Силаев Андрей Николаевич

магистр

Поляков Евгений Алексеевич

магистр

Самохвалов Платон Романович

магистр

ФГАОУ ВО «Московский политехнический университет»

г. Москва

DOI 10.21661/r-588563

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ОБНАРУЖЕНИЯ ТОЧЕЧНЫХ, КОНТЕКСТНЫХ И ЭПИЗОДИЧЕСКИХ АНОМАЛИЙ В ДАННЫХ ТРАНСПОРТНОЙ ТЕЛЕМЕТРИИ

Аннотация: *в статье представлен сравнительный анализ методов обнаружения аномалий в данных транспортной телеметрии на примере реального открытого набора навигационно-временных данных городского пассажирского транспорта. В настоящем исследовании выполнена практическая оценка на реальном открытом датасете Dublin Bus Delay Dataset применимости методов Isolation Forest, Local Outlier Factor и One-Class SVM для выявления трёх классов аномалий: точечных, контекстных и эпизодических. Рассматриваются особенности построения признаков для разных уровней представления данных и экспериментально оценивается эффективность методов по метрикам качества и вычислительным характеристикам. На основе полученных результатов сформулированы рекомендации по выбору алгоритма в зависимости от типа аномалии и условий практического применения в системах транспортного мониторинга.*

Ключевые слова: *обнаружение аномалий, точечные аномалии, контекстные аномалии, эпизодические аномалии, транспортная телеметрия, го родской пассажирский транспорт, Isolation Forest, Local Outlier Factor, One-Class SVM, сравнительный анализ.*

Спасибо Платону за вклад в работу!

Введение.

Современные транспортные системы характеризуются быстрым ростом объёмов и скорости поступления навигационно-телеметрических данных [1]. Аналитические платформы городского пассажирского транспорта непрерывно собирают GPS-координаты, скорость и курс движения, временные метки, информацию о прохождении остановок и отклонениях от расписания. В таких потоках наряду с закономерностями неизбежно присутствуют аномалии – отклонения от ожидаемого поведения, выявление которых важно для контроля соблюдения маршрутов, повышения регулярности движения, сокращения простоев и раннего обнаружения эксплуатационных и технических проблем.

К практически значимым классам транспортных аномалий относятся аномалии качества данных и позиционирования, в том числе скачки координат, невозможные скорости, ошибки временных меток и низкая точность позиционирования; операционные отклонения движения, включая отклонение от маршрута, пропуски остановок и внеплановые стоянки; а также контекстуальные аномалии регулярности, например аномальные задержки или опережения относительно типичного профиля для конкретного маршрута и временного интервала. В общем виде задача сводится к выделению наблюдений и эпизодов, существенно отклоняющихся от нормального поведения в соответствующем контексте при отсутствии надёжной полной разметки.

Обнаружение аномалий в транспортной телеметрии осложняется потоковым характером данных, контекстной зависимостью нормы, высокой неоднородностью поведения на разных маршрутах и временных интервалах, а также сочетанием точечных и коллективных отклонений. Поэтому применение универсального алгоритма без учёта класса аномалии и уровня представления данных оказывается недостаточными [2].

Целью настоящей работы является экспериментальное сравнение методов Isolation Forest, Local Outlier Factor и One-Class SVM на реальных данных

транспортной телеметрии при обнаружении точечных, контекстных и эпизодических аномалий.

Материалы и методы исследования.

В качестве эмпирической базы исследования использован открытый набор данных Dublin Bus Delay Dataset из репозитория UCI Machine Learning Repository. Датасет содержит телеметрические данные автобусов городской транспортной сети Дублина, включая временные метки, GPS-координаты, идентификаторы транспортных средств и рейсов, сведения об остановках и задержках движения. Для экспериментов было сформировано рабочее подмножество маршрутов с наиболее полным набором телеметрических и временных признаков. После очистки и отбора итоговый массив составил 1 842 317 записей, что позволило оценить методы на выборке, близкой к условиям реального транспортного мониторинга.

Исходные данные были упорядочены по идентификатору транспортного средства и времени регистрации. На этапе предварительной обработки удалялись дублирующиеся и некорректные записи, а также наблюдения с ошибочными координатами и временными метками. Для последовательных телеметрических точек рассчитывались расстояние между соседними позициями, временной интервал, мгновенная скорость и ускорение, что позволило привести данные к единому аналитическому виду и выделить признаки, потенциально связанные с аномальным поведением.

Далее были сформированы три представления данных, соответствующие различным уровням анализа. Первое включало отдельные телеметрические точки и использовалось для поиска точечных аномалий. Второе строилось на уровне остановок и сегментов маршрута и применялось для анализа задержек и отклонений в контексте конкретного рейса. Третье представление формировалось в виде скользящих временных окон продолжительностью 5 минут и предназначалось для выявления более длительных эпизодов аномального поведения. Для методов, чувствительных к масштабу признаков, выполнялась стандартизация количественных переменных.

В исследовании использована прикладная типология аномалий, ориентированная на особенности транспортной телеметрии. Все отклонения были разделены на три класса: точечные, контекстные и эпизодические. К *точечным аномалиям* относились отдельные наблюдения, резко отклоняющиеся от основной массы данных, например скачки координат, невозможные значения скорости и аномальные временные интервалы между сообщениями. *Контекстные аномалии* определялись несоответствием условий движения и включали нетипичные задержки на остановках и аномальное время прохождения сегментов относительно маршрута, времени суток и сопоставимых рейсов. *Эпизодические аномалии* представляли собой последовательности взаимосвязанных наблюдений, формирующих нетипичный поведенческий паттерн, например продолжительную внеплановую стоянку, затяжное движение с аномально низкой скоростью или устойчивый рост задержки в пределах временного окна.

Такое разделение позволило сравнить рассматриваемые методы с учётом реальных сценариев, характерных для данных городского пассажирского транспорта.

Формирование признакового пространства для обнаружения точечных, контекстных и эпизодических аномалий.

Для корректного сравнения методов обнаружения аномалий в транспортной телеметрии было сформировано единое признаковое пространство, учитывающее различия между точечными, контекстными и эпизодическими отклонениями. Поскольку каждый из указанных типов аномалий проявляется на своём уровне наблюдения, признаки строились не только по исходным телеметрическим полям, но и по агрегированным характеристикам движения. Такой подход позволил перейти от анализа отдельных сообщений к более содержательному описанию поведения транспортного средства в контексте маршрута и во времени.

Для выявления точечных аномалий использовались признаки, рассчитываемые на уровне отдельных телеметрических точек и их ближайшего окружения. В данную группу вошли текущая скорость v_t, приращение скорости Δv_t, ускорение

a_t, временной интервал между сообщениями Δt_t, расстояние между соседними точками Δs_t, а также мгновенная скорость, рассчитанная по координатам. Дополнительно учитывались изменение курса $\Delta\theta_t$, признак нахождения транспортного средства на остановке и текущее отклонение по задержке. Эти признаки позволяют фиксировать единичные грубые выбросы, связанные со скачками координат, невозможными значениями скорости, нарушениями временной последовательности сообщений и иными локальными артефактами телеметрии.

Для анализа контекстных аномалий использовались признаки, формируемые на уровне остановок и сегментов маршрута. В эту группу были включены время прохождения сегмента, длительность стоянки на остановке и задержка относительно предыдущей остановки. Для учёта условий движения дополнительно рассчитывались медианное время прохождения аналогичного сегмента в том же контексте, z-оценка времени прохождения в группе «маршрут – направление – интервал суток – тип дня» и относительное отклонение задержки от медианы контекста. Также в признаковое описание включались порядковый номер остановки на маршруте и бинарные признаки пикового и непикового режима. Использование таких характеристик позволило учитывать, что норма в транспортных данных зависит не только от абсолютных значений, но и от операционного контекста, в котором происходит движение.

Для выявления эпизодических аномалий данные дополнительно агрегировались в скользящие окна длиной 5 минут. На этом уровне рассчитывались средняя и медианная скорость, стандартное отклонение скорости, максимальное и минимальное значения скорости, доля времени с нулевой или близкой к нулю скоростью, суммарная длительность стоянок и число остановок в пределах окна. Кроме того, использовались суммарное отклонение по задержке, максимальное изменение задержки внутри окна и длина непрерывного интервала аномально низкой скорости. Такое оконное представление позволяло выявлять не отдельные выбросы, а устойчивые эпизоды нетипичного поведения, включая продолжительные внеплановые стоянки, затяжное замедление движения и устойчивое нарастание задержки.

Предлагаемое признаковое пространство обеспечило сопоставимое представление данных на трёх уровнях анализа и позволило корректно сравнить методы Isolation Forest, Local Outlier Factor и One-Class SVM применительно к различным типам аномалий в транспортной телеметрии.

Рассматриваемые методы обнаружения аномалий.

Эффективное обнаружение аномалий в больших неразмеченных данных телеметрии городского транспорта требует масштабируемых методов, устойчивых к многомерности и отсутствию разметки [3; 4]. В данной статье рассмотрим три фундаментальных метода, представляющих различные методы обнаружения аномалий без опоры на размеченные примеры аномалий, что делает их особенно ценными для работы с большими неразмеченными массивами информации.

Isolation Forest (iForest) основан на идее изоляции аномалий случайными разбиениями пространства признаков. Поскольку аномальные объекты, как правило, являются редкими и отличаются по значениям признаков, они изолируются за меньшее число разбиений по сравнению с нормальными объектами.

Для объекта xитоговая оценка аномальности задаётся выражением:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}},$$

где $E(h(x))$ – средняя длина пути изоляции по всем деревьям ансамбля, $c(n)$ – нормировочный коэффициент.

В экспериментах использовались параметры: n_estimators = 200, max_samples = 256, contamination = 0.03 для точечных аномалий и 0.05 для агрегированных уровней.

Local Outlier Factor (LOF) основан на сравнении локальной плотности объекта с плотностью его ближайших соседей. Для объекта pкоэффициент аномальности определяется как

$$LOF_k(p) = \frac{1}{\mid N_k(p) \mid} \sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)},$$

где $N_k(p)$ – множество k ближайших соседей, lrd_k – локальная достижимая плотность.

Если значение LOF существенно превышает 1, объект рассматривается как локальная аномалия. В работе использовались параметры: n_neighbors = 20 для точечного уровня, n_neighbors = 35 для контекстного и n_neighbors = 25 для оконных последовательностей.

One-Class SVM (SVM) строит границу нормального поведения в преобразованном признаковом пространстве, отделяя основную массу нормальных объектов от атипичных наблюдений. Решающее правило имеет вид:

$$f(z) = \text{sign}\left(\sum_i \alpha_i K(x_i, z)\text{-}\rho\right),$$

где $K(x_i, z)$ – ядерная функция, α_i- коэффициенты модели, ρ – смещение.

В исследовании использовалось RBF-ядро с параметрами nu = 0.05, gamma = scale. Метод применялся на всех трёх уровнях, но основной интерес представляло его поведение на оконных представлениях для эпизодических аномалий.

Экспериментальная постановка.

Для сопоставимого сравнения методов Isolation Forest, Local Outlier Factor и One-Class SVM была использована единая экспериментальная схема, включающая формирование эталонной разметки, проведение вычислительных экспериментов в общей программно-аппаратной среде и оценку результатов по набору стандартных метрик качества. Такая постановка позволила сравнить методы не только по способности обнаруживать аномалии разных типов, но и по эксплуатационным характеристикам, важным для практического применения в системах транспортного мониторинга.

Так как исходный телеметрический датасет не содержал полной надёжной разметки аномалий, для оценки качества была применена гибридная схема [5]:

− первичная автоматическая разметка на основе жёстких транспортных правил;

− ручная верификация подвыборки экспертным способом;

− формирование тестового эталонного набора для расчёта метрик.

Для предварительной разметки использовались следующие правила:

− скорость, рассчитанная по координатам, превышает 110 км/ч для городского автобуса;

− перемещение между точками не согласуется с допустимой скоростью за интервал Δt;

− время прохождения сегмента превышает 95-й перцентиль своего контекста более чем в 1,5 раза;

− стоянка вне зоны остановки длится более 4 минут;

− задержка в окне возрастает монотонно и превышает локальную медиану более чем на 2,5 стандартных отклонения.

После экспертной проверки был сформирован тестовый набор, включающий:

− 12 000 точечных объектов, из них 3,1% аномальных;

− 8 400 объектов уровня «остановка/сегмент», из них 4,7% аномальных;

− 4 800 оконных последовательностей, из них 5,4% аномальных.

Все вычислительные эксперименты выполнялись в единой среде:

− Python 3.11;

− scikit-learn 1.4;

− M1 Pro / 32 GB RAM;

− операционная система MacOS.

Это позволило корректно сравнить не только качество детектирования, но и эксплуатационные характеристики методов.

Для оценки использовались:

− Precision;

− Recall;

− F1-score;

− PR-AUC;

− время обучения;

− время обработки 10 000 объектов;

− потребление оперативной памяти.

Использование PR-AUC оправдано сильной несбалансированностью классов, характерной для задач обнаружения аномалий. Включение временных и ресурсных показателей дополнительно позволило оценить практическую пригодность каждого метода для анализа больших потоков транспортной телеметрии.

Таким образом, выбранная экспериментальная постановка обеспечила единые условия сравнения алгоритмов и позволила оценить их эффективность применительно к различным типам аномалий в реальных транспортных данных.

Результаты исследования.

Результаты для точечных аномалий. На задаче выявления точечных аномалий наилучший результат показал *Isolation Forest*. Метод эффективно обнаруживал грубые телеметрические выбросы: скачки координат, нереалистичную скорость и аномальные интервалы между сообщениями. LOF также выявлял часть таких наблюдений, но его результат заметно зависел от структуры локальной плотности. One-Class SVM оказался менее устойчивым на этом уровне представления данных и потребовал более значительных вычислительных затрат. Результаты для точечных аномалий представлены в таблице 1.

Таблица 1

Результаты для точечных аномалий

Метод	Precision	Recall	F1-score	PR-AUC	Время обучения, с
Isolation Forest	0.91	0.87	0.89	0.92	14.6
LOF	0.83	0.77	0.80	0.84	41.3
One-Class SVM	0.78	0.74	0.76	0.81	68.9

Полученные значения подтверждают, что для первичного скрининга больших потоков телеметрии Isolation Forest является наиболее подходящим выбором.

Результаты для контекстных аномалий. На уровне контекстных аномалий лидирующие результаты продемонстрировал *Local Outlier Factor.* Его преимущество проявилось при анализе времени прохождения сегментов и задержек на остановках относительно сходных рейсов в аналогичных условиях. Именно локальное сравнение с соседними объектами позволило корректно учитывать неоднородность транспортных данных.

Isolation Forest на этом уровне хуже различал тонкие локальные отклонения, если абсолютные значения признаков не выходили за глобально допустимые пределы. One-Class SVM показал близкие результаты, однако был более чувствителен к настройке параметров и менее интерпретируем. Результаты для контекстных аномалий представлены в таблице 2.

Таблица 2

Результаты для контекстных аномалий

Метод	Precision	Recall	F1-score	PR-AUC	Время обучения, с
Isolation Forest	0.76	0.72	0.74	0.79	9.8
LOF	0.84	0.80	0.82	0.87	27.4
One-Class SVM	0.81	0.77	0.79	0.83	46.2

Таким образом, LOF оказался наиболее эффективным именно для тех аномалий, где понятие нормы зависит от локального транспортного контекста.

Результаты для эпизодических аномалий. Для эпизодических аномалий на оконных признаках наиболее устойчивый результат показал *One-Class SVM*. Метод лучше моделировал границу нормального поведения в пространстве интегральных характеристик движения и точнее выделял продолжительные внеплановые стоянки, устойчивые эпизоды низкой скорости и окна с нетипичным ростом задержки.

Isolation Forest на оконных признаках также показал высокий результат, но чаще дробил продолжительные аномальные эпизоды на несколько отдельных срабатываний. LOF уступил двум другим методам из-за чувствительности к структуре локального соседства оконных объектов. Результаты для эпизодических аномалий представлены в таблице 3.

Таблица 3

Результаты для эпизодических аномалий

Метод	Precision	Recall	F1-score	PR-AUC	Время обучения, с
Isolation Forest	0.82	0.79	0.80	0.85	7.1
LOF	0.74	0.70	0.72	0.77	18.6
One-Class SVM	0.86	0.82	0.84	0.88	31.7

Эти результаты показывают, что при переходе от отдельных наблюдений к последовательностям поведение алгоритмов меняется, и модель, уступавшая на точечном уровне, может оказаться лучшей на эпизодическом.

С точки зрения эксплуатационных параметров наиболее экономичным оказался *Isolation Forest*. Он быстрее обучался, требовал меньше памяти и лучше масштабировался на больших объёмах данных. LOF оказался наиболее чувствительным к росту числа объектов из-за необходимости поиска ближайших соседей. One-Class SVM демонстрировал наибольшую вычислительную стоимость, но на умеренных по размеру оконных выборках эта стоимость оставалась приемлемой.

Таблица 4

Сравнение методов по эксплуатационным характеристикам

Критерий	Isolation Forest	Local Outlier Factor	One-Class SVM
Скорость обучения	высокая	средняя	низкая
Масштабируемость по объёму данных	очень высокая	ограниченная	низкая
Потребление памяти	низкое	среднее/высокое	высокое
Устойчивость к высокой размерности	высокая	низкая	средняя
Чувствительность к шуму	средняя	высокая	низкая
Интерпретируемость	средняя	высокая	низкая
Наиболее подходящий тип аномалий	точечные	контекстные	эпизодические

Проведённое исследование показало, что в транспортной телеметрии не существует универсального метода, одинаково эффективного для всех типов аномалий. Результат определяется не только свойствами самого алгоритма, но и тем, на каком уровне представлены данные и как сформировано признаковое пространство.

Для точечных аномалий ключевую роль играют грубые выбросы, которые хорошо выявляются методом Isolation Forest за счёт механизма случайной изоляции. Этот алгоритм практически не зависит от локальной плотности и поэтому особенно удобен для первичного онлайн-скрининга больших потоков GPS-телеметрии.

Для контекстных аномалий решающее значение имеет сравнение объекта с релевантным окружением, а не со всей совокупностью данных. В этой постановке наиболее адекватным оказался LOF, так как он позволяет учитывать, что норма на одном и том же маршруте может существенно различаться в зависимости от времени суток, направления и операционной ситуации.

Для эпизодических аномалий важно представление данных в виде окон и агрегированных последовательностей. Здесь преимущества получил One-Class SVM, способный строить нелинейную границу нормального поведения в пространстве интегральных характеристик движения. Несмотря на более высокую вычислительную стоимость, метод показал наилучшее соотношение точности и полноты для длительных отклоняющихся эпизодов.

Практически значимым выводом является целесообразность каскадного подхода в реальных системах транспортного мониторинга.

1. Isolation Forest – для первичной фильтрации грубых точечных выбросов.

2. LOF – для поиска контекстных отклонений на остановках и сегментах.

3. One-Class SVM – для анализа оконных последовательностей и поиска продолжительных эпизодов аномального поведения.

Такой подход обеспечивает баланс между скоростью, чувствительностью к локальному контексту и способностью выделять длительные нарушения в работе транспорта.

Заключение.

В статье выполнен сравнительный анализ методов Isolation Forest, Local Outlier Factor и One-Class SVM на реальных данных транспортной телеметрии. Результаты показали, что эффективность алгоритмов зависит от типа аномалии и уровня представления данных.

Isolation Forest наиболее эффективен для выявления точечных аномалий и первичного скрининга телеметрического потока. Local Outlier Factor показал лучшие результаты при обнаружении контекстных отклонений, а One-Class SVM – при выявлении эпизодических аномалий на оконных представлениях данных.

Таким образом, универсального метода для всех типов транспортных аномалий не существует, а наиболее рациональным является выбор алгоритма с учётом характера отклонений и структуры признакового пространства.

Список литературы

1. Белов С.Д. Методы и технологии Больших данных для решения научных задач в распределённой вычислительной среде: автореф. дис. ... канд. техн. наук / С.Д. Белов; Объединённый ин-т ядер. исследований. – Дубна, 2024. – 29 с. – URL: https://rusneb.ru/catalog/000199_000009_013238453/ (дата обращения: 25.03.2026).

2. Булыгин М.В. Анализ транспортных данных – новые модель и программная платформа: автореф. дис. ... канд. техн. наук / М.В. Булыгин; Моск. физ.-техн. ин-т (нац. исслед. ун-т). – М., 2025. – 23 с. – URL: https://rusneb.ru/catalog/000199_000009_013564724/ (дата обращения: 26.03.2026).

3. Краева Я.А. Масштабируемые методы и алгоритмы поиска аномалий во временных рядах: автореф. дис. ... канд. физ.-мат. наук / Я.А. Краева; Южно-Уральский гос. ун-т (нац. исслед. ун-т). – Челябинск, 2024. – 23 с. – URL: https://rusneb.ru/catalog/000199_000009_012696951/ (дата обращения: 26.03.2026).

4. Рыжиков А.С. Глубокие порождающие модели для поиска аномалий: автореф. дис. ... канд. физ.-мат. наук / А.С. Рыжиков; Нац. исслед. ун-т «Высшая школа экономики». – М., 2024. – 24 с. – URL: https://rusneb.ru/catalog/000199_000009_013209249/ (дата обращения: 27.03.2026).

5. Сивак М.А. Робастное обучение нейронных сетей с простой архитектурой для решения задач классификации: автореф. дис. ... канд. техн. наук / М.А. Сивак; Новосиб. гос. техн. ун-т. – Новосибирск, 2022. – 20 с. – URL: https://rusneb.ru/catalog/000199_000009_011112275/ (дата обращения: 27.03.2026).