён сравнительный анализ регрессионных моделей прогнозирования энергопотребления; ансамблевые методы (Gradient Boosting) демонстрируют наивысшую точность (R^2 = 0,873 и MAE = 3,61 МВт·ч) по сравнению с линейными моделями. Практическая значимость работы заключается в универсальном подходе к анализу производственных данных для поддержки управленческих решений и оптимизации бизнес-процессов на промышленных предприятиях.

Ключевые слова: большие данные, комплексный статистический анализ, ОАО «МАЗ», метод главных компонент, кластерный анализ, регрессионное моделирование, энергоэффективность, машиностроение, предиктивная аналитика, Python

Введение. В современных условиях цифровой трансформации объём генерируемых данных растёт экспоненциально. По данным аналитиков, глобальный объём данных достигнет 394 зеттабайт к 2028 году [1], при этом до 90 % корпоративных данных остаются неструктурированными [2]. Мировой рынок больших данных, оценивавшийся в 262,87 млрд долларов в 2024 году, по прогнозам достигнет 1 019 млрд долларов к 2035 году со среднегодовым темпом роста 13,10 % [3].

Динамика роста рынка представлена на рисунке 1.

В этих условиях применение изолированных статистических методов часто оказывается недостаточным для извлечения значимых закономерностей из многомерных массивов данных. Необходим интегрированный подход, объединяющий несколько этапов статистического анализа в единый аналитический конвейер [4].

Сегмент продвинутой аналитики (Advanced Analytics) прогнозируемо займёт свыше 68,56 % глобального рынка Big Data в ближайшее десятилетие.

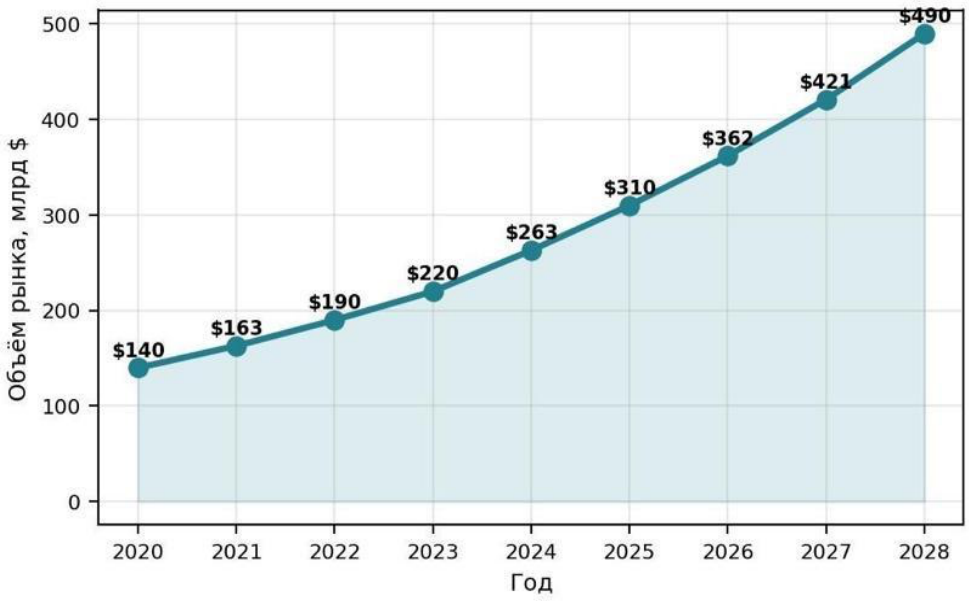

Рисунок 1. Динамика мирового рынка больших данных, млрд $ [3]

Целью данной работы является разработка и экспериментальная апробация методологии комплексного статистического анализа больших данных, интегрирующей методы описательной статистики, корреляционного анализа, снижения размерности, кластеризации и регрессионного моделирования для оптимизации бизнес-решений.

Методология комплексного статистического анализа. Предлагаемая методология представляет собой последовательный аналитический конвейер (pipeline), состоящий из семи взаимосвязанных этапов. Каждый этап решает определённую задачу, а результаты предыдущего этапа служат входными данными для последующего [5]. Общая схема методологии представлена на рисунке 2.

Рисунок 2. Схема аналитического конвейера комплексного статистического анализа

Характеристика каждого этапа представлена в таблице 1.

Таблица 1. Этапы комплексного статистического анализа больших данных

№	Этап анализа	Применяемые методы	Ожидаемый результат
1	Предобработка данных	Очистка, нормализация, обработка пропусков	Чистый и структурированный датасет
2	Описательная статистика	Меры центральной тенденции, вариации, распределения	Общая характеристика данных
3	Корреляционный анализ	Коэффициент Пирсона, тепловая карта	Матрица корреляций признаков
4	Снижение размерности	Метод главных компонент (PCA)	Сжатое пространство признаков
5	Кластерный анализ	K-Means, метод локтя, силуэтный анализ	Сегменты (кластеры) данных
6	Регрессионное моделирование	Linear, Ridge, Lasso, RF, GBM	Предиктивные модели
7	Валидация	Кросс-валидация, R^2, MAE, RMSE	Оценка качества моделей

Ключевым преимуществом данного подхода является синергетический эффект: комбинация методов позволяет компенсировать ограничения каждого отдельного метода и обеспечивает более полное извлечение информации из данных [6].

Корреляционный анализ многомерных данных. На этапе корреляционного анализа вычисляется матрица парных коэффициентов корреляции Пирсона для всех числовых признаков набора данных. Данная матрица позволяет выявить линейные зависимости между переменными и определить мультиколлинеарные признаки, подлежащие агрегации или исключению [7].

Для экспериментальной апробации был использован синтетический набор данных, моделирующий бизнес-показатели ОАО «МАЗ» за 60 месяцев (январь 2021 – декабрь 2025 гг.). На этапе корреляционного анализа вычислена матрица парных коэффициентов корреляции Пирсона для всех восьми производственных показателей: объёма производства (X_1), энергопотребления (Y), простоя оборудования (X_2), доли брака (X_3), расходов на сырьё (X_4), производительности труда (X_5), загрузки оборудования (X_6) и выручки (X_7).

Тепловая карта корреляций представлена на рисунке 3.

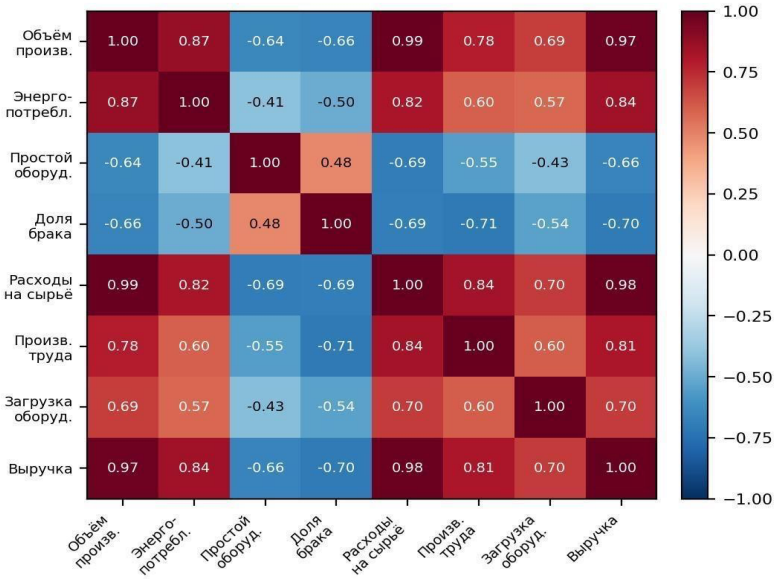

Рисунок 3. Тепловая карта корреляций производственных показателей ОАО «МАЗ»

Анализ матрицы корреляций выявил следующие значимые зависимости: сильная положительная корреляция между объёмом производства и энергопотреблением (r = 0,87), объёмом и выручкой (r = 0,97), объёмом и расходами на сырьё (r = 0,99).

Сильная отрицательная корреляция обнаружена между простоем оборудования и расходами на сырьё (r = −0,69), долей брака и производительностью труда (r = −0,71). Наличие групп коррелированных признаков обосновывает применение PCA.

Снижение размерности методом главных компонент. Метод главных компонент (Principal Component Analysis, PCA) является одним из наиболее эффективных инструментов снижения размерности для больших данных. PCA преобразует исходные показатели в новый набор показателей так, чтобы они не дублировали друг друга и шли в порядке убывания важности: первые компоненты объясняют наибольшую часть разброса данных.

Математически сначала для данных строится ковариационная матрица Σ, которая описывает, как признаки связаны друг с другом.

Затем для этой матрицы находят специальные направления (собственные векторы) и соответствующие им числа (собственные значения); каждое направление задаёт главную компоненту, а связанное с ним число показывает, какую долю общего разброса данных объясняет эта компонента: чем больше собственное значение, тем важнее компонента.

Результаты применения PCA к исследуемому набору данных представлены на рисунке 4 и в таблице 2.

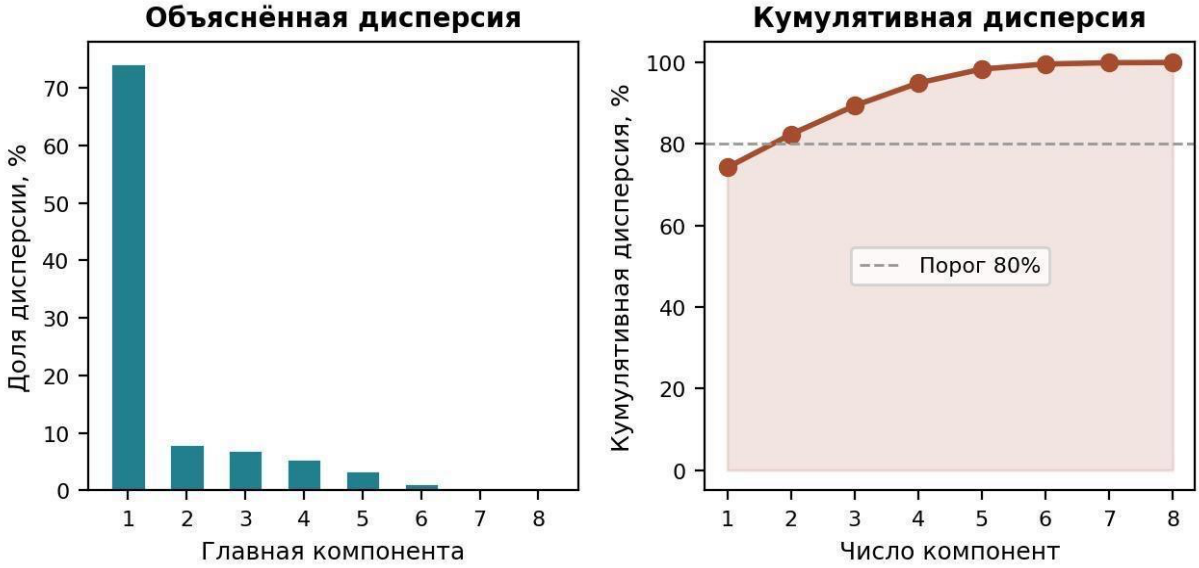

Рисунок 4. Объяснённая и кумулятивная дисперсия главных компонент

Таблица 2 . Результаты анализа главных компонент

Компонента	Собственное значение	Доля дисперсии, %	Кумулятивная доля, %
PC1	6,038	74,22	74,22
PC2	0,662	8,13	82,35
PC3	0,572	7,02	89,38
PC4	0,453	5,56	94,94
PC5	0,279	3,43	98,38
PC6	0,100	1,24	99,61
PC7	0,026	0,32	99,93
PC8	0,006	0,07	100,00

Как видно из таблицы 2, первая главная компонента объясняет 74,22% всего разброса данных и фактически отражает общий уровень производственной активности.

Она объединяет изменения объёма производства, расходов на сырьё, выручки и энергопотребления, то есть тех показателей, которые сильнее всего связаны между собой (коэффициенты корреляции от 0,87 до 0,99). Большое значение $\lambda_1 = 6{,}038$ по сравнению с остальными говорит о том, что за поведением системы ОАО «МАЗ» стоит один ведущий скрытый фактор, определяющий её основную динамику.

Кластерный анализ в пространстве главных компонент. После снижения размерности был проведён кластерный анализ методом K-Means в пространстве первых двух главных компонент .

Оптимальное число кластеров (k = 3) определено с использованием метода «локтя» и силуэтного анализа.

Результаты кластеризации визуализированы на рисунке 5.

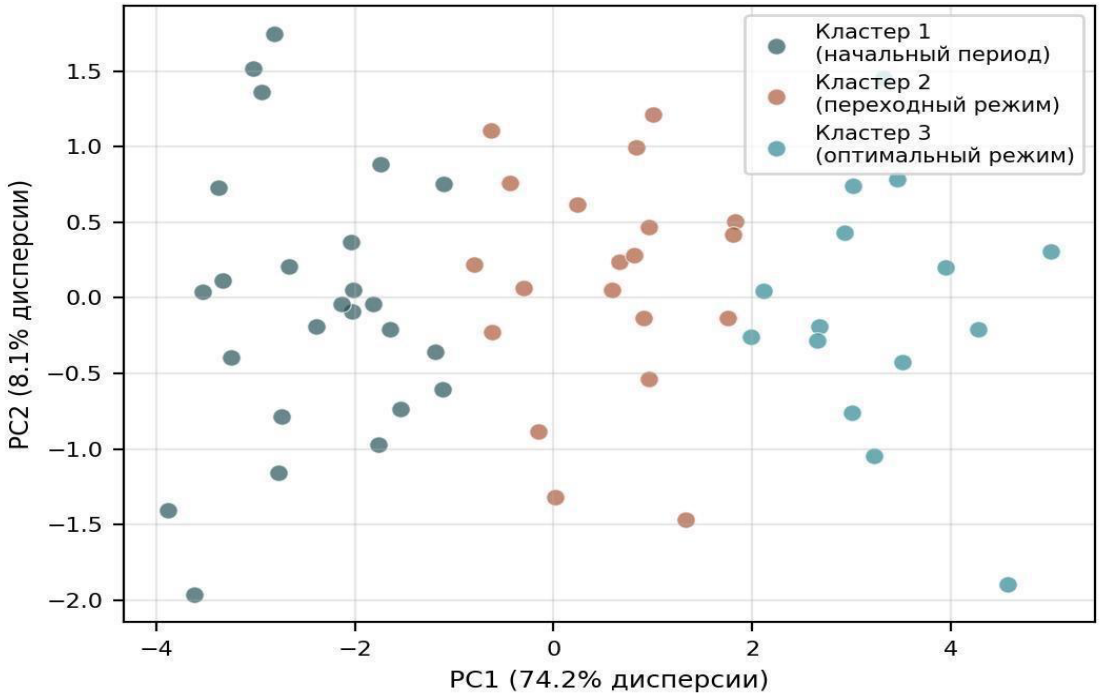

Рисунок 5. Результаты кластерного анализа в пространстве PC1–PC2

Характеристика выявленных кластеров представлена в таблице 3.

Таблица 3. Характеристика кластеров производственных периодов ОАО «МАЗ»

Кластер	**n**	**Произв., ед.**	**Энергия, МВт·ч**	**Простой, ч**	**Загрузка, %**
C1 (начальный)	25	1286	220	112	74,7
C2 (переходный)	20	1434	235	95	78,3
C3 (оптимальный)	15	1573	253	81	83,5

Кластер 1 («начальный период», 2021 – начало 2022 гг.) включает 25 месяцев и соответствует наименее эффективному состоянию: выпуск невысокий (1 286 ед./мес.), простой большой (112 ч/мес.), загрузка низкая (74,7%), энергопотребление минимально (220 МВт·ч/мес.).

Кластер 2 («переходный период», 2022–2023 гг.) объединяет 20 месяцев и отражает промежуточный этап: выпуск растёт до 1 434 ед./мес., простои снижаются до 95 ч/мес., загрузка повышается до 78,3%, что соответствует постепенному наращиванию мощностей.

Кластер 3 («оптимальный режим», 2024–2025 гг.) содержит 15 месяцев и описывает наиболее эффективную работу: выпуск 1 573 ед./мес., простои 81 ч/мес., загрузка 83,5%, при максимальном энергопотреблении (253 МВт·ч/мес.) удельная энергоёмкость снижается, что говорит о росте энергоэффективности.

Переход от кластера 1 к кластеру 3 показывает последовательную модернизацию и внедрение IoT-систем на ОАО «МАЗ»; выделение трёх режимов позволяет давать практические рекомендации по оптимизации загрузки оборудования и сокращению простоев.

Регрессионное моделирование энергопотребления. На заключительном этапе проведено регрессионное моделирование энергопотребления предприятия на основе пяти производственных показателей (объём производства, простой оборудования, доля брака, производительность труда, загрузка оборудования). Сравнительный анализ выполнен для пяти методов с использованием 5-кратной кросс-валидации [9, 10]. Результаты представлены в таблице 4 и на рисунке 6.

Таблица 4. Сравнительный анализ регрессионных моделей прогнозирования энергопотребления

Модель	R^2	MAE, МВт·ч	RMSE, МВт·ч	Время обуч., с
Линейная регрессия	0,768	6,42	7,34	0,08
Ridge-регрессия	0,782	5,98	7,02	0,10
Lasso-регрессия	0,774	6,15	7,18	0,09
Random Forest	0,847	4,27	5,12	1,54
Gradient Boosting	0,873	3,61	4,38	2,81

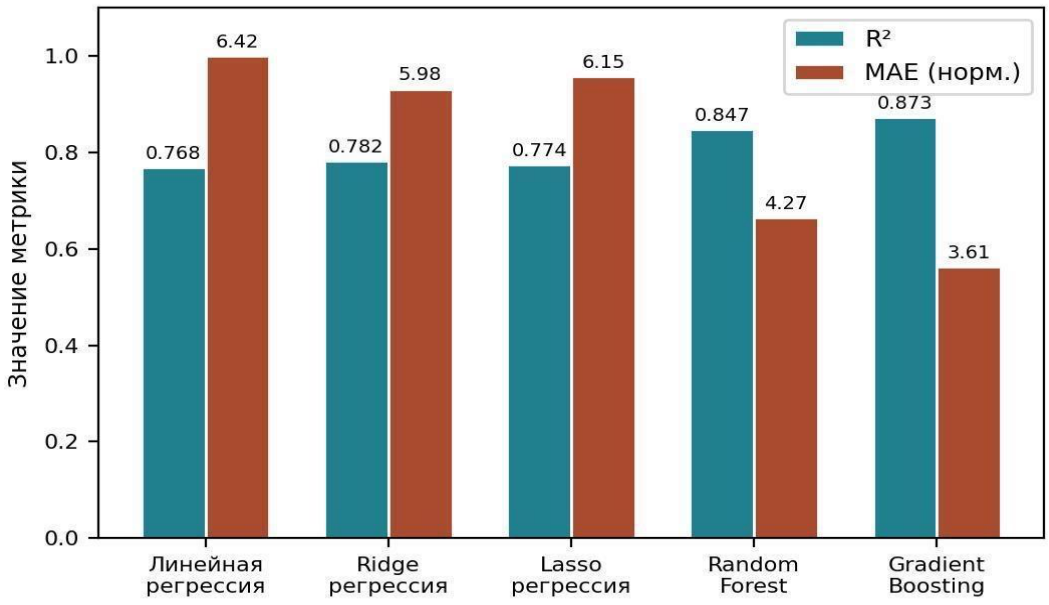

Рисунок 6. Сравнение метрик качества регрессионных моделей

Ансамблевые методы лучше справились с задачей, чем обычные линейные модели. Gradient Boosting дал лучшие результаты (R^2 = 0,873, MAE = 3,61 МВт·ч) и улучшил качество по сравнению с простой линейной регрессией на 13,7%, что указывает на выраженную нелинейность связи между энергопотреблением и производственными показателями.

Модель Random Forest оказалась второй по точности (R^2 = 0,847) и дополнительно позволяет оценивать вклад каждого признака, что упрощает интерпретацию.

Регуляризованные линейные модели (Ridge, Lasso) немного улучшили результаты базовой линейной регрессии, подтверждая наличие тесно связанных между собой признаков, выявленных на этапе корреляционного анализа.

Обсуждение результатов. Предложенный комплексный подход обладает рядом преимуществ для анализа производственных данных промышленного предприятия:

1) PCA позволил сократить размерность с 8 до 2 признаков при сохранении 82,35 % информации, что значительно упрощает визуализацию и интерпретацию многомерных производственных данных ОАО «МАЗ».

2) Кластерный анализ выявил три характерных режима работы предприятия, соответствующих этапам модернизации и повышения энергоэффективности.

Переход от кластера 1 к кластеру 3 сопровождается снижением простоя на 28 % и ростом загрузки оборудования на 12 %.

3) Ансамблевые модели обеспечивают точность прогнозирования энергопотребления с ошибкой 3,61 МВт·ч (1,5 % от среднемесячного потребления), что достаточно для оперативного планирования энергоресурсов.

4) Результаты согласуются с данными о достигнутом снижении энергопотребления ОАО «МАЗ» на 6,9 % в H1 2025 г. и подтверждают эффективность внедрения систем мониторинга и IoT-технологий.

Ограничением исследования является использование синтетических данных.

Вместе с тем, параметры генерации откалиброваны по открытым отчётным данным предприятия и отраслевой статистике, что обеспечивает реалистичность результатов [10, 11].

Заключение. В данной работе предложена и экспериментально апробирована методология комплексного статистического анализа производственных данных промышленного предприятия ОАО «МАЗ». Основные результаты:

– корреляционный анализ восьми производственных показателей выявил группы тесно связанных признаков (r до 0,99), обосновывающие необходимость снижения размерности;

– метод главных компонент обеспечил сокращение размерности на 75 % (с 8 до 2 компонент) при сохранении 82,35 % информации;

– кластерный анализ идентифицировал три режима работы предприятия, отражающих динамику модернизации за 2021–2025 гг.;

– ансамблевый метод Gradient Boosting показал наивысшую точность прогнозирования энергопотребления (R^2 = 0,873, MAE = 3,61 МВт·ч), превзойдя линейные модели на 10–14 %;

– предложенный аналитический конвейер является масштабируемым и может быть адаптирован для других предприятий машиностроительного комплекса Республики Беларусь.

Перспективным направлением является интеграция методов глубокого обучения (LSTM-сетей) для прогнозирования временных рядов энергопотребления, а также подключение данных IoT-датчиков для анализа в режиме реального времени.

Список литературы

[1] Ассоциация больших данных; Б1; TAdviser. Исследование рынка больших данных и искусственного интеллекта в России, 2025 [Электронный ресурс]. URL: https://b1.ru/insights/news/b1-materials-in-media/b1-kommersant-big-data-survey-13-november-2025/ (дата обращения: 02.03.2026).

[2] Ассоциация больших данных; Б1; TAdviser. Рынок больших данных и искусственного интеллекта в России 2025 [Электронный ресурс]. URL: https://www.tadviser.ru/index.php/Статья:Исследование_рынка_Больших_данных_и_Искусственного_интеллекта_в_России (дата обращения: 02.03.2026).

[3] Попов В.В. Применение искусственного интеллекта и больших данных в практике российских организаций. Прикладная статистика и искусственный интеллект. 2024. № 4. С. 47–60.

[4] Вишняков В.А. Specialized IoT systems: Models, Structures, Algorithms, Hardware, Software Tools: монография. Минск: БГУИР, 2023. 184 с.

[5] Домакур О.В. Методы анализа больших данных. Веснік сувязі. 2020. № 6(164). С. 50–55.

[6] Официальный сайт ОАО «МАЗ». Отчетность ОАО «МАЗ» – управляющая компания холдинга «БЕЛАВТОМАЗ» [Электронный ресурс]. URL: https://maz.by/about/reporting (дата обращения: 10.03.2026).

[7] Сташевская М.П. Анализ применения больших данных в Республике Беларусь в контексте перехода к цифровой экономике. Экономическая наука сегодня: сб. науч. ст. Минск: БНТУ, 2024. Вып. 19. С. 70–78.

[8] Montgomery D.C., Peck E.A., Vining G.G. Introduction to Linear Regression Analysis. 5th ed. Hoboken: Wiley, 2012. 672 p.

[9] Кузнецов В.О. Использование метода главных компонент для анализа надежности цепей поставок. Логистика и управление цепями поставок. 2018. № 4(87). С. 27–33.

[10] Измайлова М.В. Факторный и кластерный анализ основных показателей производственной деятельности предприятий промышленности и транспортного комплекса. Российское предпринимательство. 2012. № 24. С. 52–60 [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/faktornyy-i-klasternyy-analiz-osnovnyh-pokazateley-proizvodstvennoy-deyatelnosti-predpriyatiy-promyshlennosti-i-transportnogo (дата обращения: 15.03.2026).

[11] Волошин Т.А., Зайцев К.С., Дунаев М.Е. Применение адаптивных ансамблей методов машинного обучения к задаче прогнозирования временных рядов. International Journal of Open Information Technologies. 2023. Т. 11. № 5. С. 32–41.

Авторский вклад

Полоско Екатерина Ивановна – постановка задачи исследования, разработка методологии комплексного статистического анализа, генерация имитированных данных, реализация аналитического конвейера в среде Python, интерпретация результатов, подготовка текста статьи.

COMPLEX STATISTICAL ANALYSIS OF BIG DATA: INTEGRATION OF DIMENSIONALITY REDUCTION, CLUSTERING AND REGRESSION METHODS

E.I. Polosko

Senior Lecturer, Department of Economic Informatics, BSUIR

e.i.polosko@gmail.com

Abstract. The article presents a methodology for complex statistical analysis that combines descriptive statistics, correlation analysis, dimensionality reduction (Principal Component Analysis), clustering (K-Means) and regression modelling for big data processing. A step-by-step data analysis approach is used, implemented in Python and tested on simulated production and energy data of OJSC “MAZ” for 2021–2025. A comparative analysis of regression models for forecasting energy consumption is carried out; ensemble methods (Gradient Boosting) demonstrate the highest accuracy (R^2 = 0,873 и MAE = 3,61 МВт·ч) compared to linear models. The practical significance of the study lies in a universal approach to the analysis of production data to support managerial decision-making and optimize business processes at industrial enterprises.

Keywords: big data, complex statistical analysis, OJSC “MAZ”, principal component analysis, cluster analysis, regression modelling, energy efficiency, mechanical engineering, predictive analytics, Python.