

УДК 519.876.5

Методы прогнозирования энергопотребления центров обработки данных на основе анализа временных рядов

Мироненко Ирина Александровна

Кузьмина Инна Анатольевна (*научный руководитель*)

МГТУ им. Н.Э. Баумана, Москва, Россия

Представлен сравнительный анализ методов прогнозирования энергопотребления центров обработки данных, где высокая точность прогноза критически важна для эффективного управления ресурсами и минимизации эксплуатационных затрат на основе анализа временных рядов. Рассмотрены регрессионные модели временных рядов, включая авторегрессию, скользящее среднее, ARMA, ARIMA, SARIMA и метод Хольта — Винтерса. Проведена оценка эффективности моделей с использованием метрики MASE (Mean Absolute Scaled Error) и разработанной асимметричной метрики ошибки, учитывающей специфику задачи энергопотребления.

Ключевые слова: прогнозирование энергопотребления, временные ряды, регрессионные модели, ARIMA, метрика MASE, асимметричная метрика ошибки, центры обработки данных

Введение. Эффективное управление энергопотреблением центров обработки данных (ЦОД) является критически важной задачей как с экономической, так и с экологической точки зрения. Точное прогнозирование энергопотребления позволяет оптимизировать нагрузку, планировать ресурсы и предотвращать аварийные ситуации. В данной работе исследуются различные методы прогнозирования временных рядов с целью определения наиболее эффективных подходов для решения задачи прогнозирования энергопотребления ЦОД.

Обоснование выбора методов исследования. Для решения задачи прогнозирования энергопотребления ЦОД существует множество методов, среди которых можно выделить следующие группы: статистические методы (регрессионные модели, экспоненциальное сглаживание), методы машинного обучения (деревья решений, случайные леса, градиентный бустинг) и нейросетевые модели (рекуррентные нейронные сети, LSTM, GRU, Transformer).

В рамках данного исследования было принято решение сосредоточиться на группе регрессионных моделей временных рядов, включая авторегрессию (AR), скользящее среднее (MA), ARMA, ARIMA и SARIMA. Данное решение обусловлено следующими факторами.

Интерпретируемость моделей. В отличие от нейросетевых моделей, регрессионные модели временных рядов позволяют однозначно определить влияние различных компонентов на конечный результат прогнозирования. Как указывают Hyndman и Athanasopoulos [1], это особенно важно при принятии управленческих решений, когда необходимо понимать причинно-следственные связи.

Возможность адаптации и настройки. Регрессионные модели позволяют легко модифицировать параметры для учета изменений в характере данных. По мнению Brockwell и Davis [2], это критически важно при работе с энергопотреблением ЦОД, которое может меняться в зависимости от множества факторов.

Высокая точность при относительно небольших объемах данных. Согласно исследованиям Abraham и Ledolter [3], регрессионные модели часто показывают высокую эффективность при прогнозировании временных рядов даже на небольших выборках, в то время как нейросетевые модели требуют значительно большего объема обучающих данных.

Вычислительная эффективность. Регрессионные модели требуют меньших вычислительных ресурсов для обучения и применения по сравнению с нейросетями, что отмечают Shumway и Stoffer [4].

Основное концептуальное различие между регрессионными моделями временных рядов и нейросетевыми подходами заключается в природе их работы с данными и интерпретируемости результатов. Нейронные сети часто рассматриваются как «черные ящики», поскольку их внутреннее функционирование сложно для интерпретации [5]. В противоположность этому, регрессионные модели являются прозрачными, позволяя исследователю понимать, какие компоненты временного ряда и с каким весом влияют на прогноз.

Методы и данные исследования. В настоящем исследовании используются данные энергопотребления контроллера SL-202A2-2-2-11, полученные с интервалом в одну секунду за период в два дня, что составляет 65 533 наблюдения. Данные представляют собой временной ряд значений потребления электроэнергии в киловаттах (кВт).

Высокая частота дискретизации (1 с) позволяет детально отслеживать динамику энергопотребления и выявлять как краткосрочные колебания, так и более длительные тренды. Согласно Shabou [6], такая детализация данных создает благоприятные условия для построения точных прогнозных моделей, особенно для регрессионных методов, которые могут эффективно улавливать различные паттерны в данных.

Пример структуры исходных данных представлен на рис. 1.

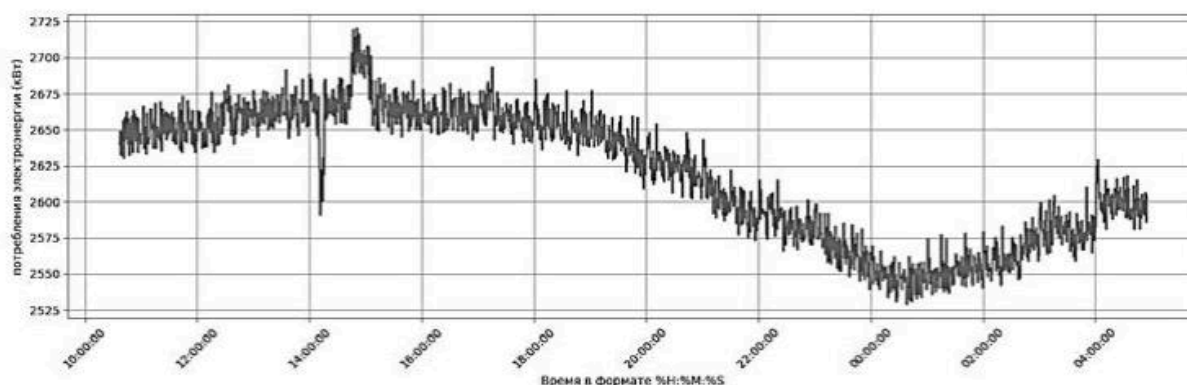


Рис. 1. Структура данных энергопотребления

Перед применением моделей временных рядов необходимо провести предварительную обработку данных, которая включает следующие этапы.

Проверка на стационарность. Для регрессионных моделей временных рядов важно, чтобы данные были стационарными, т. е. их статистические свойства не менялись со временем. Для этой цели используется расширенный тест Дики — Фуллера (ADF), который позволяет определить, является ли временной ряд стационарным.

Дифференцирование. Если тест ADF показывает, что ряд не стационарен, применяется дифференцирование — вычисление разности между последовательными наблюдениями. Это один из способов преобразования нестационарного ряда в стационарный, что особенно важно для моделей ARIMA и SARIMA.

Анализ автокорреляции и частичной автокорреляции. Для определения параметров моделей (порядков AR и MA) проводится анализ функций автокорреляции (ACF) и частичной автокорреляции (PACF).

Графики ACF и PACF для анализируемых данных энергопотребления представлены на рис. 2.

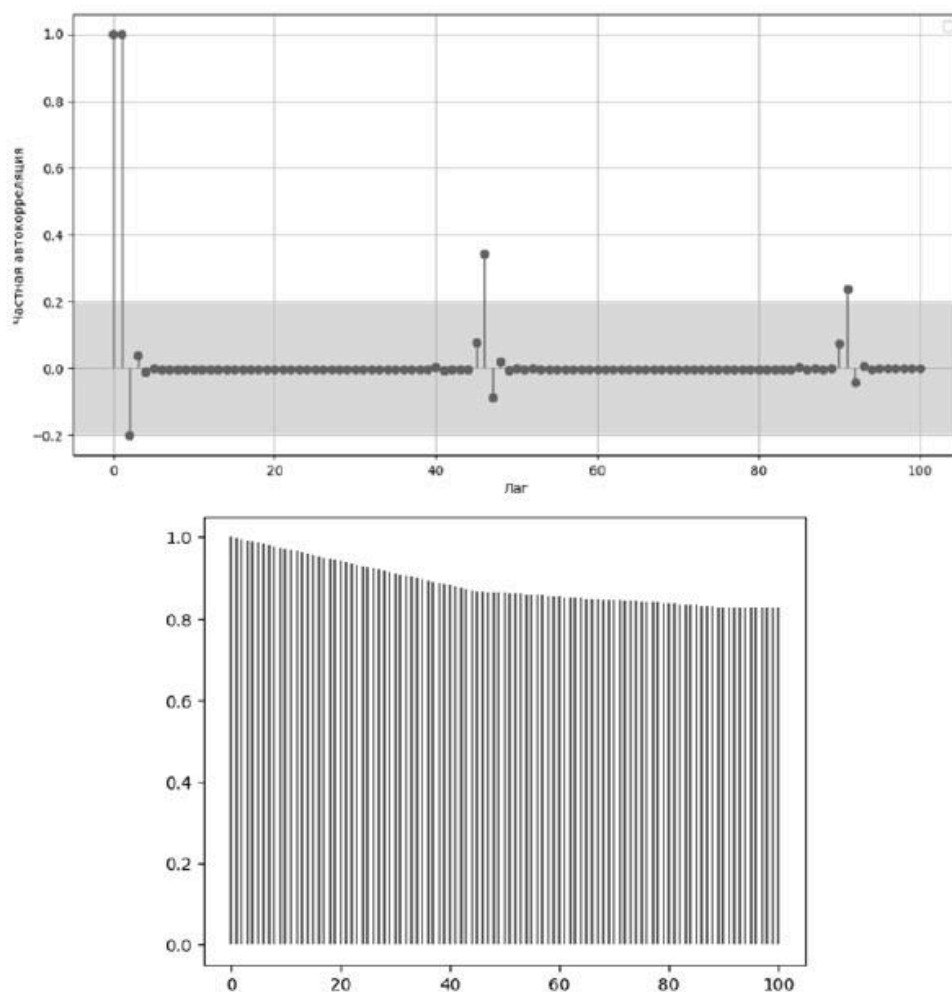


Рис. 2. Автокорреляция (PACF) и частичная автокорреляция (ACF) временного ряда энергопотребления

Для проверки стационарности временного ряда энергопотребления применялся ADF-тест. Из исходного временного ряда случайным образом выбирались подвыборки (окна) произвольной длины, в данном случае — 11 часов. Для каждой такой подвыборки вычислялось значение ADF-статистики. Процедура повторялась для уменьшенных размеров окна, чтобы оценить устойчивость результата.

На основании полученных результатов был построен график (рис. 3), который показал, что во всех случаях ADF-статистика находится ниже критического уровня ($p\text{-value} < 0,05$).

Временной ряд энергопотребления является стационарным, что подтверждает возможность применения регрессионных моделей. Минимальное стационарное окно составляет 540 наблюдений.

Формула оценки имеет вид

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^p \delta_i \Delta y_{t-i} + \epsilon_t,$$

где γ — критический параметр (если $\gamma \approx 0$, ряд нестационарен); p — количество лагов (зависит от n); n — количество наблюдений в окне.

Метрики оценки точности прогнозирования временных рядов. При разработке прогностических моделей для временных рядов критически важно иметь надежные метрики оценки точности, позволяющие объективно сравнивать эффективность различных методов. Как отмечает Hyndman в своей статье [7], существует несколько категорий метрик оценки точности прогнозов.

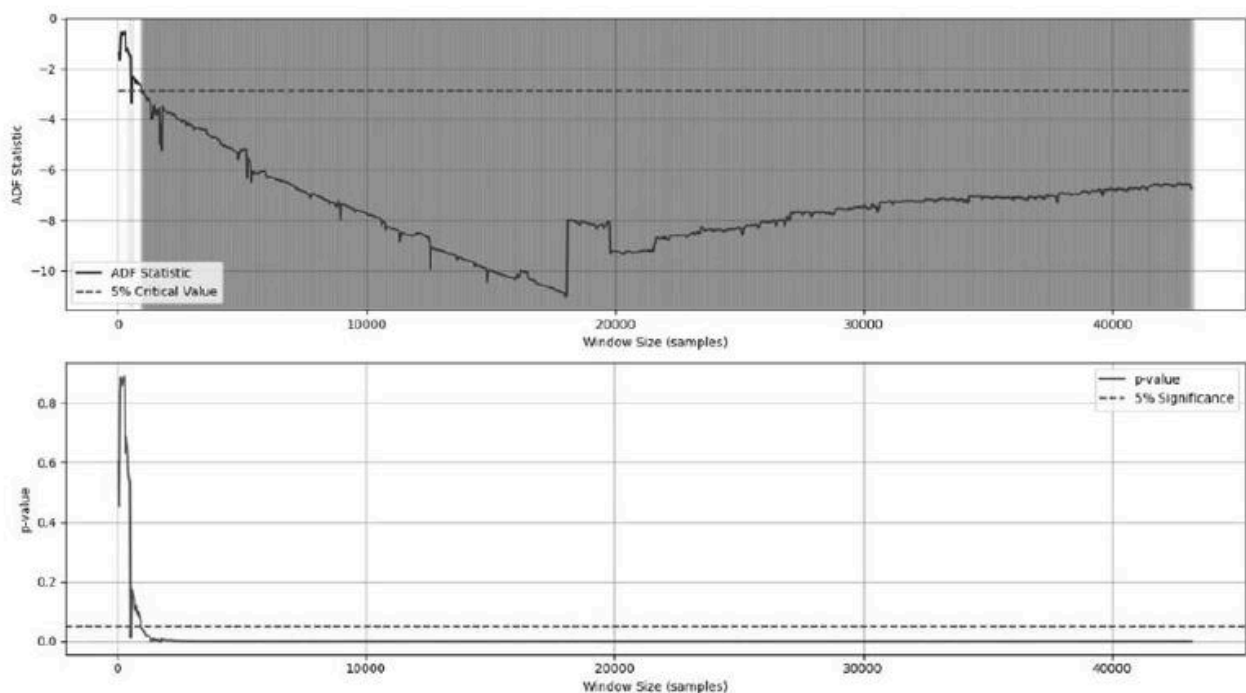


Рис. 3. Динамика ADF-статистики в зависимости от размера скользящего окна

Процентные метрики ошибок — метрики, выражающие ошибку в процентах от фактического значения, например средняя абсолютная процентная ошибка (MAPE).

Масштабно-зависимые метрики — метрики, зависящие от масштаба данных, такие как средняя абсолютная ошибка (MAE) и среднеквадратическая ошибка (MSE).

Относительные метрики ошибок — метрики, сравнивающие ошибки модели с ошибками некоторого базового метода.

Безразмерные метрики ошибок — метрики, не зависящие от масштаба данных.

В рамках данного исследования была выбрана метрика MASE (средняя абсолютная масштабированная ошибка), предложенная Hyndman [7] как универсальное решение для оценки точности прогнозов, свободное от недостатков других метрик:

$$MAPE = mean(|q_t|),$$

где масштабированная ошибка определяется как

$$q_t = \frac{e_t}{(\frac{1}{n-1} * \sum_{i=2}^n |Y_i - Y_{i-1}|)}.$$

Ключевые преимущества MASE, определившие выбор этой метрики.

Независимость от масштаба — MASE не зависит от масштаба данных, что позволяет сравнивать точность прогнозов для разных временных рядов.

Устойчивость к нулевым значениям — MASE корректно работает с прерывистыми временными рядами, где некоторые значения могут быть равны нулю.

Интерпретируемость — $MASE < 1$ означает, что модель прогнозирует лучше, чем наивный метод, $MASE > 1$ — хуже.

Универсальность — MASE может использоваться в различных ситуациях прогнозирования.

Дополнительно была разработана асимметричная метрика ошибки, учитывающая специфику задачи прогнозирования энергопотребления, где перепотребление и недопотребление имеют разные последствия.

Методы прогнозирования временных рядов и их практическая реализация. В рамках исследования были реализованы и протестированы следующие методы прогнозирования временных рядов.

AR (авторегрессия) — метод, использующий линейные зависимости между текущим значением и предыдущими значениями ряда.

MA (скользящее среднее) — метод, представляющий текущее значение как функцию от ошибок прогнозов прошлых периодов.

ARMA (авторегрессия и скользящее среднее) — комбинация методов AR и MA.

ARIMA (авторегрессия и интегрированное скользящее среднее) — расширение ARMA для нестационарных рядов, включающее дифференцирование.

SARIMA (сезонная ARIMA) — расширение ARIMA, учитывающее сезонные паттерны.

Метод Хольта — Винтерса — метод экспоненциального сглаживания, способный учитывать тренд и сезонность.

Для исследования данных методов был использован язык программирования Python с применением библиотек statsmodels, pandas, numpy и matplotlib.

Подбор оптимального размера окна обучения. Одним из ключевых аспектов прогнозирования временных рядов является определение оптимального размера окна обучения — количества исторических точек данных, используемых для обучения модели. В данной работе был использован следующий подход к определению оптимального размера окна.

Определение диапазона возможных размеров окна от 2100 до 8100 точек с шагом 300.

Для каждого размера окна и каждого метода прогнозирования.

Разделение данных на обучающие и тестовые наборы с использованием скользящего окна.

Обучение модели на обучающем наборе.

Прогнозирование на тестовом наборе.

Расчет метрики MASE для оценки точности прогноза.

Выбор размера окна, обеспечивающего минимальное значение MASE для каждого метода.

Результаты анализа показали, что оптимальный размер окна варьируется в зависимости от метода прогнозирования. Наилучшие результаты были получены при следующих размерах окна:

AR: min MASE = 37,1262 при окне 2100,

MA: min MASE = 38,3275 при окне 2100,

ARMA: min MASE = 37,2744 при окне 2100,

ARIMA: min MASE = 37,1069 при окне 2100,

SARIMA: min MASE = 40,7831 при окне 2400,

Holt-Winters: min MASE = 42,4010 при окне 5100.

Разработка асимметричной метрики ошибки. В контексте прогнозирования энергопотребления перепотребление энергии может привести к серьезным авариям и сбоям в энергосистеме, в то время как недопотребление обычно имеет менее критические последствия. Поэтому была разработана асимметричная метрика ошибки прогнозирования с учетом установленного лимита потребления.

Предложенная метрика позволяет различать ошибки перепотребления и недопотребления, а также дополнительно штрафовать случаи превышения заданного лимита. При расчете ошибки предусмотрены два коэффициента штрафа: за недопотребление и за перепотребление, при этом нарушение лимита усиливает размер штрафа в два раза.

На основе проведенного анализа с использованием метрик MASE и асимметричной ошибки были получены следующие результаты (табл. 1, 2).

Таблица 1

Минимальные значения MASE и соответствующие размеры окна для различных методов прогнозирования

Модель	Минимальное MASE	Оптимальный размер окна
AR	37,1262	2100
MA	38,3275	2100
ARMA	37,2744	2100
ARIMA	37,1069	2100
SARIMA	40,7831	2400
Holt-Winters	42,4010	5100

Таблица 2

Минимальные значения асимметричной ошибки и соответствующие размеры окна для различных методов прогнозирования

Модель	Минимальная ошибка	Оптимальный размер окна
ARIMA	9,039580	2400
AR	9,146623	2400
SARIMA	745,947530	2400
Holt-Winters	11,160382	4200
MA	9,179376	2100
ARMA	9,026930	2400

Дальнейшее развитие проекта. При учете долгосрочных зависимостей можно заметить, что ARIMA ограничен моделированием только первого лага, как видно из графика 2, где значимость корреляции быстро убывает. Это связано с тем, что ARIMA предполагает линейную зависимость текущего значения ряда от предыдущих (p) значений и ошибок прогноза. Для учета более сложных паттернов, таких как сезонность или нелинейные зависимости, можно использовать LSTM-сети, которые способны запоминать долгосрочные зависимости благодаря механизму «вентилей» (например, forget gate), что критично для прогнозирования суточных или недельных колебаний энергопотребления. Также эффективным решением являются гибридные модели ARIMA-LSTM, сочетающие ARIMA для линейных компонент и LSTM для нелинейных остатков. Для обработки больших объемов данных с тысяч датчиков в реальном времени предлагается внедрение распределенных вычислений на платформе Apache Spark. Кроме того, важным направлением развития является реализация динамической адаптации моделей к изменениям через онлайн-обучение (online learning) для LSTM, что позволит системе

адаптироваться к изменениям нагрузки ЦОД без необходимости полного переобучения моделей.

График RACF (рис. 2) подтверждает, что значимые корреляции присутствуют только на первом лаге, что обосновывает переход к методам, способным учитывать долгосрочные зависимости. Например, LSTM может анализировать сотни предыдущих значений, что особенно важно для прогнозирования пиковых нагрузок в ЦОД.

Таким образом, дальнейшее развитие проекта направлено на повышение точности прогнозов за счет современных методов машинного обучения и их интеграции с существующей инфраструктурой.

Заключение. В работе проведен сравнительный анализ различных методов прогнозирования временных рядов для задачи прогнозирования энергопотребления ЦОД. Наиболее эффективными оказались модели ARIMA и ARMA, показавшие наименьшие значения как по метрике MASE, так и по разработанной асимметричной метрике ошибки. Оптимальный размер окна обучения для большинства моделей составляет около 2100—2400 наблюдений.

Установлено, что для большинства регрессионных моделей временных рядов меньшие размеры обучающего окна приводят к лучшим результатам прогнозирования, что объясняется способностью модели лучше адаптироваться к локальным паттернам временного ряда и меньшим влиянием выбросов на параметры модели.

Разработанная асимметричная метрика ошибки учитывает специфику задачи прогнозирования энергопотребления и может быть использована в системах мониторинга и управления энергопотреблением ЦОД для минимизации рисков перепотребления энергии.

Литература

- [1] Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice. Monash University, OTexts, 2021, 422 p.
- [2] Brockwell P.J., Davis R.A. Introduction to Time Series and Forecasting. New York, Springer, 2016, 477 p.
- [3] Abraham B., Ledolter J. Statistical Methods for Forecasting. New York, Elsevier, 1983, 352 p.
- [4] Shumway R.H., Stoffer D.S. Time Series Analysis and Its Application with R examples. New York, Springer, 2011, 202 c.
- [5] Zhang G., Patuwo B.E., Hu M.Y. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 1998, vol. 14, pp. 35–62.
- [6] Shabou S. Time Series with R. URL: <https://s-ai-f.github.io/Time-Series> (accessed 01.12.2024).
- [7] Hyndman R.J. Another look at forecast-accuracy metrics for intermittent demand. Foresight: The International Journal of Applied Forecasting, 2006, vol. 4, pp. 43–46.

Energy consumption prediction in data centers using time series analysis techniques

Mironenko Irina Alexandrovna

Kuzmina Inna Anatolievna (*scientific advisor*)

BMSTU, Moscow, Russia

The article presents a comparative analysis of energy consumption forecasting methods for data centers (DCs), where high prediction accuracy is critical for efficient resource management and operational cost reduction. The study focuses on time series regression models, including autoregression (AR), moving average (MA), ARMA, ARIMA, SARIMA, and the Holt-Winters method. The models' performance is evaluated using the Mean Absolute Scaled Error (MASE) metric and a custom asymmetric error metric designed to account for the specific challenges of energy consumption forecasting.

Keywords: energy consumption forecasting, time series, regression models, ARIMA, MASE metric, asymmetric error metric, data centers