

Обнаружение зон аномальных изменений в многомерных рядах коммерческого учета тепловой энергии

О.В. Стукач^{1,2}, О.О. Еремина³

¹Новосибирский государственный технический университет, Новосибирск, Россия

²Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

³Национальный исследовательский Томский государственный университет Томск, Россия

Аннотация – Обнаружение аномальных значений в таблицах данных, полученных с тепловычислителей, необходимо для повышения качества учёта потребления тепловой энергии. К зонам аномальных значений приводят поломки приборов учёта, ошибки передачи, человеческий фактор и множество других причин. И если обнаружение грубых ошибок измерений статистическими методами сейчас не вызывает особых сложностей, обнаружение точек аномальных изменений и резких изменений в рядах – это более сложная задача, которая привлекает внимание исследователей во многих предметных областях, где активно внедряется интеллектуальный анализ данных. В конструировании погодных регуляторов отопления многоэтажных жилых зданий одним из задающих воздействий является температура наружного воздуха с резкими перепадами, вызванными, главным образом, глобальным изменением климата. В статье предлагается использовать адаптивную состоятельную меру обнаружения аномальных изменений ряда, основанную на выборе функций, наименее уклоняющихся от многомерных рядов. Мера основана на различиях в распределении вероятностей до и после зон аномальных изменений. Предложенный подход обнаружения устойчив к погрешностям измерения и может быть использован для повышения точности оценки работы регуляторов тепловой энергии. На опубликованных датасетах, находящихся в открытом доступе, проведено тестирование разработанного алгоритма. Поскольку аномалии в рядах зачастую встречаются в данных автоматических систем контроля и передачи данных учёта энергии, для тестирования взяты распространенные датасеты. Разработанный алгоритм по времени анализа данных сопоставим с популярным алгоритмом SVM, а по F-статистике демонстрирует преимущество перед другими известными методами. Предложенный способ обнаружения зон изменения в рядах является многообещающим в задачах экономии энергии и продемонстрирован с помощью обширных вычислительных экспериментов на реальных наборах данных теплоучёта.

Ключевые слова – качество данных, оценка параметров, алгоритм обработки сигналов, идентификация системы управления, коммерческий учет потребления энергоресурсов.

ВВЕДЕНИЕ

Задача анализа дискретных временных рядов с целью определения природы ряда и прогнозирования значений часто встречается на практике и достаточно хорошо исследована для рядов с разными особенностями поведения [1]. Разработаны многочисленные методы расчёта трендов, оценки периодичности, выделения компонент ряда и др [2]. Менее исследованы способы обнаружения точек и зон аномального поведения рядов, если ошибки или особые значения не превышают обычной дисперсии данных [3]. Особую актуальность приобретают методы анализа связанных рядов, или многомерных рядов [4, 5] и модели обнаружения связи между временными рядами [6]. Многие практические задачи в моделировании рядов приводят к необходимости не только прогнозирования, но и разного рода статистических оценок в пределах ряда [7]. Также важно учитывать ситуацию, когда источник данных нестационарный, поскольку характер источника данных может меняться с течением времени в реальных условиях. На практике достаточно сложно получить длинные ряды, а зашумленные зоны аномалий имеют большую продолжительность [8]. Известные алгоритмы поиска аномалий плохо справляются с их точным анализом при большой зашумленности и в условиях слабых статистических различий между измерениями, особенно когда их достаточно много.

Актуальной задачей анализа аномалий в рядах является коммерческий учёт тепловой энергии, где должно контролироваться качество данных, используемых в автоматизированных системах начислений платы за потреблённые энергоресурсы [9]. Зачастую некорректные показания приборов учёта не выявляются или выявляются не своевременно. Увеличению ошибок и снижению качества способствует интеграция данных из разных источников. Иногда данные просто не передаются на протяжении некоторого времени [10].

Аномалии скрыты шумом и обычными вариациями. Это характерно для многих данных, например учёта потреблённой электрической и тепловой энергии [11].

Для повышения качества данных в статье [12] предлагается двухэтапная очистка временных рядов показаний приборов учёта: сначала применяется иерархическая кластеризация для выделения грубых ошибок, а затем в результате статистической обработки выявляются выбросы. Но окончательное решение об ошибочности наблюдений принимается экспертным путём. Поскольку полная автоматизация предполагает исключение экспертных оценок, необходимо разрабатывать новые способы выявления аномалий и последующей очистки данных.

I. ПОДХОДЫ К ОБНАРУЖЕНИЮ АНОМАЛИЙ

Известные подходы к обнаружению аномалий можно разделить на три класса: методики, основанные на подпространстве [13], на соседстве [14] и комбинированные [15]. Среди последних можно выделить способ оценки плотности аномалий – Local Outlier Factor (LOF), метод машинного обучения Support Vector Machine (SVM), Elliptic Envelope – это обнаружение аномалий по оценке многомерного нормального распределения, Subspace Outlier Detection (SOD) – метод кластеризации с выделением аномалий.

Многообещающим методом машинного обучения для обнаружения точек аномальных изменений для многомерных рядов является анализ стационарного подпространства (stationary subspace analysis, SSA) [16]. В этом методе многомерные ряды разлагаются на стационарные и нестационарные компоненты, а затем в нестационарном подпространстве могут быть обнаружены точки изменения. Поскольку метод SSA позволяет уменьшить размерность данных без потери зон аномальных изменений, он может значительно повысить скорость обнаружения. Недостатком метода SSA является то, что при малой размерности пространства значений рядов необходимо вычислять логарифм сингулярной ковариационной матрицы. Для точной факторизации рядов на компоненты требуется относительно большое число обучающих выборок. Можно с уверенностью считать, что метод SSA непригоден для анализа ограниченного объёма данных учёта тепловой энергии для выбранного теплосчётчика, так как число значений рядов равно числу дней отопительного сезона.

Идеология применения методов машинного обучения с учителем развита в работе [17], где ставится задача построения классификатора, дискриминирующая ряды, если они произошли от разных источников. Предложенный авторами алгоритм работает на синтетических данных с биномиальным распределением, но для реальных данных потребления тепловой энергии классификатор пропускает больше половины случаев аномалий.

Если предположить, что существует точка аномалии $q(t)$ в момент t , распределение вероятностей данных $q(t-1)$, $q(t-2)$,... в прошлом $t-1$, $t-2$,... должно отличаться от распределения $q(t+1)$, $q(t+2)$,... в будущем $t+1$, $t+2$. Если это различие значимо, можно считать, что зона аномалии обнаружена. На основе этого предположения были предложены различные подходы к обнаружению точки изменения с кумулятивной суммой (likelihood ratio test) и на основе общеизвестного отношения правдоподобия [18]. Здесь логарифм отношения правдоподобия между двумя распределениями вероятностей используется в качестве меры обнаружения изменения, где каждая плотность вероятности оценивается независимо. Но поскольку сама по себе оценка плотности вероятности в условиях погрешностей измерения является непростой задачей, известные подходы работают плохо, особенно на реальных данных коммерческого учёта тепловой энергии [19]. Кроме того, методы оценки плотности вероятности в зависимых рядах довольно сложны и непрактичны для применения в регуляторах, так как требуют сложных компьютерных вычислений, а результат получается весьма приближённым [20]. К тому же регулятор может отработать такую ошибку без потери устойчивости.

Чтобы избежать использования оценки плотности вероятности, были предложены прямые подходы к определению зоны изменения, основанные на состоятельных мерах зависимости [21, 22], а также оценке соотношения вероятностных распределений непосредственно без использования оценки плотности [23, 24] или методами машинного обучения [25]. Они дают оптимальную скорость работы алгоритмов, но относительно низкую точность из-за влияния погрешности измерений.

II. ПОСТАНОВКА ЗАДАЧИ

Перспективным представляется подход, использующий критерий независимости Гильберта-Шмидта (КНГШ) на основе ядра [26]. КНГШ работает на основе выборок объекта и соответствующих ему бинарных переменных $+1$ (будущие данные) и -1 (прошлые данные) с разными распределениями вероятностей. Преимущество оценки КНГШ по сравнению с рассмотренными ранее мерами обнаружения заключается в том, что он может включать выбор признака во время оценки меры обнаружения. Сначала выбираются значения, которые отвечают за резкое изменение ряда, а затем вычисляется оценка КНГШ, используя эти выбранные значения. Зашумленность данных оказывает меньшее влияние на результат, чем рассмотренные методы, и предлагаемый метод перспективен в рассматриваемой предметной области.

Пусть $q(i)$ – m -размерная выборка ряда в момент i . Если выборка сделана из одного и того же

распределения, – в ряде аномального изменения нет, в противном случае есть. Методом скользящего окна в момент i из $q(i)$ извлекаются n -мерные выборки $q(i-j+1)|_{j=1}^n$. Две непересекающиеся последовательности $q(t)$ и $q(t+n)$ отметим как $r(i)=1$ и $r(i+n)=-1$ соответственно. Рассмотрим объединённый ряд $Q(i): \{q(i+n-j+1)\}_{j=1}^{2n}$, для которого $r(i)=\pm 1$. Обнаружение зон аномального изменения ряда заключается в вычислении меры M зависимости между q и r из $Q(i)$, такой, что

$$\begin{cases} Q(i) < s \Rightarrow \text{изменений нет;} \\ Q(i) \geq s \Rightarrow \text{аномалия обнаружена,} \end{cases}$$

где s – порог обнаружения зоны аномалии. Поскольку данные, как правило, измеряются с большими погрешностями и зашумлены, факт обнаружения сильно зависит от s .

Представим КНГШ $H[\cdot]$ в качестве меры зависимости $Q(i)$ в виде [23]:

$$Q(i) = \sum_{k=1}^n \xi_k H[B_k(i), C],$$

$$\text{где } \xi_k > 0, \quad B(i) = \begin{bmatrix} q(i-n+1) \\ q(i-n+2) \\ \dots \\ q((i+n-1)) \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 1 \\ \dots \\ -1 \end{bmatrix},$$

$H[B_k(i), C] = \text{tr}(D_k^H E^H)$, $D^H = \Gamma D \Gamma$, $E^H = \Gamma E \Gamma$, D^H , E^H – нормализованные центрированные матрицы ядра, Γ – матрица Грама.

КНГШ является состоятельной мерой зависимости случайных величин x и y , удовлетворяющей условию $0 \leq H[x, y] \leq 1$ и равен нулю тогда и только тогда, когда x и y статистически независимы. Если используется гауссово ядро, $H[x, y] \rightarrow 0$ [27].

При разработке эвристических алгоритмов поиска аномалий можно принять $\xi_k \equiv 1$. Но для автоматического поиска аномалий целесообразно использовать какой-либо регуляризующий метод для поиска ξ_k , поскольку равные весовые коэффициенты оценок КНГШ не являются хорошим выбором для многомерных рядов, особенно при зашумленных данных, например

$$\min_{\xi} \left\| E^H - \sum_k \xi_k D_k^H \right\|^2 + \beta \|\xi\|, \quad \xi_{k=1}^n \geq 0,$$

где β – параметр регуляризации.

Здесь матрица ядра E^H аппроксимируется линейной функцией элементов матрицы D_k^H . После нахождения ξ_k эти элементы нормализуются. Поскольку

$$\min_{\xi} \left\| E^H - \sum_k \xi_k D_k^H \right\|^2 = H[C, C] - 2 \sum_{k=1}^n \xi_k H[B_k, C] + \sum_{i,j} \xi_i \xi_j H[B_i B_j]$$

а первое и третье слагаемые равны нулю ввиду взаимной независимости, для минимизации целевой функции необходимо найти максимум

$$Q(i) = \sum_{k=1}^n \xi_k H[B_k(i), C].$$

В некоторой степени этот вывод очевиден, ведь если k -е наблюдение $B_k(i)$ имеет сильную зависимость от C , КНГШ $H[B_k(i), C]$ принимает большое значение и, следовательно, это наблюдение сохраняется в процессе регуляризации. Если же $B_k(i)$ и C взаимно независимы, КНГШ стремится к нулю, и такое наблюдение вскоре удаляется. В результате остаются только те наблюдения, которые приводят к резкому изменению результирующей переменной. Избыточные значения, соответствующие слагаемым $\xi_i \xi_j H[B_i B_j]$, также удаляются в процессе регуляризации.

III. АПРИОРНАЯ ИНФОРМАЦИЯ О ФУНКЦИЯХ

Если имеется предварительная информация о свойствах рядов, её можно использовать для выбора функций в КНГШ. Например, при наличии в базе данных нескольких энергоэффективных домов резкое изменение теплотребления происходит в каждом из них в близкий момент времени [28]. Поэтому для повышения эффективности регуляризации вполне разумно выбрать ряды, соответствующие кластеру энергоэффективных домов, исходя из каких-то предварительных предположений. Этому может помочь знание города [29], и с этой целью можно использовать регуляризатор группового типа эластичной сети:

$$\min_{\xi} \frac{1}{2} \left\| E^H - \sum_k \xi_k D_k^H \right\|^2 + \mu \sum_{l=1}^L \|\gamma_l\| \quad (1)$$

где $\gamma = [\gamma_1, \gamma_2, \dots, \gamma_L]^T$ – l -я группа переменных, L – количество групп, μ – параметр регуляризации. Для

минимизации (1) предлагается следующий пошаговый алгоритм (Рис. 1).

1 Пусть $D^H \in R$ – вектор начальных приближений
2 $D_0^H \leftarrow \min \sum_{l=1}^L \ \gamma_l\ $
3 Цикл $l=1, L$
Если $\left\ E^H - \sum_k^n \xi_k D_k^H \right\ ^2 \leq \mu \sum_{l=1}^L \ \gamma_l\ $
$\zeta_l \leftarrow 0$
иначе $\zeta \leftarrow \min \sum_{l=1}^L \ \gamma_l\ $
конец цикла
4 Повторить шаги 2, 3 до достижения критерия

Рис. 1. Алгоритм регуляризации

IV. ВЫБОР ЯДРА

КНГШ изменяется в зависимости от параметров и вида ядра. Поэтому входные параметры ядра для всех функций должны быть фиксированы. Сначала проводится нормализация характеристик q к

распределению с нулевым средним q_0 и стандартным отклонением σ , равным единице. Затем для всех ядер используются одни и те же параметры. Для входных переменных q используется гауссово ядро:

$K(q, q_0) = \exp\left[-\frac{(q - q_0)^2}{2\sigma^2}\right]$, где $\sigma = 1$. Для переменной r используется δ -ядро,

$$K(r, r_0) = \begin{cases} 1, & r = r_0 \\ 0, & r \neq r_0 \end{cases}$$

V. ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ С НАБОРОМ СИНТЕТИЧЕСКИХ ДАННЫХ

Для проверки работоспособности алгоритма были выбраны датасеты из открытых источников [28, 30–35]. Для КНГШ использовалось гауссово ядро, параметр регуляризации $\mu = 0,1$.

Для оценки точности C и полноты R предсказания аномалий используется F -метрика, известная из задач классификации:

$$F = 2CR/(C + R), F \in [0,1],$$

где $C = TP/(TP + FP)$, $R = FP/(TN + FN)$, а значения параметров TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative) соответствуют Табл. I.

ТАБЛИЦА I
F-МЕТРИКИ

Параметры	Действительное наличие аномалии	Предсказание аномалии
TP – правильно определённые аномалии	1	1
TN – правильно определённые нормальные данные	0	0
FP – правильные значения, которые неправильно определены как аномальные	0	1
FN – аномалии, которые определены как правильные значения	1	0

Известно, что чем ближе значение F к единице, тем больше вероятность того, что будут корректно обнаружены аномалии в данных. Результаты тестирования представлены на Рис. 2.

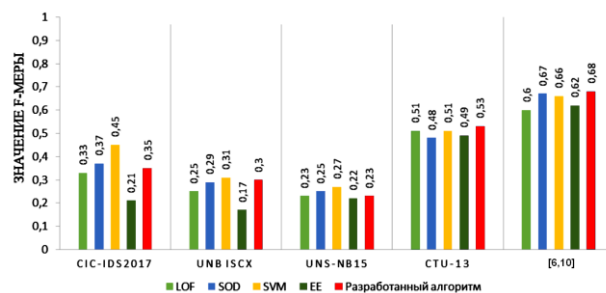


Рис. 2. Значения F-мер

Из Рис. 2 видно, что по F -оценке разработанный алгоритм в среднем обнаруживает аномалии чуть менее точно, чем алгоритм SVM, но по крайней мере в 1,1 раз лучше, чем LOF.

Кроме точности обнаружения аномалий в режиме реального времени важна оценка скорости работы алгоритмов. Производительность оценивалась по времени выполнения программы до полного обнаружения аномалий (Рис. 3).

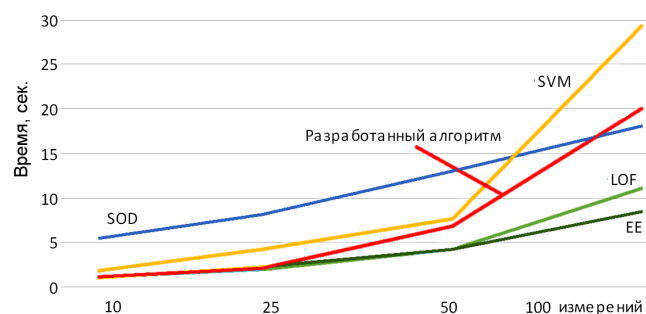


Рис. 3. Зависимости времени анализа данных от количества измерений

Из графиков на Рис. 3 видно, что разработанный алгоритм в целом сопоставим по производительности с другими алгоритмами поиска аномалий, и наряду с

ними может использоваться при исследовании многомерных рядов данных.

VI. ОБСУЖДЕНИЕ

Из проведённого тестирования на опубликованных размеченных данных из открытых источников с указанием мест аномалий [30–35] можно сделать следующие выводы.

1) Алгоритм имеет преимущество в производительности по сравнению с алгоритмом SVM, который по рассчитанной F -метрике сравним с разработанным.

2) Крайне неэффективным показал себя алгоритм Elliptic Envelope. По-видимому, он совсем непригоден для выявления аномалий. Его точность почти в половину ниже, чем у SVM и предлагаемого алгоритма.

3) Предложенный алгоритм вполне может конкурировать с другими популярными подходами к обнаружению аномалий, как по точности обнаружения, так и по времени анализа данных.

ЗАКЛЮЧЕНИЕ

Обнаружение зон аномальных изменений в многомерных рядах привлекает значительное внимание в области интеллектуального анализа данных. Эта задача имеет большое количество практических приложений, в частности, повышение качества учёта потребления тепловой энергии.

В статье исследован подход к обнаружению аномальных изменений в рядах на основе [17], использующий критерий независимости Гильберта-Шмидта (КНГШ) на основе ядра [26], и который работает на основе выборок и соответствующих ему бинарных данных с разными распределениями вероятностей. Преимущество предлагаемого подхода перед существующими заключается в том, что оценка наличия аномалии проводится с использованием признаков, которые важны для резких изменений. Предлагаемый подход более устойчив к зашумлению данных, чем существующие методы. Проведением численных экспериментов с синтетическими и реальными наборами данных показана перспективность предлагаемого подхода к обнаружению зон аномальных изменений.

ЛИТЕРАТУРА

- [1] Крюков Ю.А., Чернягин Д.В. ARIMA – модель прогнозирования значений трафика // Информационные технологии и вычислительные системы. – 2011. – № 2. – С. 41–49. – eLIBRARY ID: 17805262. – EDN: OZQLLR.
- [2] Газиров Д. И. Обзор методов статистического анализа временных рядов и проблемы, возникающие при анализе нестационарных временных рядов // Научный журнал: Олимп (Иваново). – 2016. – № 3(4). – С. 9–14. – eLIBRARY ID: 25631242.
- [3] Zhang W., Li T., He P., Yang Y., Wang S. An Outlier Suppression and Adversarial Learning Model for Anomaly Detection in Multivariate Time Series // Entropy. – 2025. – Vol. 27. – № 11. – P. 1151. – Doi: 10.3390/e27111151.
- [4] Swami A., Giannakis G., Shamsunder S. Multichannel ARMA processes // IEEE Transactions on Signal Processing. – 1994. – Vol. 42. – № 4. – Pp. 898–913. – Doi: 10.1109/78.285653.
- [5] Ivanovic M., Kurbalija V. Time series analysis and possible applications / 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). – Opatija, Croatia, May 30–June 3, 2016. – Pp. 473–479. – Doi: 10.1109/MIPRO.2016.7522190.
- [6] Усманова К.Р., Стрижов В.В. Модели обнаружения зависимостей во временных рядах в задачах построения прогностических моделей // Системы и средства информатики. – 2019. – Т. 29. – № 2. – С. 12–30. – eLIBRARY ID: 41273551.
- [7] Осовский А.В., Кутузов Д.В., Стукач О.В. Анализ моделей трафика, создаваемого устройствами интернета вещей / Динамика систем, механизмов и машин. – 2019. – Т. 7. – № 4. – С. 220–226. – ISSN: 2310-9793. – eLIBRARY ID: 41432222.
- [8] Кузнецов М.П., Ивкин Н.П. Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию // Машинное обучение и анализ данных. – 2015. – Т. 1. – № 11. – С. 1471–1483. – eLIBRARY ID: 24931576.
- [9] Савкина А.В., Федосин А.С. Проблемы качества данных в автоматизированных системах коммерческого учёта потребления энергоресурсов // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 2. – С. 158–164. – eLIBRARY ID: 21716788. – EDN: SHFYGT.
- [10] Стукач О.В., Габитова О.О. Демонстрация результатов энергосберегающего регулирования тепловой энергии на примере одного здания // Автоматика и программная инженерия. – 2023. – № 2(44). – Pp. 158–164. – eLIBRARY ID: 54207953.
- [11] Попов И.Ю., Стукач О.В., Зорин П.А. Оценка качества данных коммерческого учёта тепловой энергии методом полной вариации распознавания ошибок // Динамика систем, механизмов и машин. – 2021. – Т. 9. – № 3. – С. 117–121. – DOI: 10.25206/2310-9793-9-3-117-121.
- [12] Федосин А.С., Федосин С.А. Очистка входных данных в автоматизированных системах контроля и учёта энергоресурсов // Инфокоммуникационные технологии. – 2016. – Т. 14. – № 2. – С. 162–168.
- [13] Bacher M., Ben-Gal I., Shmueli E. Subspace selection for anomaly detection: An information theory approach. / IEEE International Conference on the Science of Electrical Engineering (ICSEE). – 16–18 Nov. 2016. – Eilat. – Israel, id.50. – P. 1–5. – Doi: 10.1109/icsee.2016.7806077.
- [14] Dutta J.K., Banerjee B., Reddy C.K. RODS: Rarity based Outlier Detection in a Sparse Coding Framework // IEEE Transactions on Knowledge and Data Engineering. – 2016. – Vol. 28. – № 2. – Pp. 483–495. – Doi: 10.1109/TKDE.2015.2475748.
- [15] Shebuti R. Sequential Ensemble Learning for Outlier Detection: A Bias-Variance Perspective // 2016 IEEE 16th International Conference on Data Mining (ICDM). – 2016. – Pp. 1167–1172. – arXiv:1609.05528. – doi: 10.48550/arXiv.1609.05528.
- [16] Blythe D., Bunau P., Meinecke F.C., Muller K.R. Feature Extraction for Change-Point Detection using Stationary Subspace Analysis // IEEE Transactions on Neural Networks and Learning Systems. – 2012. – Vol. 23. – № 4. – Pp. 631–643. – arXiv:1108.2486v1. – Doi: 10.1109/TNNLS.2012.2185811.
- [17] Hido S., Ide T., Kashima H., Kubo H., Matsuzawa H. Unsupervised change analysis using supervised learning / Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). – 2008. – Part of the book series: Lecture Notes in Computer Science (LNAI, vol. 5012). – Pp. 148–159.
- [18] Gustafsson F. The marginalized likelihood ratio test for detecting abrupt changes // IEEE Transactions on Automatic Control. – 1996. – Vol. 41. – № 1. – Pp. 66–78. – Doi: 10.1109/9.481608.
- [19] Попов И.Ю., Стукач О.В., Зорин П.А. Применение метода полной вариации распознавания ошибок к исследованию

- данных коммерческого учета тепловой энергии // Автоматика и программная инженерия. – 2021. – № 4(38). – С. 55-61.
- [20] Лагутин М.В. Наглядная математическая статистика. – М.: ВИНОМ. Лаборатория знаний, 2009. – 472 с.
- [21] Ershov I.A., Stukach O.V., Aimagambetova R.Z. The Peculiar Measure Identifying of the Temperature Leap in the Distributed Raman Sensors // Dynamics of Systems, Mechanisms and Machines (Dynamics). – 2021, 9-11 Nov., Omsk, Russian Federation. – Pp. 1-4. – Doi: 10.1109/Dynamics52735.2021.9653721.
- [22] Ершов И.А., Стукач О.В., Цыденжапов И.Б., Сычев И.В. Идентификация слабого температурного перехода в оптоволокне при помощи дивергенции Цаллиса / Электронные средства и системы управления: материалы докладов XVI Международной научно-практической конференции (18-20 ноября 2020 г.): в 2 ч. – Ч. 1. – Томск: В-Спектр. – 2020. – С. 146-148.
- [23] Nguyen X.L., Wainwright M.J., Jordan M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization // IEEE Transactions on Information Theory. – 2010. – Vol. 56. – № 11. – Pp: 5847–5861. – arXiv:0809.0853v2. – Doi: 10.1109/TIT.2010.2068870.
- [24] Kawahara Y., Sugiyama M. Change-point detection in time-series data by direct density-ratio estimation / SIAM International Conference on Data Mining (SDM 2009). – April 30 - May 2. – 2009. – Sparks, Nevada, USA. – Pp. 389-400. – Doi:10.1137/1.9781611972795.34.
- [25] Liu S., Yamada M., Collier N., Sugiyama M. Change-point detection in time-series data by relative density ratio estimation // Neural Networks. – 2013. – 16 Jan. – arXiv:1203.0453v2. – Doi: 10.48550/arXiv.1203.0453.
- [26] Gretton A., Bousquet O., Smola A., Scholkopf B. Measuring statistical dependence with Hilbert-Schmidt norms // ALT papers. – 2005. – Pp. 63-77.
- [27] Steinwart I. On the influence of the kernel on the consistency of support vector machines // Journal of Machine Learning Research. – 2001. – № 2. – Pp. 67-93.
- [28] Zorin P., Stukach O. Data of heating meters from residential buildings in Tomsk (Russia) for statistical modeling of the thermal characteristics of buildings / IEEE Dataport. – 2020. – [Online]. Available: <http://dx.doi.org/10.21227/3r4e-ch18>.
- [29] Зорин П.А., Стукач О.В. Анализ влияния погодных условий на динамику тепловой энергии в жилом фонде города Томск / Инновационные, информационные и коммуникационные технологии: сборник трудов XVII Международной научно-практической конференции. – М.: Ассоциация выпускников и сотрудников ВВИА им. проф. Жуковского. – 2020. – С. 365-368. – eLIBRARY ID: 45557868.
- [30] Набор данных CIC-IDS2017. Режим доступа: <https://www.unb.ca/cic/datasets/ids-2017.html>.
- [31] Stukach O., Zorin P. Long-Term Data from the Heat Meters in Residential Buildings Depending on the Outside Temperature and Characteristics of Buildings / IEEE Dataport. – April 13, 2021. – Doi: 10.21227/cw53-rr81.
- [32] Набор данных UNB ISCX. Режим доступа: <https://www.unb.ca/cic/datasets/ids.html>.
- [33] Набор данных UNS-NB15. Режим доступа: <https://research.unsw.edu.au/projects/unsw-nb15-dataset>.
- [34] Набор данных CTU-13. Режим доступа: <https://www.kaggle.com/datasets/doogla/ctu13>.
- [35] Зорин П.А., Стукач О.В. База данных потребления тепловой энергии многоэтажными жилыми зданиями в зависимости от метеорологических факторов и характеристик зданий. – Рег. № 25452. – Doi: 10.12731/ofernio.2025.25452.

Информация об авторах

Стукач Олег Владимирович, д.т.н., профессор
кафедры Защиты информации Новосибирского

государственного технического университета,
г. Новосибирск, Россия, профессор департамента
Электронной инженерии Московского института
электроники и математики им. А.Н. Тихонова
Национального исследовательского университета
«Высшая школа экономики», г. Москва, Россия, e-mail:
tomsk@ieee.org, ORCID: 0000-0001-6845-4285.

Еремина Олеся Олеговна, студент Высшей ИТ-
школы Национального исследовательского Томского
государственного университета.

Detection of zones of abnormal changes in multidimensional series of commercial thermal energy metering

Oleg Stukach ^{1,2}, Olesya Eremina³

¹National Research University Higher School of
Economics, Moscow, Russia

²Novosibirsk State Technical University, Novosibirsk,
Russia

³National Research Tomsk State University, Tomsk, Russia

Abstract – The detection of abnormal values in the data tables from heat meters is necessary to improve the quality of accounting for heat energy consumption. Meter failures, transmission errors, human factors, etc. lead to abnormal value zones. Unless the detection of rough measurement errors by statistical methods is not particularly difficult now, the detection of points of abnormal changes and abrupt changes in the series is a more difficult problem. It attracts the attention of researchers in many fields where data analysis is actively being implemented. In the design of weather regulators for heating multi-storey residential buildings, one of the input is the temperature of the outside air with sudden fluctuations mainly follows from global climate change. We suggest the use an adaptive consistent measure for detecting abnormal series changes based on the selection of functions that least deviate from multidimensional series. The measure is based on differences in the probability distribution before and after the zones of abnormal changes. The proposed detection is resistant to measurement errors and can be used to improve the accuracy of thermal energy regulators. The designed algorithm has been tested on open datasets. One of the most common sets was used for testing because anomalies in the series are often found in the data of automatic systems for monitoring and transmitting energy metering data. The designed algorithm was comparable to the popular SVM algorithm in terms of data analysis time. It demonstrates an advantage over other well-known methods in terms of F-statistics. The proposed method for detecting zones of variation in series is promising in energy-saving tasks and has been demonstrated using extensive computational experiments on real heat accounting data sets.

Keywords – data quality, parameter estimation, algorithm of signal processing, system identification, energy resources accounting.

REFERENCES

- [1] Kryukov Yu.A., Chernyagin D.V. An ARIMA model for forecasting a values of network traffic // *Informatisionnie tekhnologii i vichislitelnie sistemi.* – 2011. – № 2. – Pp. 41-49. – eLIBRARY ID: 17805262. – EDN: OZQLLR (in Russian).
- [2] Gazirov D. I. Review of methods of statistical analysis of time series and problems arising in the analysis of non-stationary time series // *Scientific journal: Olympus (Ivanovo).* – 2016. – № 3(4). – Pp. 9-14. – eLibrary ID: 25631242 (in Russian).
- [3] Zhang W., Li T., He P., Yang Y., Wang S. An Outlier Suppression and Adversarial Learning Model for Anomaly Detection in Multivariate Time Series // *Entropy.* – 2025. – Vol. 27. – № 11. – P. 1151. – Doi: 10.3390/e27111151.
- [4] Swami A., Giannakis G., Shamsunder S. Multichannel ARMA processes // *IEEE Transactions on Signal Processing.* – 1994. – Vol. 42. – № 4. – Pp. 898-913. – Doi: 10.1109/78.285653.
- [5] Ivanovic M., Kurbalija V. Time series analysis and possible applications / 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). – Opatija, Croatia, May 30–June 3, 2016. – Pp. 473-479. – Doi: 10.1109/MIPRO.2016.7522190.
- [6] Usmanova K.R., Strizhov V.V. Models for detecting dependencies in time series in the tasks of constructing predictive models // *Computer science systems and tools.* – 2019. – Vol. 29. – № 2. – Pp. 12-30. – eLIBRARY ID: 41273551 (in Russian).
- [7] Osovskiy A.V., Kutuzov D.V., Stukach O.V. Analyze traffic patterns generated by IoT devices // *Dynamics of Systems, Mechanisms and Machines.* – 2019. – Vol. 7. – № 4. – Pp. 220-226. – ISSN: 2310-9793. – eLIBRARY ID: 41432222 (in Russian).
- [8] Kuznetsov M.P., Ivkin N.P. Time series classification algorithm using combined feature description // *Mashinnoe obuchenie i analiz dannykh.* – 2015. – Vol. 1. – № 11. – Pp. 1471-1483. – eLIBRARY ID: 24931576 (in Russian).
- [9] Savkina A.V., Fedosin A.S. Data quality issues for energy management systems // *Prikladnyy zhurnal: upravlenie i vysokie tekhnologii.* – 2014. – № 2. – Pp. 158-164. – eLIBRARY ID: 21716788. – EDN: SHFYGT (in Russian).
- [10] Stukach O.V., Gabitova O.O. Outcomes of Energy-Saving Regulation of Thermal Energy on the Example of One Building // *Automatics & Software Engineering.* – 2023. – № 2(44). – Pp. 158-164. – eLIBRARY ID: 54207953 (in Russian).
- [11] Popov I.Yu., Stukach O.V., Zorin P.A. Evaluation of the Quality of Commercial Thermal Energy Accounting Data by Total Variation Outlier Recognizer // *Dynamics of Systems, Mechanisms and Machines.* – 2021. – Vol. 9. – № 3. – Pp. 117-121. – Doi: 10.25206/2310-9793-9-3-117-121 (in Russian).
- [12] Fedosin A.S., Fedosin S.A. Input data cleaning in automatic systems for commercial measurement of power consumption // *Infokommunikacionnye tekhnologii.* – 2016. – Vol. 14. – № 2. – Pp. 162-168 (in Russian).
- [13] Bacher M., Ben-Gal I., Shmueli E. Subspace selection for anomaly detection: An information theory approach. / *IEEE International Conference on the Science of Electrical Engineering (ICSEE).* – 16–18 Nov. 2016. – Eilat. – Israel, id.50. – P. 1-5. – Doi: 10.1109/icsee.2016.7806077.
- [14] Dutta J.K., Banerjee B., Reddy C.K. RODS: Rarity based Outlier Detection in a Sparse Coding Framework // *IEEE Transactions on Knowledge and Data Engineering.* – 2016. – Vol. 28. – № 2. – Pp. 483-495. – Doi: 10.1109/TKDE.2015.2475748.
- [15] Shebuti R. Sequential Ensemble Learning for Outlier Detection: A Bias-Variance Perspective // *2016 IEEE 16th International Conference on Data Mining (ICDM).* – 2016. – Pp. 1167–1172. – arXiv:1609.05528. – doi: 10.48550/arXiv.1609.05528.
- [16] Blythe D., Bunau P., Meinecke F.C., Muller K.R. Feature Extraction for Change-Point Detection using Stationary Subspace Analysis // *IEEE Transactions on Neural Networks and Learning Systems.* – 2012. – Vol. 23. – № 4. – Pp. 631-643. – arXiv:1108.2486v1. – Doi: 10.1109/TNNLS.2012.2185811.
- [17] Hido S., Ide T., Kashima H., Kubo H., Matsuzawa H. Unsupervised change analysis using supervised learning / *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD).* – 2008. – Part of the book series: Lecture Notes in Computer Science (LNAI, vol. 5012). – Pp. 148–159.
- [18] Gustafsson F. The marginalized likelihood ratio test for detecting abrupt changes // *IEEE Transactions on Automatic Control.* – 1996. Vol. 41. – № 1. – Pp. 66–78. – Doi: 10.1109/9.481608.
- [19] Popov I.Yu., Stukach O.V., Zorin P.A. The Use of Total Variation Outlier Recognizer Method to the Study of Commercial Thermal Energy Accounting Dataset // *Automatics & Software Engineering.* – 2021. – № 4(38). – Pp. 55-61 (in Russian).
- [20] Lagutin M.V. Visual mathematical statistics. – Moscow: VINOM. Laboratory of Knowledge. – 2009. – 472 p. (in Russian).
- [21] Ershov I.A., Stukach O.V., Aimagambetova R.Z. The Peculiar Measure Identifying of the Temperature Leap in the Distributed Raman Sensors // *Dynamics of Systems, Mechanisms and Machines (Dynamics).* – 2021, 9-11 Nov., Omsk, Russian Federation. – Pp. 1-4. – Doi: 10.1109/Dynamics52735.2021.9653721.
- [22] Ershov I.A., Stukach O.V., Tsydenzhapov I.B., Sychev I.V. Identification of a weak temperature transition in optical fiber using the Tsallis divergence / *Electronic tools and control systems: Proceedings of the XVI International Scientific and Practical Conference (November 18-20, 2020).* – Part 1. – Tomsk. – V-Spekt, – 2020. – Pp. 146-148 (in Russian).
- [23] Nguyen X.L., Wainwright M.J., Jordan M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization // *IEEE Transactions on Information Theory.* – 2010. – Vol. 56. – № 11. – Pp: 5847–5861. – arXiv:0809.0853v2. – Doi: 10.1109/TIT.2010.2068870.
- [24] Kawahara Y., Sugiyama M. Change-point detection in time-series data by direct density-ratio estimation / *SIAM International Conference on Data Mining (SDM 2009).* – April 30 - May 2. – 2009. – Sparks, Nevada, USA. – Pp. 389-400. – Doi:10.1137/1.9781611972795.34.
- [25] Liu S., Yamada M., Collier N., Sugiyama M. Change-point detection in time-series data by relative density ratio estimation // *Neural Networks.* – 2013. – 16 Jan. – arXiv:1203.0453v2. – Doi: 10.48550/arXiv.1203.0453.
- [26] Gretton A., Bousquet O., Smola A., Scholkopf B. Measuring statistical dependence with Hilbert-Schmidt norms // *ALT papes.* – 2005. – Pp. 63-77.
- [27] Steinwart I. On the influence of the kernel on the consistency of support vector machines // *Journal of Machine Learning Research.* – 2001. – № 2. – Pp. 67-93.
- [28] Zorin P., Stukach O. Data of heating meters from residential buildings in Tomsk (Russia) for statistical modeling of the thermal characteristics of buildings / *IEEE Dataport.* – 2020. – [Online]. Available: <http://dx.doi.org/10.21227/3r4e-ch18>.
- [29] Zorin P.A., Stukach O.V. Analiz vliyaniya pogodnykh uslovii na dinamiku teplovoi energii v zhilom fonde goroda Tomsk / *Innovative, information and comms tech: Proceedings of the XVII Int. sci.-pract. conf. – Moscow.* – 2020. – C. 365-368. – eLIBRARY ID: 45557868 (in Russian).
- [30] The CIC-IDS2017 dataset. Access mode: <https://www.unb.ca/cic/datasets/ids-2017.html>.
- [31] Stukach O., Zorin P. Long-Term Data from the Heat Meters in Residential Buildings Depending on the Outside Temperature and Characteristics of Buildings / *IEEE Dataport.* – April 13, 2021. – Doi: 10.21227/cw53-rr81.
- [32] The UNB ISCX dataset. Access mode: <https://www.unb.ca/cic/datasets/ids.html>.
- [33] The UNS-NB15 dataset. Access mode: <https://research.unsw.edu.au/projects/unsw-nb15-dataset>.
- [34] The CTU-13 dataset. Access mode: <https://www.kaggle.com/datasets/doogla/ctu13>.
- [35] Zorin P.A., Stukach O.V. Database of heat energy consumption by multi-storey residential buildings depending on meteorological factors and building characteristics. – Reg. № 25452. – Doi: 10.12731/ofernio.2025.25452 (in Russian).

Information about the authors

Oleg V. Stukach is with National Research University Higher School of Economics, Moscow, Russia; Novosibirsk State Technical University, Novosibirsk, Russia, e-mail: tomsk@ieee.org, ORCID: 0000-0001-6845-4285.

Olesya O. Eremina is with National Research Tomsk State University, Tomsk, Russia.