

Operational Forecasting of Wind Turbine Power Generation using Clustering and Anomaly Detection

¹Matrenin P.V., ²Khamitov R.N.

¹Ural Federal University, Ekaterinburg, Russian Federation

²Industrial University of Tyumen, Tyumen, Russia Federation

Abstract. Operational forecasting of wind turbine power output is a critical task for power systems with a high share of renewable energy sources, as the accuracy of short-horizon power estimation directly affects system stability, the demand for balancing resources, and overall economic performance. A significant challenge in developing models based is the presence of anomalies, measurement distortions, and pronounced heterogeneity of wind turbine operating regimes, which degrades the performance of unified forecasting approaches. The objective of this study is to develop and analyze an approach to operational wind turbine power forecasting based with explicit consideration of anomalies and operating regimes. A two-stage method is proposed, including anomaly detection and clustering using density-based algorithms, followed by the construction of separate regression models for the identified clusters, which enables accounting for operational heterogeneity and can reduce forecasting errors in specific operating regimes. The key result is the demonstrated dependence of the effectiveness of cluster-oriented modeling on the expressive capacity of the underlying regression model. For models with limited flexibility, accounting for operating regimes leads to a substantial reduction in typical prediction error under high-power operating conditions, whereas for highly expressive models a unified approach provides comparable or superior performance. The practical relevance of the proposed approach lies in its applicability to operational wind power forecasting assuming the availability of short-horizon wind speed forecasts, as well as in supporting data quality assessment and analysis of wind turbine operating regimes, thereby improving the reliability and efficiency of wind energy system operation.

Keywords: wind turbine; power forecasting; short-term forecast; SCADA data; clustering; DBSCAN; decision tree ensembles; anomaly detection

DOI: <https://doi.org/10.52254/1857-0070.2026.1-69.12>

UDC: 621.311.24:004.852

Proгноза операціональна а генерації де енергії електричної а турбін еолієних використовуючи кластеризацію і виявлення аномалій

¹Matrenin P.V., ²Khamitov R.N.

¹Уральський федеральний університет, Єкатеринбург, Російська Федерація

²Індустріальний університет Тюме́ні, Тюмень, Російська Федерація

Rezumat. Прогноза операціональна а потужності турбін еолієних є важливою задачею для систем енергетики з високим частком відновлюваних джерел енергії, оскільки точність оцінки потужності на короткому горизонті безпосередньо впливає на стабільність системи, потребу в ресурсах для балансування і загальну економічну ефективність. Основним викликом при розробці моделей є наявність аномалій, спотворень вимірювань і вираженої гетерогенності режимів роботи турбін еолієних, що погіршує ефективність єдиних підходів до прогнозування. Мета дослідження – розробити і проаналізувати підхід до операціонального прогнозування потужності турбін еолієних з явною увагою до аномалій і режимів роботи. Пропонується двоступінчастий метод, що включає виявлення аномалій і кластеризацію за допомогою алгоритмів на основі густоти, за яким слідує побудова окремих моделей регресії для виявлених кластерів, що дозволяє врахувати операціональну гетерогенність і знизити помилки прогнозування в певних режимах роботи. Основним результатом є демонстрація залежності ефективності моделювання, орієнтованого на кластери, від виразності базової моделі регресії. Для моделей з обмеженою гнучкістю врахування режимів роботи призводить до суттєвого зменшення типової помилки прогнозування в умовах високої потужності, тоді як для висловлюваних моделей єдиний підхід забезпечує порівняльну або кращу ефективність. Практичне значення запропонованого підходу полягає в його застосуванні до операціонального прогнозування потужності вітряків при наявності короткогоризонтних прогнозів швидкості вітру, а також в підтримці оцінки якості даних і аналізу режимів роботи вітряків, що покращує надійність і ефективність експлуатації систем вітряної енергетики.

precum și în susținerea evaluării calității datelor și a analizei regimurilor de funcționare ale turbinelor eoliene, îmbunătățind astfel fiabilitatea și eficiența funcționării sistemului de energie eoliană.

Cuvinte-cheie: turbină eoliană; prognoză energetică; prognoză pe termen scurt; date SCADA; grupare; DBSCAN; ansambluri de arbori decizionali; detectarea anomaliilor.

Оперативное прогнозирование выработки ветроэнергетических установок с использованием кластеризации и детектирования выбросов

¹Матренин П.В., ²Хамитов Р.Н.

¹Уральский федеральный университет, Екатеринбург, Российская Федерация

²Тюменский индустриальный университет, Тюмень, Российская Федерация

Аннотация. Оперативное прогнозирование выработки ветроэнергетических установок является важной задачей для электроэнергетических систем с высокой долей возобновляемых источников энергии, поскольку точность оценки мощности на оперативном горизонте непосредственно влияет на устойчивость энергосистемы, потребность в балансирующих ресурсах и экономические показатели эксплуатации. Важной задачей при построении прогнозных моделей является наличие в данных выбросов, искажений измерений и выраженная неоднородность режимов работы ветроустановок. Целью настоящего исследования является разработка и анализ подхода к оперативному прогнозированию выработки ветроэнергетических установок по с учетом выбросов и эксплуатационных режимов работы оборудования. В работе предложен двухэтапный метод, включающий детектирование выбросов и кластеризацию эксплуатационных состояний на основе плотностных алгоритмов, а также построение отдельных регрессионных моделей для выделенных кластеров режимов, что позволяет учитывать неоднородность эксплуатации и снижать ошибку прогнозирования в отдельных режимах работы ветроустановки. Наиболее значимым результатом исследования является установленная зависимость эффективности кластер-ориентированного моделирования от аппроксимирующей способности используемой модели. Установлено, что для моделей с ограниченной обобщающей способностью учет режимов работы позволяет существенно снизить типичную ошибку прогноза в режимах высокой мощности, тогда как для моделей с высокой обобщающей способностью единая модель демонстрирует сопоставимое или лучшее качество по интегральным метрикам. Дополнительно показано, что кластеризация эксплуатационных режимов может использоваться для автоматического выявления выбросов и нештатных состояний. Практическая значимость работы заключается в возможности применения предложенного подхода для оперативного прогнозирования выработки ветроустановок при наличии прогноза метеорологических параметров на малом горизонте, а также для анализа качества данных и режимов эксплуатации оборудования, что повышает надежность и эффективность управления ветроэнергетическими объектами.

Ключевые слова: ветроэнергетическая установка, ветроэлектрическая станция, прогнозирование генерации, кластеризация, DBSCAN, детектирование выбросов.

ВВЕДЕНИЕ

Прогнозирование генерации ветроэлектрических станций (ВЭС) является важной задачей для электроэнергетических систем с высокой долей возобновляемых источников энергии (ВИЭ) из-за сложности учета стохастической генерации в режимно-балансовых расчетах [1–3]. Теоретическая кривая мощности ветроэнергетической установки (ВЭУ) представляет собой модельную зависимость выходной мощности ВЭУ от скорости ветра при стандартных условиях. Однако данная кривая не учитывает турбулентность, инерционные факторы, топографию, деградацию оборудования и алгоритмы управления ВЭУ, вследствие чего прогнозы на ее основе могут содержать систематические ошибки [4, 5].

В этой связи наблюдается перенос внимания исследователей с прогнозирования скорости и направления ветра на прогнозирование

выработки ВЭУ и ВЭС, что позволяет адаптировать модель к фактическому поведению конкретной установки в конкретных условиях эксплуатации [6, 7].

В настоящей работе предполагается, что на оперативном горизонте прогноз скорости ветра доступен с приемлемой точностью, и задача прогнозирования выработки ВЭУ сводится к моделированию зависимости выходной мощности от текущего или прогнозируемого состояния.

Качество данных является ключевым узким местом [8, 9]. Реальные массивы содержат выбросы и ошибочные записи, возникающие из-за отказов датчиков, сбоях связи, обледенения, оперативных ограничений выработки и иных факторов. Если такие аномалии не устранять до обучения прогнозной модели, они снижают точность прогноза, что показано в ряде исследований [10–12].

Среди методов автоматической очистки данных особое место занимает алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN выделяет плотные области, соответствующие типичному режиму работы ВЭУ, и определяет разреженные изолированные точки как шум (выбросы), не предполагая заранее форму кривой мощности [8–10].

Помимо DBSCAN применяются и другие алгоритмы кластеризации, чаще всего k-средних, а также гибридные методы [10, 13, 14]. Кластеризация используется не только для очистки данных, но и для сегментации режимов работы ВЭУ с последующим обучением прогнозных моделей. Вместо одной глобальной модели данные делятся на более однородные группы (по условиям ветра, состоянию турбины, погодным сценариям), и для каждой группы обучается отдельная модель, что позволяет точнее отразить нелинейные и нестационарные зависимости при условии адекватного выбора модели и достаточной однородности выделенных кластеров

В работе [14] кластеризация выполняется на уровне отдельных ВЭУ, то есть кластером является множество установок. В исследовании [15] рассматриваются кластеры дней, схожих по метеорологическим условиям, для формирования отдельных прогнозных моделей; аналогичный подход представлен в статье [16], где также использован алгоритм k-средних. В работе [17] предложено использование как метеорологических условий, так и значений выработки для кластеризации, с применением обратной ковариационной матрицы Теплица для поиска схожих фрагментов временных рядов.

Еще одним направлением применения кластеризации в рассматриваемой задаче является восстановление пропущенных значений метеорологических условий и данных ВЭС [18].

Заметный прогресс демонстрируют гибридные физико-статистические модели, совмещающие детерминированные математические модели ВЭУ с гибкостью моделей машинного обучения. Включение в вектор признаков теоретической мощности, рассчитанной по кривой зависимости выработки от скорости ветра, позволяет повысить точность прогноза генерации [4, 7, 18–20].

Другим принципиальным аспектом оперативного и краткосрочного прогноза выработки ВЭУ является учет инерционности ветряной

турбины. Поскольку отклик мощности не мгновенен, информативным признаком оказывается «предыдущая мощность» (мощность на предыдущем шаге, например за 10 минут до текущего момента) [21–25].

В данной статье в задачу оперативного прогнозирования мощности ВЭУ внесен следующий вклад:

- предложено использование кластеризации на уровне отдельных записей набора данных, а не более крупных интервалов, как в аналогичных исследованиях;
- на основе DBSCAN разработан алгоритм одновременной кластеризации для применения нескольких отдельных прогнозных моделей и детектирования выбросов для повышения робастности моделей;
- исследовано влияние на точность прогнозирования выработки ВЭУ расширенного пространства признаков, включающего в себя расчетное значение мощности, а также предыдущие измеренные значения выходной мощности ВЭУ для учета инерции;
- исследованы условия применимости кластер-ориентированного подхода к прогнозированию выработки ВЭУ и показана зависимость эффективности кластеризации от аппроксимирующей способности используемой модели.

I. МЕТОДЫ ИССЛЕДОВАНИЯ

Для решения задачи оперативного прогнозирования выработки ветроэнергетических установок в работе предложен метод, включающий следующие этапы:

- 1) формирование и преобразование вектора признаков;
- 2) детектирование выбросов в данных с использованием плотностного алгоритма кластеризации DBSCAN;
- 3) кластеризация эксплуатационных режимов работы ВЭУ на очищенных данных;
- 4) построение и сравнение прогнозных моделей для различных режимов работы и единой модели без учета кластеризации;
- 5) оценка качества моделей с использованием интегральных и условных метрик.

А. Формирование и преобразование признаков

Вектор используемых признаков включает: текущую скорость ветра (м/с); синус и косинус направления ветра; синус и косинус часа суток; синус и косинус номера месяца; теоре-

тическую (расчетную) мощность ВЭУ при текущей скорости ветра (кВт); предыдущее измеренное значение выходной мощности ВЭУ; разность расчетной и измеренной мощностей ВЭУ.

Применение тригонометрических преобразований позволяет корректно учитывать циклическую природу направления ветра и временных признаков.

Использование расчетной мощности, определяемой по теоретической кривой, позволяет включить в модель физическую информацию о поведении установки, а использование предыдущего значения мощности – учесть инерционность ветряной турбины и запаздывание отклика на изменение скорости ветра.

Принципиально важным элементом используемого пространства признаков является включение признака, отражающего согласованность между расчетной и фактической выработкой ВЭУ. Существенное расхождение между расчетной и измеренной мощностью в большинстве случаев свидетельствует не о штатном режиме работы установки, а о наличии технологических ограничений, нештатных состояний или ошибок измерений. Использование данного признака позволяет алгоритмам плотностной кластеризации эффективно отделять физически допустимые режимы работы от аномальных наблюдений без введения жестких пороговых критериев.

Численные признаки подвергаются стандартизации по формуле

$$F' = \frac{F - \text{mean}(F)}{\text{std}(F)},$$

где F – вектор значений признака в наборе данных, mean – функция вычисления среднего значения, std – функция вычисления среднеквадратического отклонения.

В. Детектирование выбросов на основе DBSCAN

На первом этапе обработки данных применяется алгоритм DBSCAN для выявления выбросов и аномальных записей в данных. Алгоритм используется исключительно для очистки данных и не применяется для сегментации режимов прогнозирования.

Множество записей (точек в терминах кластеризации) $X = \{x_1, \dots, x_n\}$ рассматривается в пространстве признаков, и для каждой точки x_i определяется ε -окрестность

$$N_\varepsilon(x_i) = \{x_j \in X \mid d(x_i, x_j) \leq \varepsilon\}.$$

Если $N_\varepsilon(x_i) \geq \min_{pts}$, то точка x_i относится к плотной области данных; точки, не принадлежащие ни одной плотной области, рассматриваются как выбросы. Такой подход позволяет автоматически выявлять ошибочные измерения, связанные с отказами датчиков, ограничениями выработки и иными нештатными ситуациями, без задания априорной формы зависимости мощности от скорости ветра.

Параметры DBSCAN выбираются на основе анализа плотностной структуры данных. Значение \min_{pts} задается с учетом размерности пространства признаков и объема выборки, а параметр ε определяется по k -distance графику как характерный масштаб плотных областей данных.

Записи, идентифицированные как выбросы, исключаются из дальнейшего анализа и обучения прогнозных моделей.

С. Кластеризация эксплуатационных режимов

После удаления выбросов выполняется кластеризация эксплуатационных режимов работы ВЭУ на очищенном наборе данных с использованием алгоритма HDBSCAN (Hierarchical DBSCAN) [26]. В отличие от DBSCAN, применяемого на предыдущем этапе для детектирования выбросов, HDBSCAN позволяет выявлять устойчивые плотностные структуры в данных без предварительного задания масштаба кластеризации и автоматически определять число кластеров.

Кластеризация проводится на уровне отдельных записей данных в пространстве признаков, сформированном на первом этапе, без использования отклонения от теоретической кривой мощности, что позволяет интерпретировать полученные кластеры как различные эксплуатационные режимы работы ветроустановки. Записи, не отнесенные HDBSCAN к кластерам, рассматриваются как переходные состояния между режимами.

В рамках исследования предполагается, что кластеризация эксплуатационных режимов выполняется на накопленном массиве данных за полный год эксплуатации, после чего полученные метки режимов фиксируются и используются на этапе обучения и сравнения прогнозных моделей.

D. Построение прогнозных моделей

Для оценки влияния учета эксплуатационных режимов рассматриваются два варианта моделирования:

- единая модель, обучаемая на всем очищенном наборе данных без учета кластеризации;
- кластер-ориентированный подход, при котором для каждого выделенного эксплуатационного режима обучается отдельная регрессионная модель.

В качестве базовых регрессионных моделей исследуются различные классы алгоритмов. Если для некоторого режима объем обучающей выборки оказывается недостаточным, применяется единая модель.

E. Оценка качества моделей

Оценка качества прогнозирования проводится с использованием сгруппированной кросс-валидации по суткам, при которой все записи, относящиеся к одному календарному дню, целиком попадают либо в обучающую, либо в тестовую выборку. Такой подход снижает влияние автокорреляции временных рядов и соответствует сценарию оперативного применения модели при наличии прогноза метеорологических параметров.

Качество моделей оценивается с использованием общепринятых интегральных метрик (RMSE, MAE, R^2).

II. РЕЗУЛЬТАТЫ

В данном разделе анализируется, в каких условиях учет эксплуатационных режимов и предварительная плотностная очистка данных действительно улучшают качество оперативного прогнозирования выработки ВЭУ.

В исследовании использованы измерения системы сбора данных реальной ветроэнергетической установки из открытого источника [27]. Шаг дискретизации данных составляет 10 минут, и данные содержат 50 530 измерений за период с 1 января 2018 г. 00:00 до 31 декабря 2018 г. 23:50 (количество пропущенных значений составляет 3,86 %).

A. Детектирование выбросов и профили аномалий

Применение алгоритма DBSCAN выявило несколько устойчивых типов аномальных наблюдений, что позволяет рассматривать плотностную структуру данных как информа-

тивный источник для автоматической идентификации нештатных режимов работы ветроустановки.

Агрегированные профили аномалий (Таблица 1) позволяют выделить типовые сценарии несогласованности между расчетной и фактической выработкой, что упрощает интерпретацию природы выявленных выбросов и их связь с эксплуатационными и измерительными факторами.

Использованы следующие обозначения: V_{med} – медианная скорость ветра, T_{med} – медианная расчетная мощность ВЭУ, P_{med} – медианная фактическая мощность ВЭУ, ΔP_{med} – отклонение фактической скорости от расчетной.

Таблица 1
Профили выделенных аномалий
Table 1
Profiles of anomalies detected

ID	Доля Frac- tion, %	V_{med} , м/с, m/s	T_{med} , кВт, kW	P_{med} , кВт, kW	ΔP_{med} , кВт, kW
A-1	2.72	9.74	2622	84	-1053
A1	0.21	10.53	3078	0	-3031
A2	0.12	14.44	3600	0	-3600
A3	0.06	12.66	3588	358	-3213
A4	0.09	13.64	3600	0	-3600
A5	0.06	12.82	3600	0	-3600

Как видно из Рисунка 1, аномальные наблюдения формируют разреженные области в пространстве признаков, пространственно отделенные от основной плотностной структуры данных (для отображения данных на плоскости применен алгоритм UMAP [28]).

Рисунок 2 показывает примеры выявленных аномалий.

Основная плотностная структура данных (A0) соответствует физически корректным режимам работы ветроустановки, тогда как малые плотностные группы (A1–A5) отражают типовые аномальные состояния, характеризующиеся существенным расхождением между расчетной и фактической мощностью.

Сравнение UMAP-визуализаций (Рисунок 1) показывает, что включение признака ΔP обеспечивает явное пространственное отделение аномальных наблюдений, тогда как его исключение приводит к формированию устойчивых кластеров эксплуатационных режимов. Это подтверждает ключевую роль физически осмысленного признака ΔP для детектирования выбросов и обосновывает раздельное использование пространств признаков для очистки данных и анализа режимов.

Анализ профилей выбросов показывает, что аномальные наблюдения в основном связаны с существенным расхождением между расчетной и фактической мощностью при физически допустимых значениях скорости

ветра. Такие ситуации могут быть интерпретированы как следствие технологических ограничений выработки, нештатных режимов управления или ошибок измерений.

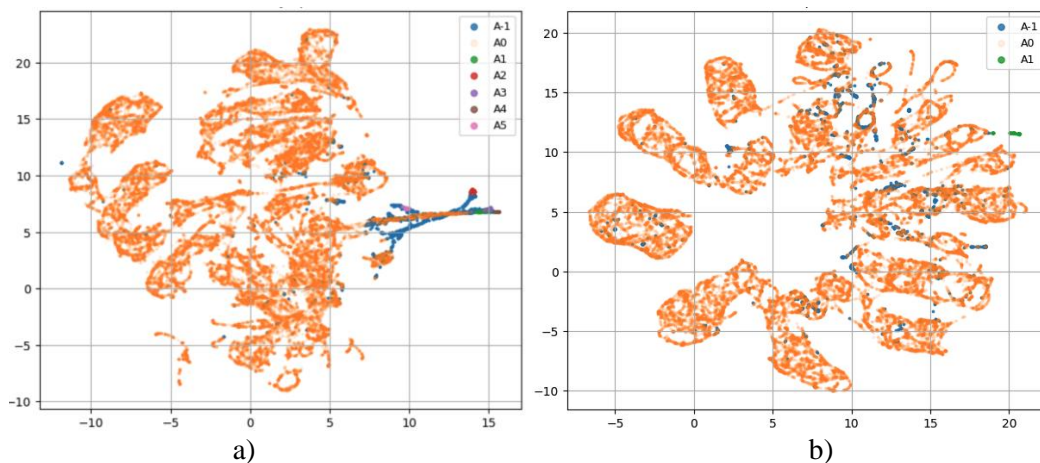


Рис. 1. Визуализация профилей аномалий в пониженном пространстве признаков, а) при использовании ΔP , б) без использования ΔP .

Fig. 1. Visualization of anomaly profiles in reduced feature space а) with ΔP , б) without ΔP .

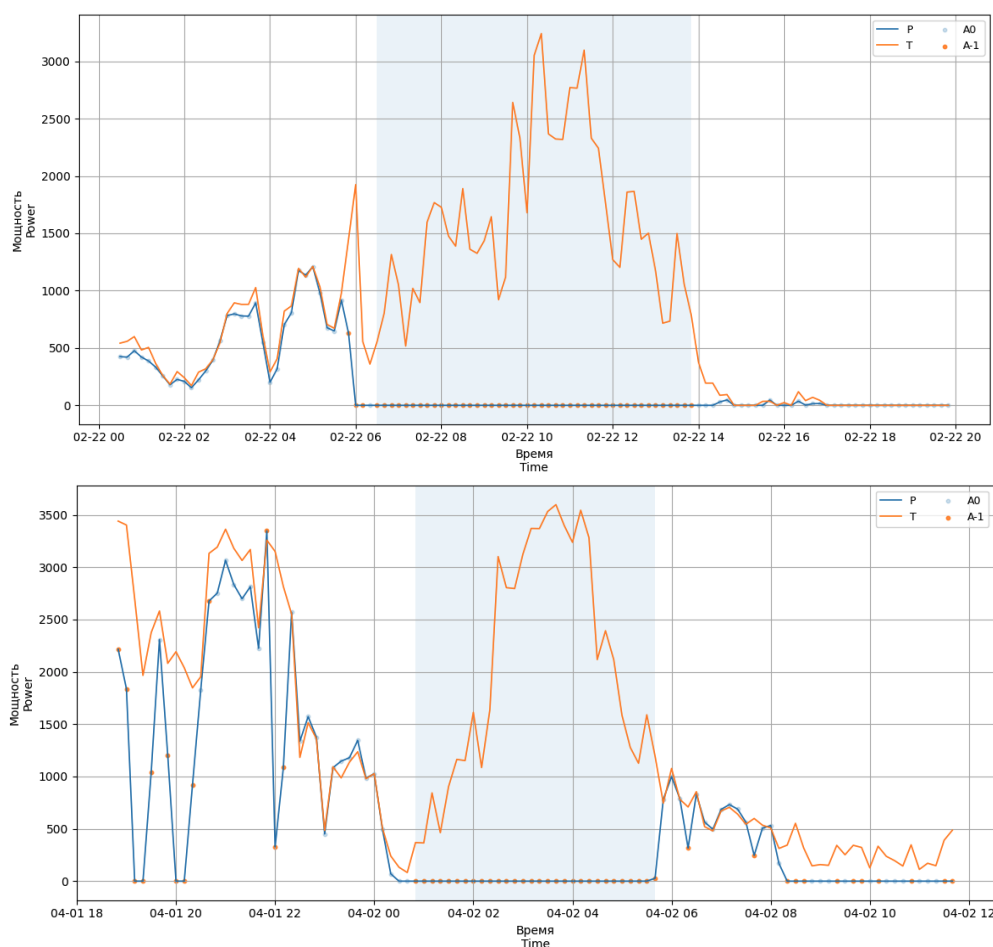


Рис. 2. Примеры выявленных аномалий (P – измеренная мощность, T – расчетная).

Fig. 2. Examples of detected anomalies (P is measured power, T is calculated one).

Ключевым фактором, определяющим принадлежность наблюдений к аномальным, является несогласованность между расчетной и измеренной мощностью, что подтверждает информативность физически обоснованного признака ΔP для автоматического детектирования выбросов.

Во всех выделенных профилях выбросов наблюдается существенное отклонение измеренной мощности от значения, ожидаемого на основе теоретической кривой мощности при заданной скорости ветра. Это подтверждает, что включение расчетной мощности в пространство признаков позволяет использовать физические соотношения в качестве неявного критерия корректности данных и существенно повышает информативность процедуры детектирования выбросов. Как показано в Таблице 1, для всех профилей выбросов характерны значительные отклонения между расчетной и измеренной мощностью при физически допустимых значениях скорости ветра.

Полученные результаты подчеркивают, что эффективность автоматического детектирования выбросов в данных существенно зависит от учета физической природы процесса, а не только от выбора алгоритма машинного обучения.

В. Кластеризация эксплуатационных режимов ветроустановки

После удаления выбросов кластеризация используется как инструмент выявления устойчивых эксплуатационных режимов, отражающих различные физические состояния работы ветроустановки. Для выявления устойчивых режимов эксплуатации применялся алгоритм HDBSCAN, позволяющий автоматически определять число кластеров и выделять плотностные структуры в данных без предварительного задания масштаба кластеризации.

Кластеризация проводилась в пространстве признаков, включающем скорость и направление ветра, временные характеристики и расчетную мощность, без использования отклонения между расчетной и фактической выработкой, что позволило интерпретировать полученные кластеры именно как режимы эксплуатации, а не как проявления аномалий.

В результате применения HDBSCAN в данных были выявлены несколько устойчивых кластеров, соответствующих различным эксплуатационным режимам работы ветроустановки, а также группа наблюдений, не принадлежащих ни одному устойчивому кластеру.

Последние интерпретируются как переходные состояния между режимами и характеризуются повышенной вариабельностью параметров. Для наглядной интерпретации структуры данных на Рисунке 3 представлена визуализация результатов кластеризации в пониженном пространстве признаков, полученном методом UMAP. Видно, что основные режимы эксплуатации формируют компактные и хорошо различимые плотностные области, тогда как переходные состояния располагаются между ними и не образуют самостоятельных кластеров. Для каждого режима были рассчитаны агрегированные профили, включающие типичные значения скорости ветра, расчетной и фактической мощности, а также характерные временные интервалы. Результаты анализа профилей режимов приведены в Таблице 2.

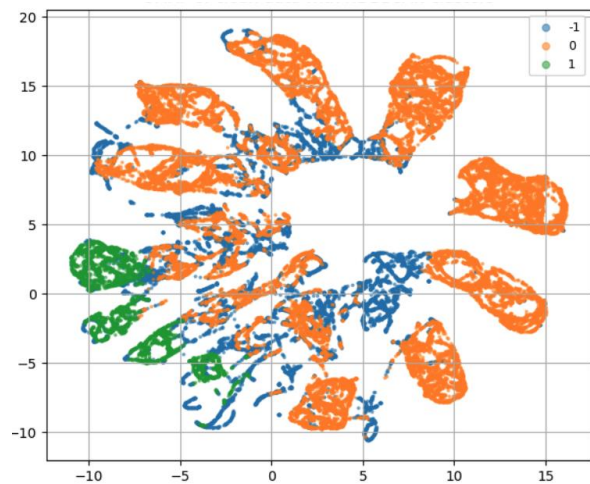


Рис. 3. Визуализация кластеров HDBSCAN.
Fig. 3. Visualization HDBSCAN clusters.

Таблица 2
Режимы работы, определенные HDBSCAN
Table 2
Operating modes defined by HDBSCAN

ID	Доля Fraction, %	V_{med} , м/с, m/s	P_{med} , кВт, kW
R ₋₁	29.52	7.24	841.35
R ₀	62.04	6.35	633.36
R ₁	8.44	13.62	3531.75

Выявленные режимы: R₋₁ – переходные состояния между устойчивыми режимами; R₀ – номинальный режим при умеренной нагрузке; R₁ – режим высокой выработки ВЭУ.

Выделенные режимы обладают выраженной физической интерпретацией и отражают как типичную эксплуатацию ветроустановки,

так и условия повышенной нагрузки, имеющие наибольшее значение для задач оперативного прогнозирования.

Важно подчеркнуть, что кластеризация эксплуатационных режимов выполняется на накопленном массиве данных за полный год эксплуатации и результаты кластеризации фиксируются перед этапом построения прогнозных моделей.

Режим R_1 характеризуется высокими значениями скорости ветра и выработки и соответствует условиям повышенной нагрузки, имеющим наибольшую практическую значимость для задач оперативного прогнозирования. В этом режиме наблюдается повышенная

вариабельность мощности, что отражает влияние инерционных и управляющих факторов на отклик ветроустановки.

Выявленные эксплуатационные режимы используются далее для построения и сравнения прогнозных моделей, обучаемых отдельно для каждого режима и на всем массиве данных без учета кластеризации

С. Анализ влияния кластеризации на точность моделей

Сравнение результатов (Таблица 3, Рисунок 4) показывает, что влияние кластеризации на качество прогнозирования существенно различается для моделей с разной аппроксимирующей способностью.

Таблица 3
Показатели точности различных моделей при использовании единой модели и при построении моделей для каждого кластера (среднее значение \pm СКО)

Table 3
Accuracy indicators of different models when using a single model and when constructing models for each cluster (mead \pm std)

Модель Model	Подход Approach	RMSE, кВт	MAE, кВт	R ²
PL	Одна модель	106.8 \pm 7.4	70.0 \pm 7.7	0.993 \pm 0.001
PL	Кластеризация	103.9 \pm 2.6	66.2 \pm 2.7	0.994 \pm 0.001
SVR	Одна модель	90.8 \pm 6.0	52.4 \pm 3.6	0.995 \pm 0.000
SVR	Кластеризация	89.1 \pm 4.8	51.3 \pm 3.3	0.995 \pm 0.000
MLP	Одна модель	96.9 \pm 7.5	61.3 \pm 5.2	0.994 \pm 0.001
MLP	Кластеризация	117.6 \pm 10.0	74.7 \pm 2.1	0.993 \pm 0.001
XGBoost	Одна модель	86.5 \pm 7.6	49.9 \pm 3.8	0.995 \pm 0.000
XGBoost	Кластеризация	86.3 \pm 7.1	50.2 \pm 2.5	0.995 \pm 0.000
LightGBM	Одна модель	83.3 \pm 8.3	46.3 \pm 2.8	0.996 \pm 0.000
LightGBM	Кластеризация	86.2 \pm 7.3	48.9 \pm 3.5	0.995 \pm 0.000
CatBoost	Одна модель	80.6 \pm 6.6	47.4 \pm 2.6	0.996 \pm 0.000
CatBoost	Кластеризация	84.5 \pm 9.1	50.0 \pm 3.1	0.995 \pm 0.000

Были использованы следующие модели: полиномиальная регрессия (PL); метод опорных векторов (SVR); многослойный перцептрон (MLP); экстремальный градиентный бустинг (XGBoost); быстрый бустинг (LightGBM); категориальный бустинг (CatBoost).

Оценка проводилась с использованием сгруппированной кросс-валидации по суткам. Для каждой модели использовался подбор гиперпараметров методом случайного поиска.

Учет эксплуатационных режимов приводит к снижению интегральной ошибки прогнозирования для полиномиальной регрессии и в меньшей степени для метода опорных векторов.

Для нейронных сетей и ансамблей деревьев решений единая модель демонстрирует сопоставимое или более низкое значение интегральных метрик по сравнению с кластер-ориентированным подходом.

Средние по всем режимам эксплуатации не отражают особенности поведения моделей в наиболее важных условиях работы ВУЭ.

В связи с этим далее проводится анализ качества прогнозирования в режиме высокой вырабатываемой мощности (R_1).

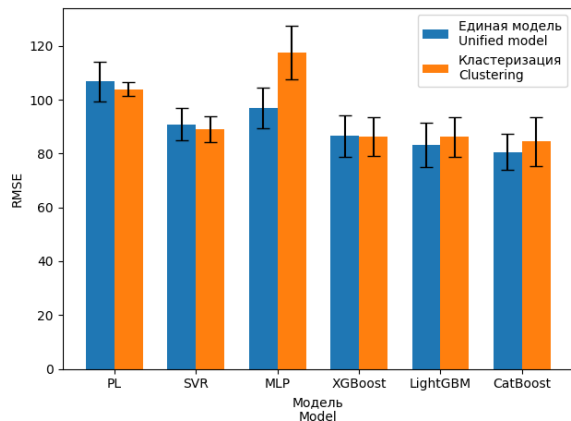


Рис. 4. Результаты единой модели и модели с кластеризацией по различным классам алгоритмов.

Fig. 4. Results for the unified model and the clustering-based model across different algorithm classes.

D. Анализ режима высокой вырабатываемой мощности

Анализ режима высокой выработки R_1 показывает (Таблица 4, Рисунок 5), что кластер-ориентированный подход особенно эффективен для моделей с ограниченной гибкостью, для которых учет режимов позволяет существенно снизить типичную ошибку прогнозирования в наиболее критических условиях эксплуатации. Для метода опорных векторов и CatBoost наблюдается преимущественное снижение MAE при сопоставимых или более высоких значениях RMSE.

Таблица 4
Результаты для режима высокой выработки ВЭУ
Table 4
Results for the high-output mode of the wind turbine

Модель Model	Подход Approach	RMSE, кВт	MAE, кВт
PL	Одна модель One model	97.9 ± 9.3	75.7 ± 8.3
PL	Кластеризация Clasterization	74.8 ± 15.8	53.1 ± 7.1
SVR	Одна модель One model	79.8 ± 8.8	53.6 ± 7.6
SVR	Кластеризация Clasterization	83.1 ± 14.6	49.5 ± 9.9
CatBoost	Одна модель One model	70.5 ± 12.4	43.5 ± 9.0
CatBoost	Кластеризация Clasterization	72.4 ± 12.8	41.3 ± 9.6

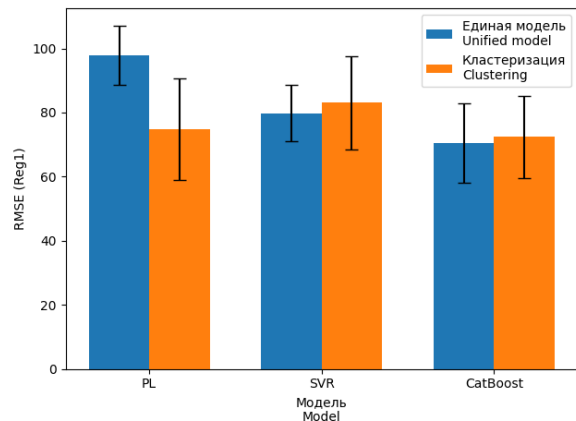


Рис. 5. Результаты для режима высокой выработки.

Fig. 5. Results for the high-output mode.

III. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Полученные результаты позволяют по-новому рассмотреть роль кластеризации в задачах оперативного прогнозирования выработки ветроэнергетических установок, сместив акцент с универсального улучшения метрик на анализ условий ее целесообразности.

Показано, что эффективность кластер-ориентированного подхода существенно зависит от аппроксимирующей способности используемой модели и режима эксплуатации ветроустановки.

В существующих исследованиях, использующих кластеризацию для повышения точности прогнозирования скорости ветра и выработки, не проводится систематический анализ того, в каких условиях кластеризация действительно улучшает прогноз. В настоящей работе данный пробел восполняется за счет одновременного анализа нескольких классов регрессионных моделей и использования как интегральных, так и условных метрик качества.

Принципиальным отличием предложенного подхода является раздельное использование плотностной кластеризации для решения двух различных задач. Алгоритм DBSCAN применяется для детектирования выбросов и очистки данных, тогда как алгоритм HDBSCAN используется для выявления устойчивых эксплуатационных режимов.

Дополнительным элементом новизны является использование физически обоснованного признака согласованности между расчетной и фактической мощностью ветроустановки. Результаты анализа профилей аномалий показывают, что именно несогласованность между расчетной и измеренной мощностью является

ключевым фактором, позволяющим автоматически выявлять ошибочные измерения.

Предложенный подход к детектированию выбросов на основе плотностной кластеризации обладает рядом преимуществ по сравнению с традиционными методами очистки данных. В частности, в отличие от ручной экспертной фильтрации, основанной на заранее заданных правилах и пороговых значениях, не требуется априорное задание или экспертный подбор допустимых диапазонов параметров. Подход не зависит от субъективных представлений эксперта о «нормальном» режиме работы ВЭУ. Это особенно важно в условиях сложных и изменяющихся режимов эксплуатации, когда фиксированные эвристические правила оказываются либо избыточно жесткими, либо недостаточно чувствительными к нетипичным, но физически допустимым состояниям.

По сравнению со статистическими методами выявления выбросов, основанными на анализе распределений отдельных признаков и использовании порогов (например, по стандартному отклонению, межквартильному размаху или квантилям), плотностная кластеризация позволяет учитывать многомерную структуру данных и взаимосвязи между признаками. Статистические подходы, как правило, требуют подбора пороговых значений для каждого признака либо их комбинаций, что затрудняет переносимость метода между различными ветроустановками и наборами данных. В противоположность этому, DBSCAN адаптируется к локальной плотностной структуре данных и позволяют автоматически выделять аномальные наблюдения как разреженные области в пространстве признаков, что повышает воспроизводимость процедуры очистки данных.

Дополнительно следует подчеркнуть, что используемые в работе методы плотностной кластеризации не опираются на предположения о нормальности распределений данных или ошибок. Алгоритмы кластеризации используют исключительно геометрическую и плотностную структуру выборки. Это обстоятельство является принципиальным преимуществом при анализе данных ВЭУ, которые характеризуются выраженной асимметрией распределений и наличием выбросов.

Показано, что кластер-ориентированный подход не является универсальным средством повышения качества модели. Для моделей с

ограниченной аппроксимирующей способностью, таких как полиномиальная регрессия, учет эксплуатационных режимов приводит к устойчивому снижению как интегральных, так и условных ошибок прогнозирования. Для моделей средней сложности (метод опорных векторов) эффект кластеризации проявляется в снижении типичной ошибки в режиме высокой мощности, но носит менее устойчивый характер. В то же время для таких моделей как нейронные сети и ансамбли деревьев решений, единая модель демонстрирует сопоставимое или лучшее качество по интегральным метрикам, что указывает на способность таких алгоритмов эффективно учитывать неоднородность данных без явной сегментации.

Особое значение имеет анализ качества прогнозирования в режиме высокой мощности, который редко рассматривается в существующих публикациях. Показано, что кластер-ориентированный подход может приводить к снижению типичной ошибки в наиболее важном режиме эксплуатации ВЭУ.

Кроме того, даже в тех случаях, когда учет эксплуатационных режимов не приводит к улучшению интегральных метрик точности прогнозирования, кластеризация данных повышает интерпретируемость модели и результатов анализа. Выделение устойчивых режимов эксплуатации позволяет связать поведение модели с физическими состояниями ветроустановки, упростить анализ ошибок прогнозирования и обеспечить более прозрачное объяснение причин расхождений между расчетной и фактической выработкой. Таким образом, кластеризация выступает не только как инструмент потенциального повышения точности, но и как средство структурирования данных и повышения объяснимости моделей машинного обучения, что имеет важное значение для практического применения прогнозных систем в энергетике.

ЗАКЛЮЧЕНИЕ

В работе рассмотрена задача оперативного прогнозирования выработки ветроэнергетических установок с учетом неоднородности эксплуатационных режимов и наличия выбросов в исходных данных. Предложен и экспериментально исследован подход, основанный на раздельном применении плотностных алгоритмов кластеризации для детектирования выбросов и выявления режимов эксплуатации. Показано, что использование алгоритма DBSCAN в

сочетании с физически обоснованным признаком согласованности между расчетной и фактической мощностью позволяет эффективно выявлять и интерпретировать выбросы в данных без введения жестких пороговых критериев. Экспериментальное сравнение различных классов регрессионных моделей показало, что эффективность кластер-ориентированного подхода существенно зависит от аппроксимирующей способности используемого алгоритма. Основным научным результатом работы является не столько демонстрация способа повышения точности прогнозирования, сколько выявление границ применимости кластер-ориентированных моделей в задачах оперативного прогнозирования выработки ветроустановок. Полученные результаты показывают, что эффективность учета эксплуатационных режимов определяется балансом между снижением систематического смещения модели и ростом дисперсии, зависящим от сложности используемого алгоритма. Данный вывод уточняет и дополняет результаты предыдущих исследований и может служить основой для обоснованного выбора архитектуры прогнозных моделей.

Литература (References)

- [1] Arslan Tuncar E., Saglam S., Oral B. A review of short-term wind power generation forecasting methods in recent technological trends. *Energy Reports*, 2024, vol. 12, suppl. 1, pp. 197–209. <https://doi.org/10.1016/j.egyr.2024.06.006>
- [2] Eroshenko S.A, Khalyasmaa A.I. Intelligent model of decision support system of distributed generation integration. *Proc. of the IEEE International Conference on Software Engineering and Service Sciences*, 2017, <https://doi.org/10.1109/ICSESS.2017.8342868>
- [3] Liu Z., Guo H., Zhang Y., Zuo Z. A Comprehensive Review of Wind Power Prediction Based on Machine Learning: Models, Applications, and Challenges. *Energies*, 2025, vol. 18(2), art. 350. <https://doi.org/10.3390/en18020350>
- [4] Lydia M., Kumar S.S., Selvakumar A.I., Kumar G.E.P. A comprehensive review on wind turbine power curve modeling techniques. *Renewable and Sustainable Energy Reviews*, 2014, vol. 30, pp. 452–460. <https://doi.org/10.1016/j.rser.2013.10.030>
- [5] Carta J.A., Velázquez S., Matías J.M. A review of power curve modelling techniques for wind turbines. *Renewable and Sustainable Energy Reviews*, 2013, vol. 21, pp. 572–581. <https://doi.org/10.1016/j.rser.2013.01.032>
- [6] Maldonado-Correa J., Martín-Villate A., de la Torre S.S. Wind power forecasting: A systematic literature review. *Wind Engineering*, 2021, vol. 45(2), pp. 413–426. <https://doi.org/10.1177/0309524X19891672>
- [7] Bazionis I.K., Georgilakis P.S. Review of Deterministic and Probabilistic Wind Power Forecasting: Models, Methods, and Future Research. *Electricity*, 2021, vol. 2(1), pp. 19–39. <https://doi.org/10.3390/electricity2010002>
- [8] Zhao Y., et al. Data-driven correction approach to refine power curve of wind farm under wind curtailment. *IEEE Transactions on Sustainable Energy*, 2018, vol. 9(1), pp. 95–105. <https://doi.org/10.1109/TSTE.2017.2717021>
- [9] Paik C., Chung Y., Kim Y.J. Power Curve Modeling of Wind Turbines through Clustering-Based Outlier Elimination. *Applied System Innovation*, 2023, vol. 6(2), art. 41. <https://doi.org/10.3390/asi6020041>
- [10] Subuh A.A.A., Kaboli S.H.A., Waqar M., Vallée F. Hybrid Model for Cleaning Abnormal Data of Wind Turbine Power Curve Based on Machine Learning Approaches. *e-Prime – Advances in Electrical Engineering, Electronics and Energy*, 2025, vol. 13, art. 101043. <https://doi.org/10.1016/j.prime.2025.101043>
- [11] Pinson P. Wind energy: forecasting challenges for its operational management. *Wind Energy*, 2013, vol. 16(6), pp. 909–926. <https://doi.org/10.1002/we.1538>
- [12] Wan C., et al. Wind power forecasting in China: Status and challenges. *Renewable and Sustainable Energy Reviews*, 2015, vol. 44, pp. 239–248. <https://doi.org/10.1016/j.rser.2014.12.011>
- [13] Mehmood Z., Wang Z. Hybrid iForest-DBSCAN for anomaly detection and wind power curve modelling. *Expert Systems with Applications*, 2025, vol. 289, art. 128381. <https://doi.org/10.1016/j.eswa.2025.128381>
- [14] Hou G., Wang J., Fan Y. Wind power forecasting method of large-scale wind turbine clusters based on DBSCAN clustering and an enhanced hunter-prey optimization algorithm. *Energy Conversion and Management*, 2024, vol. 307, art. 118341. <https://doi.org/10.1016/j.enconman.2024.118341>
- [15] Dong L., et al. Wind power day-ahead prediction with cluster analysis of NWP. *Renewable and Sustainable Energy Reviews*, 2016, vol. 60, pp. 1206–1212. <https://doi.org/10.1016/j.rser.2016.01.106>
- [16] Wang K., Qi X., Liu H., Song J. Deep belief network - based k-means cluster approach for short-term wind power forecasting. *Energy*, 2018, vol. 165, pp. 840–852. <https://doi.org/10.1016/j.energy.2018.09.118>
- [17] Fan H., Zhang X., Mei S., Zhang J. A Markov Regime Switching Model for Ultra-Short-Term Wind Power Prediction Based on Toeplitz Inverse Covariance Clustering. *Frontiers in Energy Research*, 2021, vol. 11. <https://doi.org/10.3389/fenrg.2021.638797>

- [18] Karaman Ö.A. Prediction of Wind Power with Machine Learning Models. *Applied Sciences*, 2023, vol. 13(20), art. 11455. <https://doi.org/10.3390/app132011455>
- [19] Shokrzadeh S., Jozani M.J., Bibeau E.L. Wind turbine power curve modeling using advanced parametric and nonparametric methods. *IEEE Transactions on Sustainable Energy*, 2014, vol. 5(4), pp. 1262–1269. <https://doi.org/10.1109/TSTE.2014.2345057>
- [20] Xu W., Chen X., Kang C. Short-term wind power forecasting using variational mode decomposition and deep learning. *Applied Energy*, 2018, vol. 230, pp. 714–732. <https://doi.org/10.1016/j.apenergy.2018.09.041>
- [21] Singh U., Rizwan M. SCADA system dataset exploration and machine learning based forecast for wind turbines. *Results in Engineering*, 2022, vol. 16, art. 100640. <https://doi.org/10.1016/j.rineng.2022.100640>
- [22] Krevnevičiūtė J., et al. The Forecast of the Wind Turbine Generated Power Using Hybrid (LTC + XGBoost) Model. *Applied Sciences*, 2025, vol. 15(13), art. 7615. <https://doi.org/10.3390/app15137615>
- [23] Kusiak J., Zheng H., Song Z. Short-term prediction of wind farm power: a data mining approach. *IEEE Transactions on Energy Conversion*, 2009, vol. 24(1), pp. 125–136. <https://doi.org/10.1109/TEC.2008.2006552>
- [24] Foley A.M., Leahy P.G., Marvuglia A., McKeogh E.J. Current methods and advances in wind power forecasting. *Renewable Energy*, 2012, vol. 37(1), pp. 1–8. <https://doi.org/10.1016/j.renene.2011.05.033>
- [25] Lei M., Shiyan L., Chuanwen J., Ying L., Zhao X. A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, 2009, vol. 13(4), pp. 915–920. <https://doi.org/10.1016/j.rser.2008.02.002>
- [26] Campello R., Moulavi D., Zimek A., Sander J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data*, 2015, vol. 10(1), pp. 1–51. <https://doi.org/10.1145/2733381>
- [27] Wind Turbine Scada Dataset. Available online: <https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset> (accessed on 20 December 2025).
- [28] McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*, 2018. Available online: <https://arxiv.org/abs/1802.03426> (accessed on 20 October 2025).

Сведения об авторах.



Матренин Павел Викторович, к.т.н.

Область научных интересов: системный анализ, искусственный интеллект в энергетике, методы обработки и анализа данных, прогнозирование временных рядов, объяснимый искусственный интеллект.

E-mail: p.v.matrenin@urfu.ru

ORCID: 0000-0001-5704-0976



Хамитов Рустам

Нуриманович, д.т.н., п

Область научных интересов: прогнозирование электропотребления, распределенная генерация, возобновляемые источники энергии, асинхронные электрические двигатели

E-mail: apple_27@mail.ru

ORCID: 0000-0001-9876-5471