УДК 31(47+57):622.276

Публикуется в авторской редакции

Адаптация методов поиска трендов и изменений статистических характеристик ряда в потоковых данных с промысловых объектов

Е.В. Юдин[1], к.ф.-м.н.
Э.Ф. Мурзин[2]
М.И. Гудилов[3]
Д.О. Исаев[1]

[1]Группа компаний «Газпром нефть»
[2]ООО «ОЙЛ ЭНД ГЭС ПРОДАКШН ТУЛС»
[3]ООО «НЕДРА»

Адрес для связи: muremfic@mail.ru

Ключевые слова: телеметрия, электроцентробежный насос (ЭЦН), временные ряды, алгоритмы CUSUM, ADWIN, BOCPD

В работе рассматриваются современные методы анализа временных рядов, направленные на выявление трендов и изменений статистических характеристик ряда. Цель исследования заключается в разработке адаптивных алгоритмов анализа временных рядов, обеспечивающих оперативное выявление аномалий и точек изменения статистических характеристик в режиме потоковой обработки данных со скважин. Для достижения поставленной цели проводится обзор и сравнительный анализ классических и современных методов. Среди алгоритмов поиска тренда исследуется CUSUM, рассматриваются его классическая реализация и подходы к определению конца промежутка. Предлагается применение модифицированного алгоритма CUSUM для обработки данных в реальном времени, а также использование сглаживающих методов — экспоненциального скользящего среднего, взвешенного скользящего окна и фильтра Калмана — для предварительной подготовки данных перед анализом. Для задач поиска точек изменения статистических характеристик ряда исследуются алгоритмы ADWIN и BOCPD. Предлагаются их адаптации: для ADWIN — новый способ сжатия точек, позволяющий быстрее реагировать на изменения, для BOCPD — оптимизация обработки при накоплении большого объема данных, что существенно ускоряет вычисления. Кроме того, представлен новый метод, основанный на последовательном применении информационного критерия и F-теста, позволяющий выявлять изменения дисперсии в потоковых

Adaptation of trend detection and statistical change point detection methods for streaming data from field assets

E.V. Yudin[1]
E.F. Murzin[2]
M.I. Gudilov[3]
D.O. Isaev[1]

[1]Gazprom Neft Company Group, RF, Saint Petersburg
[2]OIL AND GAS PRODUCTION TOOLS LLC, RF, Krasnogorsk
[3]NEDRA LLC, RF, Saint Petersburg

E-mail: muremfic@mail.ru

Keywords: telemetry, electric submersible pump (ESP), time series, algorithms CUSUM, ADWIN, BOCPD

This paper examines modern methods of time series analysis aimed at detecting trends and changes in the statistical characteristics of a series. The research objective is to develop adaptive time series analysis algorithms that enable real-time anomaly detection and identification of change points in statistical characteristics during the stream processing of well data. To achieve this goal, a review and comparative analysis of classical and modern methods is conducted. Among trend detection algorithms, CUSUM is investigated, considering both its classic implementation and approaches to determining the segment endpoint. The application of a modified CUSUM algorithm for real-time data processing is proposed, along with the use of smoothing methods — exponential moving average, weighted moving window, and Kalman filter — for preliminary data preparation prior to analysis. For the tasks of detecting change points in the statistical characteristics of a series, the ADWIN and BOCPD algorithms are examined. Their adaptations are proposed: for ADWIN — a new method for compressing points, allowing for faster response to changes; for BOCPD — optimization of processing with accumulating large volumes of data, which significantly accelerates computations. Furthermore, a novel method based on the sequential application of an information criterion and an F-test is presented, enabling the detection of variance changes in streaming data. The novelty of the work lies in the comprehensive adaptation of algorithms for operation under conditions of stream processing of

данных. Новизна работы заключается в комплексной адаптации алгоритмов для функционирования в условиях потоковой обработки промысловых данных с возможностью их запуска в режиме real-time на edge-устройствах. Такой подход открывает перспективы для построения локальных интеллектуальных систем, потенциально способных к автономному управлению режимами работы скважины без существенной зависимости от удаленных вычислительных ресурсов. Практическая значимость работы проявляется в возможности применения предложенных алгоритмов для автоматического контроля состояния оборудования, прогнозирования изменений производительности скважин и своевременного выявления отклонений от нормального режима работы.

field data, with the capability to run in real-time on edge devices. This approach opens prospects for building local intelligent systems, potentially capable of autonomous well operation management without significant reliance on remote computing resources. The practical significance of the work is manifested in the possibility of applying the proposed algorithms for automatic equipment condition monitoring, forecasting well productivity changes, and timely detection of deviations from normal operating conditions.

Введение

Современные нефтегазовые предприятия характеризуются высоким уровнем цифровизации и широким применением телеметрических систем. В технологических процессах используется большое количество датчиков, осуществляющих непрерывную передачу данных о давлении, температуре, дебите, вибрациях и других параметрах, отражающих состояние оборудования и процессов добычи. Однако телеметрия с датчиков чаще всего передается для анализа в корпоративные сеть передачи данных (КСПД), а не обрабатывается в реальном времени в контуре технологической сети передачи данных (ТСПД) [1]. Это создает ограничения: узкая пропускная способность каналов связи вынуждает снижать частоту или разрешение данных, что ведет к потере информативности, а задержка при передаче снижает оперативность реакции.

Перенос аналитики непосредственно в ТСПД, на периферийные устройства, позволяет проводить полноценный анализ сырых данных высокой детализации без их передачи, отправляя в центр только агрегированные показатели и события. Таким образом, ключевым условием является разработка алгоритмов анализа временных рядов, способных функционировать на ограниченных ресурсах периферийных устройств и обеспечивать принятие решений в реальном времени.

В настоящей работе рассматриваются модификации алгоритмов анализа временных рядов, направленные на их применение в условиях потоковой обработки. Предложены адаптации алгоритмов CUSUM, ADWIN и BOCPD, позволяющие снизить чувствительность к шумам, уменьшить вычислительные затраты и обеспечить работу в режиме реального времени. Кроме того, разработан новый метод выявления изменений

дисперсии, основанный на сочетании информационного критерия и F-теста. Предлагаемые подходы ориентированы на использование в интеллектуальных системах мониторинга и управления, включая периферийные вычислительные устройства.

1. Анализ трендов во временных рядах.

1.1. Постановка задачи и требования к алгоритмам

Задача выделения тренда во временных рядах заключается в определении устойчивого направления изменения сигнала на фоне флуктуаций, вызванных шумом в данных. Для телеметрических данных, поступающих в режиме реального времени, требуется не только точное, но и оперативное определение тренда, что накладывает дополнительные ограничения на применяемые методы.

Можно выделить основные требования к алгоритмам: они должны обеспечивать корректную работу в потоковом режиме, без необходимости пересчета всей истории наблюдений, быть устойчивыми к шуму и выбросам, которые характерны для промышленных сигналов и могут приводить к ложному определению направлений изменения. Также алгоритмы должны быть вычислительно эффективными и не требовать значительных ресурсов памяти.

С учетом перечисленных требований, наиболее подходящим для решения задачи выделения тренда в условиях потоковой обработки данных, представляется класс алгоритмов CUSUM. Рассмотрим его подробнее.

1.2. Обзор CUSUM и его реализации

Алгоритм CUSUM (Cumulative Sum Control Chart) [2] основан на анализе кумулятивных сумм отклонений наблюдаемых значений от ожидаемого среднего уровня при отсутствии изменений. Пусть имеется последовательность наблюдений $x_1, x_2, \ldots, x_t$. Тогда кумулятивная сумма отклонений вычисляется по рекуррентной формуле:

$$S_t = x_t - x_{t-1}, \qquad (1),$$

где S_t – значение кумулятивной суммы при приходе -того наблюдения; x_t– текущее наблюдение; x_{t-1} – предыдущее наблюдение.

Для обнаружения изменений вычисляются положительная и отрицательная ветви CUSUM:

$$S_t^+ = \boldsymbol{max}(0, S_{t-1}^+ + S_t - k), \qquad (2),$$

$$S_t^- = \boldsymbol{max}(0, S_{t-1}^- - S_t - k), \qquad (3),$$

где k – параметр инерционности, задающий чувствительность метода к малым отклонениям.

Изменение фиксируется, если одна из ветвей превышает порог h. Такой подход позволяет оперативно обнаруживать направленные изменения в сигнале. Благодаря своей простоте, низкой вычислительной сложности (O(1)) и способности работать при поступлении данных последовательно, CUSUM хорошо соответствует требованиям, предъявляемым к алгоритмам реального времени.

Важно отметить, что классическая версия алгоритма CUSUM предназначена для офлайн фиксации точки изменения среднего уровня ряда, то есть момента сдвига статистических характеристик, а не для детекции изменений тренда онлайн.

1.3. Предлагаемые адаптации

В адаптированной версии алгоритма k до момента детекции тренда равно 0. После детекции тренда k выбирается как:

$$k = \boldsymbol{max}(S_{trend}, S_{new}) \cdot \alpha, \qquad (4),$$

где S_{trend} – максимальное S за текущий тренд до детекции; S_{new} – S пришедшее до детекции; α – коэффициент, регулирующий жесткость k.

Конец тренда определяется на последующем сегменте данных: ожидается пока та сумма, благодаря которой был обнаружен тренд (S_t^+или S_t^-) не станет меньше 0. В точке, где это произошло, фиксируется момент окончания тренда, что позволяет выделить полный интервал тренда.

Также, сигнал предварительно сглаживается. В данной работе используются три подхода:

1. Экспоненциальное скользящее среднее (Streaming EMA):

$$y_t = \alpha \cdot x_t + (1 - \alpha) \cdot y_{t-1} \qquad (5),$$

где y_t – сглаженное значение; y_{t-1}– предыдущее сглаженное значение; α – коэффициент, регулирующий влияние новой точки.

2. Взвешенное скользящее среднее (Realtime Weighted MA):

$$y_t = \frac{\sum_{i=0}^{n-1} w_i x_{t-i}}{\sum_{i=0}^{n-1} w_i}, \qquad (6)$$

где w_i –веса, показывающие влияние каждого значения.

3. Фильтр Калмана для потоковых данных (Streaming Kalman Filter):

$$y_t = x_{t-1} + K_t \cdot (x_t - y_{t-1}), \qquad (7)$$

где K_t – коэффициент Калмана, показывающий влияние каждого нового значения.

Каждый из методов имеет тонкую настройку в виде трех наборов гиперпараметров («easy», «middle», «hard»). Пример разных настроек представлен на рис. 1.

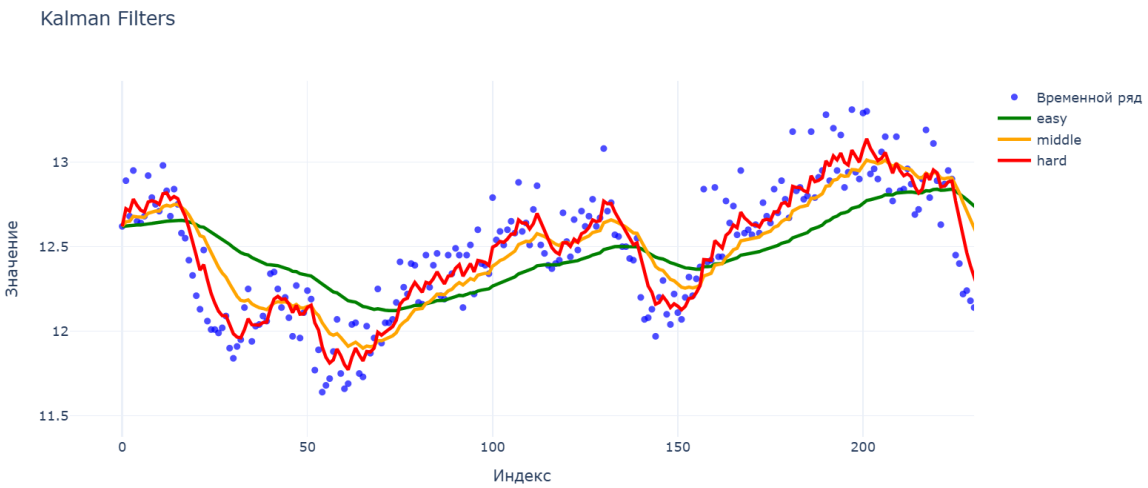

Рис. 1. Примеры различных настроек гиперпараметров фильтра Калмана

Применение предварительного сглаживания позволяет уменьшить влияние кратковременных колебаний и шумовых выбросов. После сглаживания данные поступают на вход адаптированного CUSUM-алгоритма.

На рис. 2 в качестве примера сглаживающей функции был взят фильтр Калмана с настройкой «hard».

Рис. 2. CUSUM алгоритм со сглаживанием через фильтр Калмана

2. Обнаружение изменений статистических характеристик ряда.

2.1. Постановка задачи и требования к алгоритмам

Требования к алгоритмам обнаружения изменений статистических характеристик ряда во многом совпадают с требованиями к детекции трендов. Однако ключевое отличие состоит в том, что алгоритм должен фиксировать не направление изменения, а моменты, когда изменяются базовые статистические показатели ряда — такие как среднее или стандартное отклонение.

2.2. Описание методов

Алгоритм ADWIN (Adaptive Windowing) относится к методам выявления изменений среднего значения временного ряда и основан на хранении адаптивного окна наблюдений. При поступлении нового значения алгоритм добавляет его в окно и проверяет возможность статистически значимого разбиения окна на две части с различающимися средними значениями. В случае обнаружения изменения старые наблюдения, соответствующие предыдущему состоянию системы, удаляются, а размер окна адаптируется к новым условиям.

Для обеспечения вычислительной эффективности ADWIN использует иерархическую структуру бакетов, в которых агрегируются статистики групп

наблюдений. Более старые данные хранятся в сжатом виде, тогда как недавние значения сохраняются с высоким разрешением, что позволяет балансировать между точностью анализа и ограничениями по памяти и вычислительным ресурсам. Однако такая схема агрегации приводит к избыточному сжатию информации: усреднение данных внутри бакетов снижает чувствительность алгоритма к локальным и плавным изменениям статистических характеристик. В результате возможно запаздывание обнаружения момента сдвига среднего значения, особенно в условиях высокочастотных потоков данных или постепенного дрейфа параметров временного ряда.

Для решения указанных проблем была предложена модифицированная версия алгоритма из библиотеки river [3], сохраняющая преимущества ADWIN, но уменьшающая потери информации при сжатии. Во-первых, скорректирована процедура вычисления порога определения статистически значимого различия средних между частями окна ε. Порог рассчитывается по выражению

$$\varepsilon = \sqrt{2 \cdot m \cdot \sigma_t^2 \cdot ln\left(\frac{2ln\, n}{\delta}\right)} + \frac{2}{3} m \cdot ln\left(\frac{2ln\, n}{\delta}\right), \quad (8),$$

где σ_t^2 – текущая оценка дисперсии; n – длина окна; δ – уровень значимости; m – соотношение между средними левого и правого подокон.

В предложенной версии m рассчитывается так, как описано в оригинальной статье [4], что делает механизм адаптации порога более устойчивым и обеспечивает корректную реакцию на реальные изменения статистики ряда.

Во-вторых, при сжатии данных используется преобразование n точек не в одну, а в m усредненных точек с помощью библиотеки tslearn [5].

Пример работы алгоритма HADWIN представлен на рис. 3.

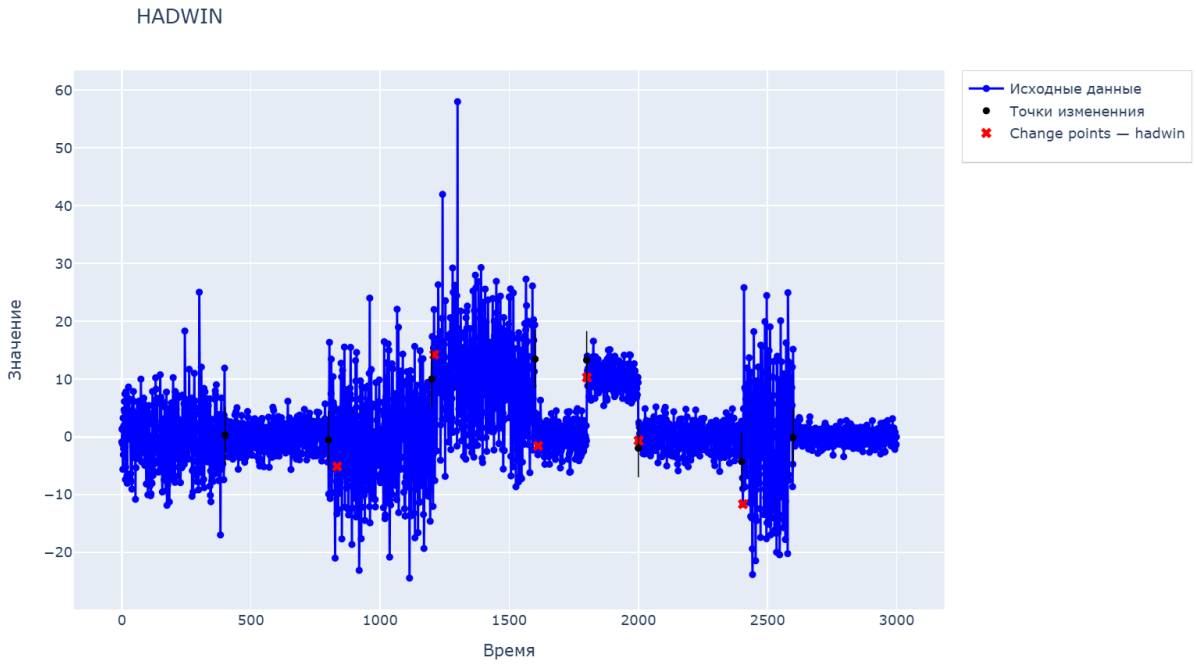

Рис. 3. Пример работы метода HADWIN

Второй тип алгоритмов является BOCPD (Bayesian Online Changepoint Detection) [6]. Он определяет точку изменения как среднего, так и стандартного отклонения. Метод оценивает вероятность смены распределения в режиме реального времени и способен обнаруживать изменения на каждой новой точке данных. Через формулу метод выражается в виде рекуррентного соотношения вероятностей.

$$\mathrm{P}(r_t, x_{1:t}) = \sum_{r_{t-1}} P(r_t|r_{t-1}) P\left(x_t \middle| r_{t-1}, x_t^{(r)}\right) P(r_{t-1}, x_{1:t-1}), \qquad (9),$$

где $P(r_t|r_{t-1})$ – вероятность ошибки; $P\left(x_t \middle| r_{t-1}, x_t^{(r)}\right)$ – вероятность наблюдать текущую точку данных x_t, исходя из длины сегмента r_{t-1} и всех предыдущих наблюдений; $P(r_{t-1}, x_{1:t-1})$ – вероятность, что длина текущего сегмента до предыдущей точки была равна r_{t-1}.

В данной реализации используется механизм очистки буфера данных следующим образом: при достижении n точек в буфере, остаются последние m точек (n>m), первые же

n-m точек удаляются, и m оставленных точек сдвигаются влево, оставляя m-n значений буфера пустыми, где и будут храниться новые точки.

Очистка буфера происходит также при нахождении точки изменения статистических характеристик ряда.

Помимо очистки буфера для корректной отработки алгоритма BOCPD каждое значение смещается на среднее, вычисленное через EMA. Необходимость масштабирования вызвана тем, что BOCPD использует NIG (Normal Inverse Gaussian) распределение для расчета вероятностей, однако тогда точки должны иметь среднее 0.

К третьему типу отнесем статистические тесты: KSWIN (Kolmogorov-Smirnov Windowing) [7] и Ftest. KSWIN использует скользящее окно, разделяя его на две части и сравнивая распределения значений в каждой части с помощью критерия Колмогорова–Смирнова. Если статистика теста превышает заданный порог, фиксируется изменение распределения или дисперсии ряда. Стоит сказать, что KSWIN основан на эмпирической функции распределения поэтому, чтобы сравнить расстояние между двумя кривыми нужно много **(**> 30**)** точек. Из-за большого объема точек, время между реальным изменением и его детектированием может быть заметно выше, чем у других оконных статистических методов. Поэтому в качестве аналога используется оконный Ftest, который использует F-статистику, меняющую распределение в зависимости от числа точек, и тем самым работающего и при малых окнах (< 30 точек).

Вместо проверки всех точек в окне методом Ftest можно находить наиболее подходящую методом наибольшего правдоподобия. В качестве критерия выбран BIC (или AIC) критерий.

Выразим поиск наиболее подходящей точки через формулу:

$$R(\tau) = n \ln \hat{\sigma} - \tau \ln \widehat{\sigma_1} - (n - \tau) \ln \widehat{\sigma_2}, \qquad (10),$$

где $R(\tau)$ — функция от номера элемента с конца; τ — номер элемента с конца;
$\widehat{\sigma}_1$— оценка дисперсии по левому подокну; $\widehat{\sigma}_2$ — оценка дисперсии по правому подокну;
n — общее число точек в буфере.

Таким образом, вначале отбирается точка, в которой наиболее вероятно произошло изменение. А уже потом данная точка передается в Ftest. Алгоритм двойной проверки был назван GrowingBufferChangePointDetector, так как его буфер данных постоянно растет. При достижении определенного числа точек, буфер сжимается до m точек, через библиотеку tslearn. Благодаря двойной проверке уменьшается число ложноположительных детекций.

3. Результаты тестирования на данных.

Помимо стандартных метрик (Precision, Recall, F1-Score) были получены:

Matched Changes - количество реальных изменений в данных, которые алгоритм корректно обнаружил и локализовал;

Detected Changes - общее количество задетектированных точек;

False Positives – количество ложных точек детектирования;

Over Detection - отношение обнаруженных изменений к реальному числу изменений.

3.1. Результаты тестирования CUSUM

Для тестирования методов было сгенерировано 100 синтетических временных рядов, каждый длиной 3000 точек с равномерной дискретизацией. Каждый ряд содержал 12 искусственно добавленных трендовых компонент с различной амплитудой и продолжительностью на фоне базового стационарного процесса, не имея выраженной сезонности.

Таблица 1

Method	Target	Matched Changes	Detected Changes	False Positives	Over Detection	Recall	Precision	F1
EMA	End	3,24	11,51	8,2	0,96	0,27	0,28	0,27
EMA	Start	2,94	17,65	14,71	1,47	0,245	0,17	0,20
Kalman	End	3,24	11,51	8,2	0,96	0,27	0,28	0,27
Kalman	Start	2,94	17,65	14,71	1,47	0,245	0,17	0,20
WMA	End	3,24	11,51	8,2	0,96	0,27	0,28	0,27
WMA	Start	2,94	17,65	14,71	1,47	0,245	0,17	0,20
Эталон	Любой	12,0	12,0	0	1	1	1	1

3.2. Результаты тестирования CPD

В табл. 2 результаты тестов на синтетических данных, моделирующих поведение данных телеметрии. Все было сгенерировано 100 временных рядов по 3000 точек в каждом, содержавших по 13 – 14 точек изменения статистических характеристик временного ряда, расположенных случайно. Основной ряд не содержал выраженной сезонности.

Таблица 2

Method	Matched Changes	Detected Changes	False Positives	Over Detection	Recall	Precision	F1
ADWIN (ориг.)	2,21	9,1	6,89	0,68	0,16	0,24	0,19
HADWIN (адапт.)	3,89	***9,42***	5,53	***0,70***	0,29	0,41	0,34
KSWIN (ориг.)	0,0	4,4	***4,4***	0,34	0,0	0,0	0,0
Ftest (адапт.)	8,8	22,3	13,5	1,68	0,65	0,39	0,49
GBCPD (адапт.)	2,0	20,7	18,7	1,61	0,15	0,1	0,12
BOCPD (ориг.)	0,0	20,9	20,9	1,58	0,0	0,0	0,0
BOCPD (адапт.)	***8,92***	19,58	10,66	1,45	***0,66***	***0,46***	***0,54***
Эталон	13,45	13,45	0,0	1	1	1	1

Выводы

1. Полученные результаты демонстрируют эффективность предложенных адаптаций алгоритмов как для выявления трендов во временных рядах (адаптированный алгоритм CUSUM), так и для обнаружения точек изменения статистических характеристик (адаптированные алгоритмы HADWIN, BOCPD, F-test и GBCPD). В рамках анализа метода CUSUM было проведено сравнение различных вариантов предварительного сглаживания входных данных. На основании результатов, представленных в табл. 1, показано, что используемые методы сглаживания в

большинстве случаев являются взаимозаменяемыми и не оказывают существенного влияния на итоговую точность обнаружения трендов.

2. Методы обнаружения изменений статистических характеристик временного ряда сравнивались как между собой, так и с их оригинальными реализациями, а также с алгоритмом KSWIN. Согласно результатам, представленным в табл. 2, наихудшие показатели продемонстрировал оригинальный алгоритм KSWIN, что обусловлено причинами, рассмотренными в разделе 2.2. Наилучшие значения используемых метрик показал адаптированный алгоритм BOCPD. Вместе с тем, остальные предложенные модификации также продемонстрировали стабильные и конкурентоспособные результаты, что позволяет рассматривать их в качестве перспективных решений для практического применения.

3. Следует отметить, что в рамках данной работы тестирование алгоритмов проводилось на синтетических наборах данных, что обусловлено сложностью объективной оценки корректности обнаружения изменений на реальных телеметрических данных, где истинные моменты сдвигов, как правило, неизвестны и могут быть определены лишь экспертным путем. В дальнейшем планируется проведение экспериментов на реальных потоках телеметрии с привлечением экспертной оценки, а также дополнительная валидация предложенных методов в условиях промышленной эксплуатации.

Список литературы

1. *Краснов А.Н.* Сбор и маршрутизация данных на удаленных кустовых площадках [Текст] / А.Н. Краснов, М.Ю. Прахова, Ю.В. Калашник [и др.] // Нефтяное хозяйство. – 2025. – № 9. – С. 101-107. – https://doi.org/10.24887/0028-2448-2025-9-101-107

2. *Page E.S.* Continuous Inspection Scheme [Текст] / E. S. Page // *Biometrika.* — 1954. — № ***41*** *(1/2)*. — С. *100–115*.

3. *River.* Online machine learning in Python [Электронный ресурс]. — URL: https://riverml.xyz/latest/api/drift/ADWIN/ (дата обращения: 23.12.2025).

4. *Bifet, A.* Learning from Time-Changing Data with Adaptive Windowing / Bifet, A. Gavalda R. // Society for Industrial and Applied Mathematics —2007. — [Электронный ресурс]. — URL: https://epubs.siam.org/doi/10.1137/1.9781611972771.42 (дата обращения: 23.12.2025).

5. tslearn [Электронный ресурс]. — URL: https://tslearn.readthedocs.io/en/stable/gen_modules/preprocessing/tslearn.preprocessing.TimeSeriesResampler.html (дата обращения: 23.12.2025).

6. *Adams R P.* Bayesian Online Changepoint Detection [Электронный ресурс] / R. P. Adams, D. J. C. MacKay — URL: https://arxiv.org/pdf/0710.3742 (дата обращения: 23.12.2025).

7. *River.* Online machine learning in Python [Электронный ресурс]. — URL: https://riverml.xyz/latest/api/drift/KSWIN/ (дата обращения: 23.12.2025).