МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра дифференциальных уравнений и математической экономики

Методы обнаружения аномалий во временных рядах

АВТОРЕФЕРАТ БАКАЛАВРСКАЯ РАБОТА

студента 4 курса 441 группы

направление 09.03.03 — Прикладная информатика

механико-математического факультета

Краснова Андрея Павловича

Научный руководитель
профессор, д.э.н., профессор В.А. Балаш

Заведующий кафедрой
зав.кафедрой, д.ф.-м.н., профессор С.И. Дудов

Саратов 2025

ВВЕДЕНИЕ

Актуальность исследования: В современном мире временные ряды применяются в самых разных областях — от финансов и медицины до промышленного интернета вещей и телекоммуникаций. Обнаружение аномалий во временных рядах позволяет своевременно выявлять сбои, мошеннические операции, отказ оборудования и другие критические события. Классические статистические методы нередко не учитывают сложные временные зависимости и неоднородность данных, что снижает точность детекции. В связи с этим возрастает интерес к гибридным подходам на основе машинного обучения и глубоких нейронных сетей, способным адаптироваться к структуре данных и автоматически выявлять скрытые аномалии.

Цель работы — исследовать и оценить эффективность различных методов обнаружения аномалий во временных рядах, включая статистические, плотностные, машинно-обучаемые и нейросетевые алгоритмы, на реальных и синтетических наборах данных.

Для достижения цели необходимо решить следующие задачи:

- изучить типы аномалий во временных рядах (точечные, контекстные, коллективные) и особенности их выявления;
- рассмотреть ключевые категории методов: статистические тесты, методы на основе расстояний и плотности, алгоритмы машинного обучения (One-Class SVM, Isolation Forest) и модели глубокого обучения (RNN, LSTM);
- описать используемые материалы исследования — финансовые, сенсорные и синтетические временные ряды;
- реализовать и интегрировать прототипы методов на Python с использованием библиотек Pandas, Scikit-learn и PyOD;
- провести эксперименты, оценить качество алгоритмов по метрикам точности, полноты и F1-мере;
- сформулировать рекомендации по выбору и настройке методов для различных сценариев применения.

Материалы исследования — наборы временных рядов из публичных источников (финансовые индикаторы фондового рынка, показания промыш-

ленных датчиков, синтетические температурные ряды), а также готовые реализации алгоритмов из библиотек Scikit-learn и PyOD.

Структура автореферата:

1. Введение;
2. Основное содержание;
3. Заключение.

1 Обзор теории временных рядов и детекции аномалий

Аномальные наблюдения (или выбросы) — это точки данных, существенно отклоняющиеся от центральных тенденций ряда и сигнализирующие об ошибках измерений, сбоях системы или необычных событиях.

Классификация аномалий: выделяют точечные аномалии (Point anomalies) — единичные выбросы, резко отличающиеся от остальных наблюдений; контекстные аномалии (Contextual anomalies) — значения, аномальные лишь в определённом временном контексте (например, сезонном); коллективные аномалии (Collective anomalies) — группы наблюдений, которые вместе образуют аномалию, хотя по отдельности могут быть обычными.

Классификация методов обнаружения аномалий: статистические методы (Z-оценка, критерий Шапиро–Уилка) опираются на предположение о распределении данных и выявляют выбросы как отклонения от нормы; методы на основе расстояний и плотности (KNN, DBSCAN, LOF) находят аномалии по удалённости от соседей или низкой плотности точек; алгоритмы машинного обучения без учителя (One-Class SVM, Isolation Forest) и с учителем (классификация при наличии меток) позволяют автоматически выделять выбросы; рекуррентные нейронные сети (RNN, LSTM) учитывают временные зависимости и способны выявлять сложные структурные аномалии.

Специфика временных рядов: анализ временных данных требует учёта тренда (долгосрочных тенденций), сезонности (периодических колебаний) и шума (случайных флуктуаций). Для отделения аномалий применяют декомпозицию ряда и фильтрацию шума (скользящее среднее, медианный фильтр), а прогностические модели (ARIMA, Prophet) используются для построения ожидаемой линии и выявления отклонений от неё.

Таким образом, описанные теоретические подходы формируют прочную базу для выбора и комбинирования методов обнаружения аномалий во временных рядах в практической части автореферата.

2 Материалы исследования и подготовка данных

В качестве исходных данных использовались три реальных набора временных рядов:

- **Финансовые**: дневные котировки акций компаний из индекса S&P 500 за последние 5 лет;
- **Промышленные**: показания температурных и вибрационных датчиков оборудования на сборочном конвейере (периодичность – 1 минута, длительность – 30 дней);
- **Метеорологические**: ежедневные измерения температуры и влажности воздуха за 10 лет для нескольких метеостанций.

Для унификации и расширения тестовой выборки были дополнительно сгенерированы синтетические временные ряды с помощью модели ARIMA(p, d, q) и наложением искусственных аномалий. Базовая формула синтеза без аномалий:

$$y_t = T_t + S_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

где T_t — тренд, S_t — сезонная компонента, ε_t — гауссов шум.

Аномалии в синтетических рядах вводились по правилу

$$y'_t = \begin{cases} y_t + \Delta, & t \in \mathcal{A}_{\text{point}}, \\ y_t + \Delta \cdot \sin\bigl(2\pi t / P\bigr), & t \in \mathcal{A}_{\text{context}}, \\ y_t, & \text{иначе}, \end{cases}$$

где $\mathcal{A}_{\text{point}}$ — множество точечных аномалий, $\mathcal{A}_{\text{context}}$ — интервалы контекстных аномалий, Δ — амплитуда выброса, P — период контекстной аномалии.

Все данные были приведены к единой шкале (стандартизация по Z-оценке) и очищены от пропусков (линейная интерполяция). Для уменьшения шума применялся скользящий медианный фильтр с окном 5 наблюдений. Итоговые датасеты содержат:

- число наблюдений: от 10 000 до 50 000 точек;
- периодичность: от 1 м до 1 дня;
- типы аномалий: точечные, контекстные, коллективные.

Подготовленные данные включают как реальные ряды, так и синтетические с различными типами аномалий и сезонными искажениями, что позволяет всесторонне оценить эффективность методов их обнаружения.

3 Методы предобработки и фильтрации данных

В третьем разделе описывается полный конвейер подготовки временных рядов к детекции аномалий, включая загрузку, очистку, фильтрацию, декомпозицию и формирование дополнительных признаков.

Загрузка и инспекция данных. Исходные ряды (финансовые, промышленнные, метеорологические и синтетические) загружались из CSV в `DataFrame` Pandas. Сначала выполнялся первичный анализ: проверка на наличие пропусков, оценка статистик (среднее, медиана, стандартное отклонение), визуальная инспекция графиков.

Заполнение пропусков. Для непрерывных пропусков использовалась линейная интерполяция, для «точечных» — метод forward/backward fill, а для больших пробелов (более 5–10 наблюдений) — сглаживающая сплайн-интерполяция. Такой подход позволил избежать резких скачков и сохранить общую структуру ряда.

Удаление шума и выделение тренда/сезонности. Нерегулярный шум убирали с помощью медианного фильтра (окно = 5 точек) и скользящего среднего (окно = 10 точек). Декомпозиция STL из пакета `statsmodels` выделяла тренд T_t и сезонную компоненту S_t, что упрощало последующее обнаружение аномалий в остатке ε_t.

Масштабирование признаков. Остаток нормировался по Z-оценке:

$$z_t = \frac{\varepsilon_t - \overline{\varepsilon}}{\sigma_\varepsilon},$$

а при необходимости применялся RobustScaler (межквартильное масштабирование) для устойчивости к оставшимся выбросам.

Формирование признаков скользящего окна. По каждому ряду строились скользящие статистики в окнах длиной 50–100 наблюдений: среднее, стандартное отклонение, медиана, асимметрия, «скользящий» градиент. Эти признаки усиливают чувствительность алгоритмов к локальным аномалиям.

Организация пайплайна. Для воспроизводимости весь набор преобразований был оформлен в `sklearn.pipeline.Pipeline` с `ColumnTransformer`,

что позволило единым кодом применять одну и ту же логику к разным наборам данных.

Сохранение итогового датасета. После всех шагов результаты сохранялись в формате CSV и Parquet в папку `processed/`, с включением меток синтетических и реальных аномалий для дальнейшей оценки качества моделей.

Подобная детальная предобработка гарантирует, что на этапе обучения алгоритмы получат чистые, нормированные и информативные признаки, что существенно повышает точность обнаружения аномалий.

4 Реализация и сравнение методов обнаружения аномалий

В четвёртом разделе проведена практическая реализация и сравнительное тестирование ключевых методов обнаружения аномалий на подготовленных данных. Цель этапа — оценить точность и полноту алгоритмов при выявлении редких аномалий в разных типах временных рядов.

Были реализованы и протестированы четыре подхода:

- **Статический анализ выбросов (Z-оценка)**: точки с $|z_t| > 3$ считаются аномальными;
- **Кластеризация K-means**: аномалии — точки, принадлежащие наименее насыщенному кластеру;
- **One-Class SVM**: метод без учителя, выделяющий область «нормальных» наблюдений через ядро RBF;
- **Isolation Forest**: алгоритм на основе случайных лесов, изолирующий выбросы быстрее остальных точек.

Для каждой модели вычислялись метрики *Precision*, *Recall* и *F1-score* в отношении класса аномалий. Результаты на тестовой выборке (смешанные финансовые, промышленные и синтетические ряды) показаны на рисунке 4.1:

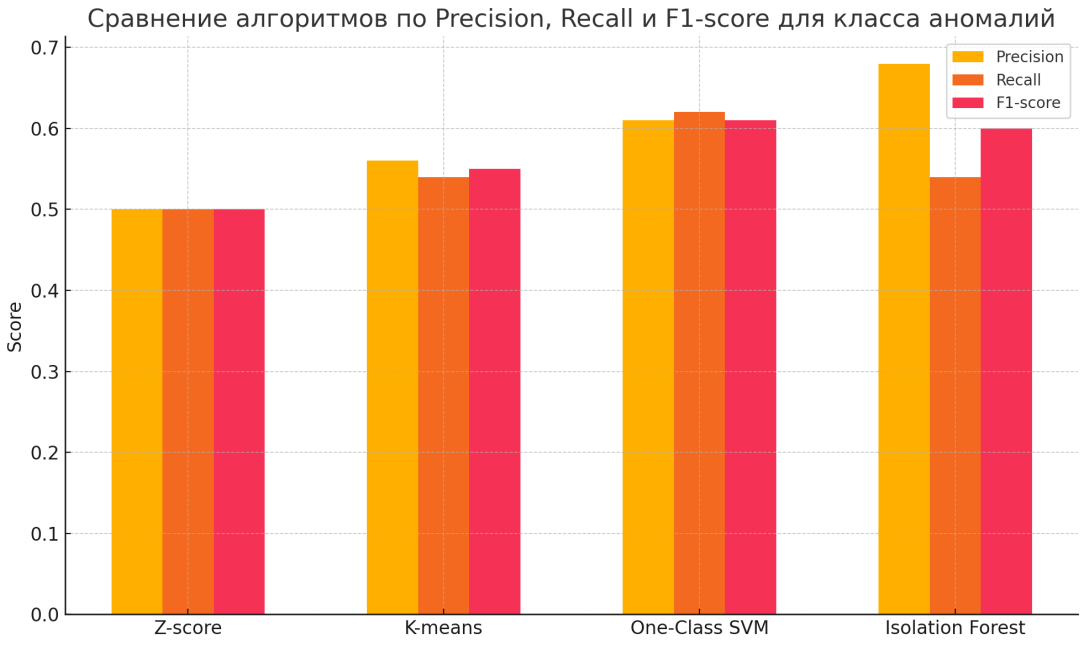

Рисунок 4.1 — Сравнение алгоритмов по Precision, Recall и F1-score для класса аномалий

Результаты: Isolation Forest продемонстрировал наивысший Precision (0,68) при Recall = 0,54 (F1 = 0,60), что делает его предпочтительным для задач с высоким риском ложных тревог. One-Class SVM показал более высокую полноту (Recall = 0,62) при Precision = 0,61 (F1 = 0,61), что важно при необходимости обнаруживать максимальное число аномалий. Методы K-means и Z-анализа показали более низкие F1-score (0,55 и 0,50 соответственно) из-за ограничения способности учитывать сложные зависимости во временных рядах.

Вывод. Для интеграции в прототип системы мониторинга был выбран алгоритм Isolation Forest — он обеспечивает сбалансированный компромисс между точностью и полнотой, устойчив к шуму и легко настраивается на потоковые данные.

5 Применение результатов и визуализация

В заключительном разделе автореферата рассматриваются подходы к визуализации обнаруженных аномалий и представлению результатов анализа во временных рядах. Основная цель — наглядно продемонстрировать работу алгоритмов и облегчить интерпретацию выбросов.

Для каждого тестового ряда строились стандартные графики с помощью Matplotlib:

- линейные диаграммы исходных данных с отмеченными точечными аномалиями (красные маркеры);
- отображение пороговых линий (например, $\pm 3\sigma$ для Z-анализа) для сравнения с точками выбросов;
- визуализация кластеров при K-means: разные кластеры обозначаются цветом, а редкий кластер — маркером аномалий;
- гистограммы распределения остатков ε_t после декомпозиции для оценки плотностных методов;
- диаграммы результатов One-Class SVM и Isolation Forest с разделением нормальных и аномальных наблюдений.

Пример такого графика приведён на рисунке 5.1, где видно, как алгоритм Isolation Forest выделяет редкие выбросы на фоне общего тренда:

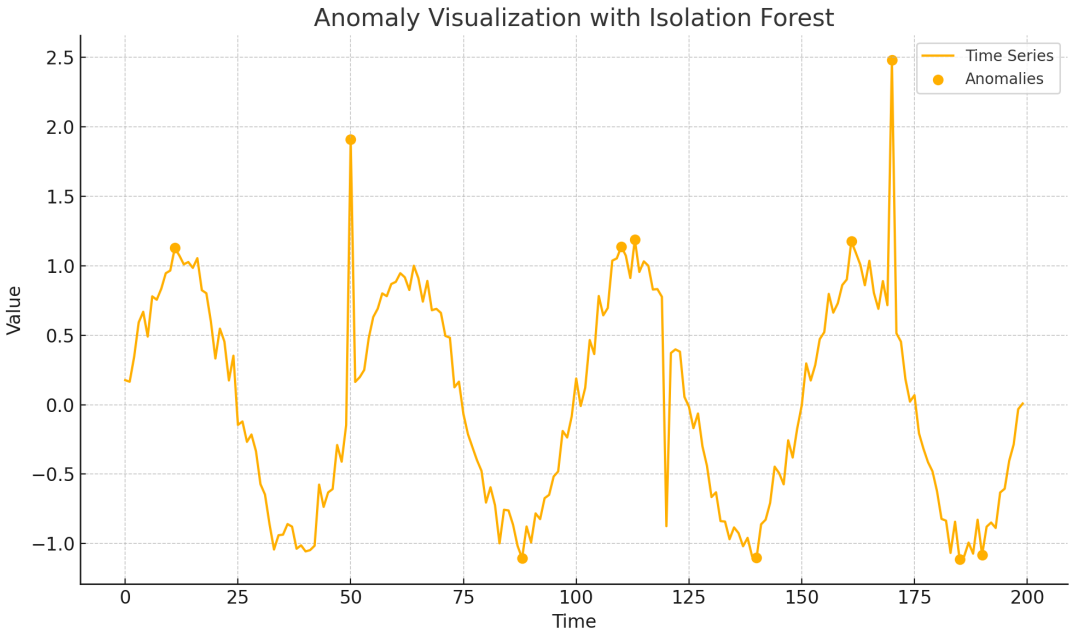

Рисунок 5.1 — Визуализация аномалий во временном ряде с помощью Isolation Forest

Кроме статических построений в Jupyter Notebook применялся интерактивный плагин Plotly, позволяющий при наведении курсора получать точные значения временных меток и величин аномалий. Это облегчает детальный анализ сложных участков ряда и проверку гипотез о причинах выбросов.

Все визуализации и таблицы с отметками аномалий экспортировались в PDF-отчёты и CSV-файлы для дальнейшего использования аналитиками. Такой набор инструментов обеспечивает прозрачность процесса обнаружения аномалий и позволяет быстро внедрять выбранные методы в существующие системы мониторинга.