DOI: 10.24412/2077-8481-2026-1-87-96

УДК 681.2

М. А. ГУНДИНА, *канд. физ.-мат. наук*, *доц.*
П. С. БОГДАН, *канд. техн. наук*, *доц.*
О. В. ЮХНОВСКАЯ
Белорусский национальный технический университет (Минск, Беларусь)

ВЫЯВЛЕНИЕ НЕЭКСТРЕМАЛЬНЫХ АНОМАЛЬНЫХ ЗНАЧЕНИЙ ИЗМЕРЕНИЙ

Аннотация

Цель исследования – представление результатов работы алгоритмов, позволяющих проанализировать визуальную интерпретацию данных, полученных измерительными приборами, для выявления неэкстремальных аномальных значений. Для достижения вышеуказанной цели используется компьютерная система Wolfram Mathematica. В исследовании приведен результат обработки данных, полученных цифровым оптическим датчиком – пульсоксиметром рефлектометрического типа.

Ключевые слова:

аномальные значения, выборка, погрешность измерения, отклонение, прибор.

Для цитирования:

Гундина, М. А. Выявление неэкстремальных аномальных значений измерений / М. А. Гундина, П. С. Богдан, О. В. Юхновская // Вестник Белорусско-Российского университета. – 2026. – № 1 (90). – С. 87–96.

Введение

Поиск аномальных значений (выбросов или аномалий) в данных, полученных измерительными приборами, – весьма актуальная задача, поскольку такие значения искажают статистические показатели (среднее значение, стандартное отклонение и др.). Их выявление, удаление или корректировка позволяет получить репрезентативную выборку исследуемого процесса. Например, если датчик температуры печи в 99 % случаев показывает 500 °C, однако, согласно значениям выборки, наблюдается значение 0 °C (соответствующее сбою оборудования), средняя температура будет рассчитана неверно. Приборы могут давать сбои (кратковременные помехи, «обрыв», разряд батареи и др.). Выявление таких значений выборки, соответствующих описанным ситуациям, позволяет среагировать на возникшую ситуацию: либо исключить их из анализа, либо откалибровать прибор.

Кроме этого, например, скачки показаний датчиков являются распространенной проблемой в промышленных системах автоматизации. Нестабильные значения могут привести к неправильным решениям системы управления, аварийным остановкам оборудования. Понимание причин возникновения таких значений важно для обеспечения надежной работы технологических процессов [1, 2].

Анализ аномальных значений также способствует оптимизации процессов. Так, выявление такого значения позволяет найти уязвимые элементы исследуемой системы.

В некоторых описанных ситуациях при обнаружении аномального значения необходимы его удаление и корректировка выборки. В одних случаях нужен учет наличия таких значений на этапе моделирования, в других – аномалии могут сигнализировать о наличии

областей интереса, которые требуют дальнейшего детального исследования. Таким образом, решение вышеописанной задачи позволяет принимать более обоснованные решения на основе полученных достоверных данных и предотвратить поломки оборудования и сбои системы [3, 4].

Инженеры сталкиваются с ситуациями наличия аномальных значений измерений, которые требуют выявления или отсутствующих данных, требующих устранения или исправления, или значений, которые сильно отличаются от оставшихся значений выборки. Неспособность обнаружения аномальных значений может поставить под угрозу выводы проведенных исследований [5].

Систематические аномальные значения измерений приборов сложны для обнаружений и последующей обработки данных. Если вовремя не проанализировать выборку и не удалить эти аномальные значения, то они повлияют на изменчивость данных. До сих пор остается актуальной задача анализа результатов эксперимента с помощью разных инструментов, одного и того же инструмента в разное время либо некоторой комбинации этих ситуаций [6].

Цель исследования

Цель исследования – представление результатов работы алгоритмов, позволяющих проанализировать визуальную интерпретацию данных, полученных измерительными приборами, для выявления неэкстремальных аномальных значений.

Для достижения вышеуказанной цели используется компьютерная система Wolfram Mathematica. В этой системе представлен широкий набор возможностей для анализа аномальных значений, применяя оптимизированные алгоритмы. Рассмотренные функции позволяют пользователю решать реальные задачи и создавать приложения обработки данных различной природы, в том числе выявлять аномальные значения выборок.

Экстремальные и неэкстремальные аномальные значения выборки

Под аномальными значениями будем понимать данные, которые сильно отличаются от общего распределения. При этом можно выделить ошибочно возникающие выбросы (ошибка ввода данных, которые вызваны человеческим фактором; погрешности измерения; ошибка эксперимента, например шум при записи голоса; ошибка обработки; ошибка получения выборки), естественные выбросы и др.

Аномальные значения могут быть экстремальными и неэкстремальными. Неэкстремальные значения умеренно далеки от остальных данных, а экстремальные имеют большее расстояние от общего массива данных [7]. Таким образом, неэкстремальные аномальные значения соответствуют событиям с вероятностью порядка $\sim 10^{-3}$, а экстремальные – порядка $\sim 10^{-5}$ и меньше. Эти цифры являются следствием прикладной интерпретации математических расчетов для нормального распределения данных, представленных в [8].

Экстремальное аномальное значение – точка данных, которая катастрофически далека от основного распределения выборки. Это не просто редкое или выделяющееся значение, а такое, которое находится за пределами даже «разумных» ожиданий для изучаемого процесса или явления. Его присутствие в выборке часто выступает следствием грубых ошибок, сбоев системы или уникальных, нетипичных событий.

Рассмотрим основные характеристики экстремального аномального значения.

1. Высокая степень удаленности: значение находится настолько далеко от

центра распределения, что его нельзя объяснить естественной изменчивостью данных.

2. Сильное влияние: оказывает чрезвычайно большое воздействие на ключевые статистики выборки.

В этом случае регрессионные модели могут быть построены неверно, т. к. модель учитывает влияние значения, значительно отличающегося от остальных значений.

Неэкстремальное аномальное значение – это наблюдение в выборке, которое заметно отклоняется от основной массы данных, но не является катастрофически далеким. Оно находится за пределами «типичного» диапазона, но всё ещё может быть объяснено естественной изменчивостью процесса, редкими, но возможными событиями или незначительными ошибками измерения.

Ключевые характеристики:

1) умеренная удалённость: значение выходит за рамки ожидаемой вариабельности, но не настолько, чтобы считаться невозможным для данной генеральной совокупности;

2) ограниченное влияние: оказывает небольшое или среднее влияние на статистические показатели. Так, например, среднее арифметическое может слегка смещаться, а стандартное отклонение немного увеличиваться. При рассмотрении статистической модели влияние присутствует, но не искажает результаты кардинально.

Такие аномальные значения часто представляют наибольший аналитический интерес, т. к. могут указывать на:

– начало нового тренда;

– редкие, но важные состояния системы (например, пиковую нагрузку);

– объекты, относящиеся к другому, смежному классу данных.

Графическое представление измерений, полученных датчиком температуры

На рис. 1 представлены результаты измерения, полученные цифровым оптическим датчиком – пульсоксиметром MAX30100 рефлектометрического типа. Для связи с датчиком применялся микроконтроллер ESP32S3. Частота выборки составляла 200 Гц. Для обработки использовались данные инфракрасного канала, представляющие собой непосредственно выходные значения АЦП без дополнительного пересчета. На горизонтальной оси отмечены номера измерений параметра, на вертикальной – фактические значения датчика (см. рис. 1).

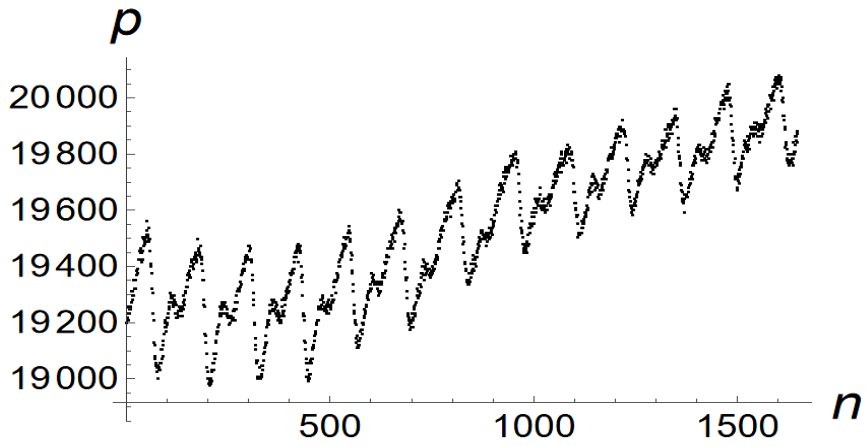

Рис. 1. Исходная выборка данных, полученных измерительным прибором

Для количественного описания и идентификации аномалии могут быть использованы квартили. Они в контексте работы с аномальными значениями применяются для количественного определения выбросов. Самая распространенная методика, использующая квартили для поиска выбросов, – это метод межквартильного размаха. Сравнивая статистики, которые чувствительны к выбросам (выборочное среднее), и статистики, которые устойчивы к ним (медиана), можно понять, насколько аномалии искажают картину.

Вычисление квартилей в Wolfram Mathematica осуществляем следующим образом:

Q1=Quantile[data,0.25]

Q3=Quantile[data,0.75]

IQR=Q3-Q1

Определяем границы для выбросов:

lowerBound=Q1-1.5*IQR

upperBound=Q3+1.5*IQR

Находим аномальные значения:

anomalies=Select[data,#<lowerBound||#>upperBound&]

Для данного набора значений наблюдается незначительное отличие выборочного среднего от медианы.

Также мощным инструментом для первичного анализа данных является гистограмма. Она представляет собой столбчатый график, показывающий распределение частот значений в выборке. Аномалии, будучи редкими и нетипичными, нарушают общую картину этого распределения.

Для сравнения представим результаты для трех выборок, полученных в процессе измерения (рис. 2).

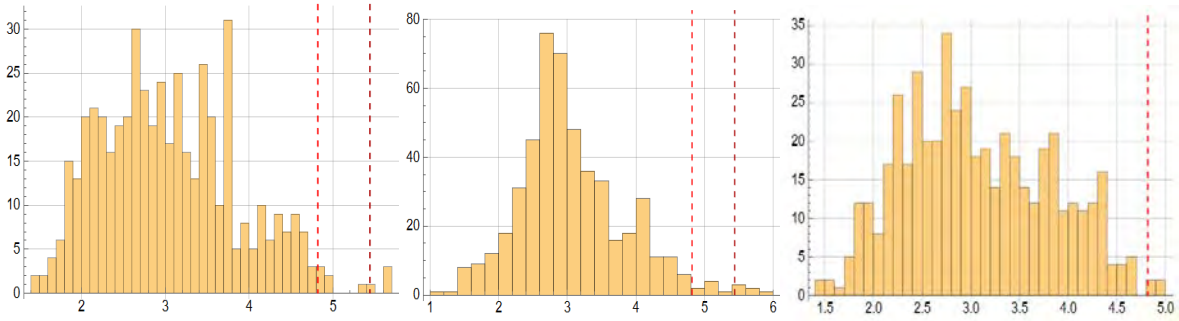

Рис. 2. Гистограмма распределения расстояний Махаланобиса для первой, второй и третьей выборок. На оси абсцисс представлены значения расстояния Махаланобиса, на оси ординат – частота

Вышеприведенная гистограмма указывает на присутствие аномальных значений. Основная масса данных образует один основной кластер и на отдалении от него с левого края появляется один или несколько очень низких и узких столбцов, которые не соединяются (или незначительно соединяются) с основным массивом. Такие «островки» представляют собой значения, которые находятся так далеко от основной части данных, что попадают в свои собственные, крайние интервалы. В первых двух выборках присутствуют экстремальные аномальные значения, которые располагаются между штриховыми линиями. В третьей выборке наблюдаются только неэкстремальные значения, т. к. граница

для экстремальных значений находится уже за границами гистограммы.

Однако использование гистограммы для выявления аномальных значений имеет недостатки. Если аномальных значений незначительное количество и они не слишком экстремальны по своим значениям, их столбец на гистограмме может быть настолько мал, что его легко не заметить. Такая задача – исследование неэкстремальных значений – и определила цель поводимых исследований.

Особенности реализации алгоритма выявления неэкстремальных аномальных значений

Исходные данные представляют собой одномерный массив $x_1, x_2, \ldots, x_n$.

Мультинормальный закон распределения описывает следующую случайную величину:

$$X = \begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix} \sim N(\mu, \Sigma),$$

где μ – вектор математических ожиданий (или математическое ожидание); Σ – ковариационная матрица для многомерного случая (или среднее квадратическое отклонение).

Для преобразования данных применяем метод временных окон.

Создаем матрицу данных

$$x_1 = \begin{pmatrix} x_1 \\ \ldots \\ x_5 \end{pmatrix} \quad x_2 = \begin{pmatrix} x_2 \\ \ldots \\ x_6 \end{pmatrix} \quad \ldots \quad x_{n-m+1} = \begin{pmatrix} x_{n-4} \\ \ldots \\ x_n \end{pmatrix}$$

при значении параметра $m = 5$.

Для оценки параметров распределения используем выборочное среднее, выборочную ковариационную матрицу.

Для каждого вектора x_i вычисляем расстояние Махаланобиса как меру удалённости от центра распределения с учетом корреляционной структуры:

$$MD_i^2 = (x_i - \hat{\mu})^T \cdot \hat{\Sigma}^{-1} \cdot (x_i - \hat{\mu}),$$

где $\hat{\mu}$ – вектор математических ожиданий; Σ – ковариационная матрица.

Для формулировки критерия выявления аномальных значений будем использовать следующую величину:

$$T_m = \sqrt{F_{x_m^2}(1 - \alpha_m)},$$

где α_m – уровень значимости, для случая неэкстремальных значений $\alpha_m = 0{,}01$; $F_{x_m^2}$ – q-квантиль распределения $\chi^2(p)$.

Граница для экстремальных аномальных значений $T_e = \sqrt{F_{x_m^2}(1 - \alpha_e)}$, где $\alpha_e = 0{,}001$.

Тогда критерий будет выглядеть следующим образом.

Для каждого *i* значение:

– нормальное, если $MD_i \le T_m$;

– неэкстремальное, если $T_m \le MD_i \le T_e$;

– экстремальное, если $MD_i > T_e$,

где MD_i – расстояние Махаланобиса.

Общая схема применения описанного теоретического подхода следующая.

1. Задаем параметры алгоритма (размер скользящего окна, уровень значимости для неэкстремальных аномальных значений, уровень значимости для экстремальных значений).

2. Задаем функцию создания скользящих окон.

3. Оцениваем параметры многомерного нормального распределения.

4. Вычисляем расстояния Махаланобиса для каждого окна.

5. Определяем порог по распределению χ^2.

6. Осуществляем классификацию окон.

7. Выводим результаты.
8. Визуализируем результаты.

Результат работы алгоритма приведен на рис. 3.

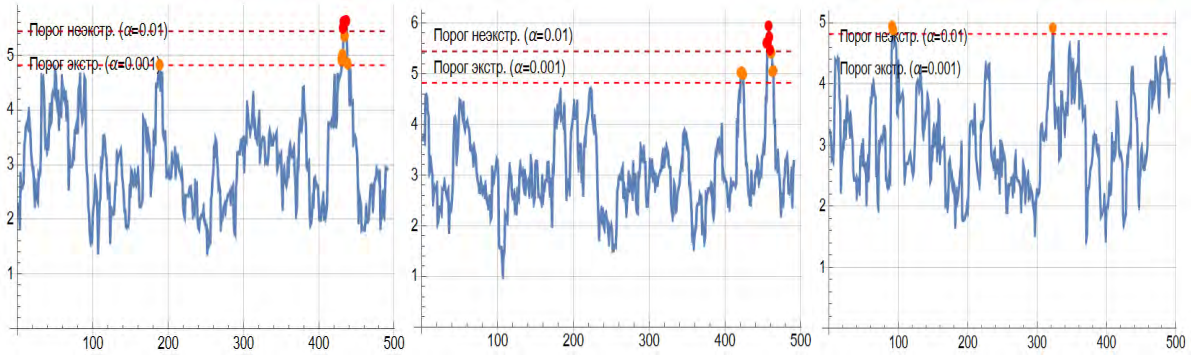

Рис. 3. График расстояний Махаланобиса для скользящих окон. На оси абсцисс представлены номера окон, на оси ординат – значения расстояния Махаланобиса

Отметим, что для первой, второй и третьей выборок выявлены неэкстремальные аномальные значения. В первом и во втором случае также выявлены и экстремальные аномальные значения, что подтверждается гистограммами, представленными на рис. 2.

Сравнение полученных результатов с результатами с использованием автокорреляционной функции для выявления аномальных значений

Существует подход к поиску аномалий с совершенно другой позиции рассуждений. Использование автокорреляционной функции позволяет обнаружить аномалии во временной структуре данных. Рассмотрим функции для данных трех выборок (рис. 4).

Команда, позволяющая построить график функции автокорреляции для исходных данных,

```
acf=ListPlot[CorrelationFunction[data,{1,1
00}],PlotRange->All,PlotLabel-
>"Автокорреляционная функция
(ACF)",GridLines->Automatic,Joined-
>True]
```

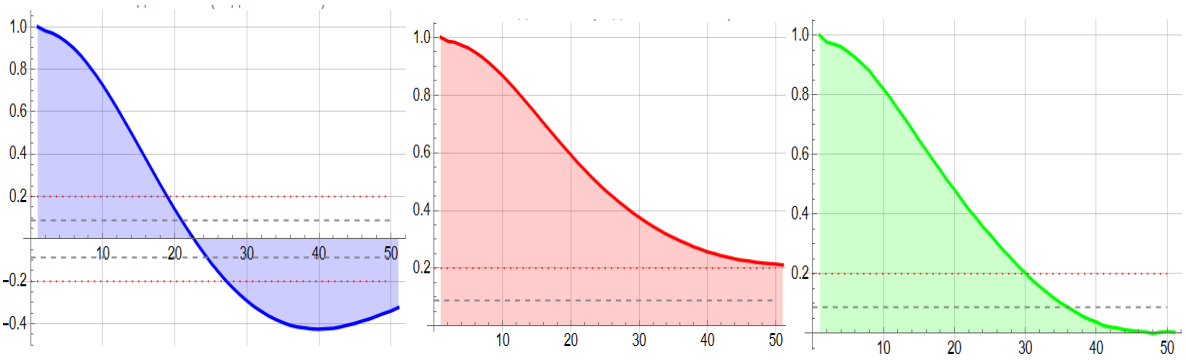

Рис. 4. График автокорреляционной функции для случая доверительной границы 95 %. На оси абсцисс располагается номер лага, на оси ординат – значение автокорреляционной функции. Критические значения для значимости представлены штриховой линией

Такая функция измеряет, насколько значения временного ряда коррелируют с их собственными предыдущими измерениями. Аномалии нарушают эту естественную для ряда корреляционную структуру.

Для стационарного временного ряда с нулевым средним для случайного процесса теоретическая автокорреляционная функция выглядит следующим образом:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{E((X_i - \mu)(X_{i+k} - \mu))}{E((X_i - \mu)^2)},$$

где $\gamma(k)$ – автоковариация с лагом k; μ – математическое ожидание процесса; E – оператор математического ожидания.

Для выборочной автокорреляционной функции формула для вычисления

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2},$$

где n – длина ряда; $\bar{x}$ – выборочное среднее.

Известно, что периодические компоненты ряда проявляются в графике такой функции в виде пиков через регулярные промежутки лагов.

Проанализировав данные функции корреляции, можно сделать следующие выводы. Данный сигнал не является белым шумом, т. к. в этом случае наблюдалось бы быстрое затухание функций. Медленное затухание указывает на сильную память процесса, наблюдается убывание по степенному закону. Отсутствуют периодические компоненты, т. к. в этом случае наблюдались бы затухания с колебаниями. Проанализировав графики подвыборок первой, второй и третьей, также можно сделать выводы, что первый график показывает быстрое затухание, а второй медленное и аномальные значения могут быть в переходной зоне от первой выборки ко второй. Действительно, на рис. 3 видно, что аномальные значения возникают к концу первой подвыборки.

Использование периодограммы для выявления аномальных значений

Для выявления аномальных значений также используется периодограмма (рис. 5). Она показывает, какая доля дисперсии (мощности) временного ряда сосредоточена на различных частотах. Всплески (см. рис. 5) указывают на наличие некоторых неэкстремальных аномалий.

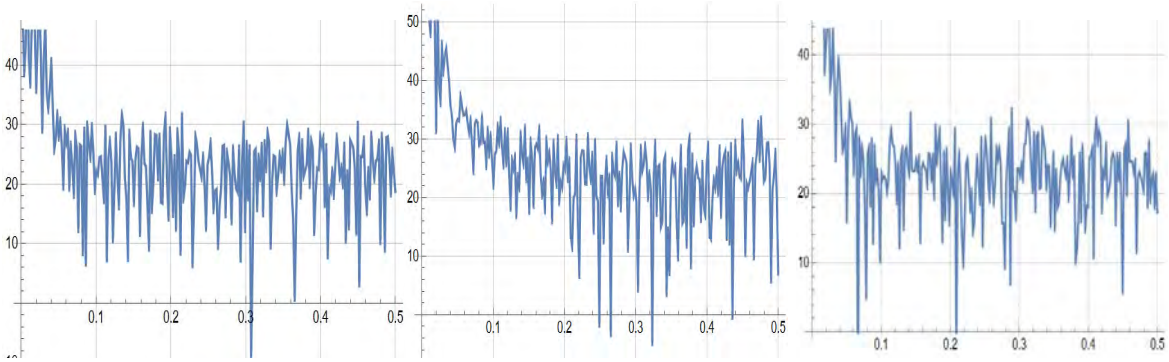

Рис. 5. Периодограмма для исходных массивов

Даже если в исходном временном ряде тренд и шум маскируют периодический компонент, периодограмма преобразует ряд в частотную область, где этот компонент становится очевидным в виде резкого пика на конкретной частоте.

Периодограмма показывает, какая доля дисперсии (мощности) сигнала приходится на различные частоты. Аномалия часто проявляется как неожиданно высокий пик на определенной частоте (или частотах), который не соответствует основному гармоническому составу нормального процесса.

Так, например, рассмотрим данные о потреблении электроэнергии. В обычном режиме периодограмма показывает пики на частотах, соответствующих суточным и недельным циклам. Если вдруг появляется новый, неожиданный пик на частоте, скажем, 2 ч (частота 1/120 Гц), то это аномалия. Она может указывать на неисправное оборудование, которое включается с таким странным интервалом, и это не видно невооруженным глазом во временной области.

Рассмотрим ряд изменений в периодограмме, которые могут указывать на наличие аномальных значений.

1. Исчезновение ожидаемого пика. Пропал суточный цикл в данных активности пользователей. Система, возможно, перестала функционировать в штатном режиме.

2. Ослабление/усиление пика. Пик суточной периодичности стал значительно слабее или, наоборот, резко усилился. Это указывает на изменение амплитуды цикла.

3. Сдвиг частоты пика. Пик периодичности сместился (например, с 24 на 23,5 ч). Это говорит об изменении длительности цикла.

Периодограмма позволяет выявить короткие, импульсные события. Внезапный резкий скачок значения (импульс) во временной области «размазывается» по всем частотам в частотной области. На периодограмме это проявляется как всплеск мощности на высоких частотах. Периодограмма по всему ряду дает усредненную картину. Чтобы найти момент, когда началась аномалия, используется скользящее окно. Оно применяется для построения периодограмм для коротких последовательных отрезков ряда для оценки спектра со временем.

Такой анализ позволяет выявить следующие ситуации при работе с приборами:

– аномальный пик на частоте, не кратной частоте вращения ротора, – признак повреждения подшипника;

– появление нового пика в спектре потребления – признак подключения неисправного или нового оборудования с собственным циклом работы.

Это дает возможность находить скрытые периодичности и аномалии, совершенно невидимые другими методами.

Заключение

Выявление неэкстремальных аномальных значений – одна из важных и сложных задач анализа данных. Такой подход позволяет повысить качество и надежность моделей, поскольку неэкстремальные аномалии могут незаметно смещать оценки параметров. Знание об их наличии заставляет выбирать устойчивые методы обработки данных.

В этом случае модель, обученная на «очищенных» от неинформативных аномалий данных, лучше улавливает общие закономерности и делает более точные прогнозы на новых данных.

При обработке изображений, полученных промышленным оборудованием, такие значения могут позволить выявить области интереса. Учет неэкстремальных аномальных значений дает

возможность выявить скрытые риски и проблемы (температура станка, стабильно выходящая на 5 % за верхнюю границу, но не вызывающая аварийный стоп, – сигнал о начинающемся износе). В медицине некритично высокий, но аномальный для пациента показатель в анализах крови может быть ранним маркером заболевания. В логистике небольшое, но систематическое увеличение времени доставки в одном узле – индикатор назревающей логистической проблемы.

При обработке промышленных изображений аномалии могут указывать на то, что выборка неоднородна и состоит из нескольких скрытых подвыборок.

Поскольку для инженера эта задача предобработки данных и построения устойчивых моделей является важной и актуальной, то необходим учет природы возникновения таких неэкстремальных аномальных значений.

Таким образом, выявление неэкстремальных аномальных значений – это не стандартная очистка, а стратегический этап анализа, который лежит на стыке статистики и предметной экспертизы. Игнорирование данного этапа ведет либо к построению хрупких моделей, либо к упущению важнейших факторов, скрытых в таких значениях.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. **Синютин, С. А.** Цифровая обработка сигналов в интеллектуальных датчиках вибрации / С. А. Синютин // Известия Южного федерального университета. Серия : Технические науки. – 2023. – Т. 32, № 3. – С. 18–26.

2. **Макарова, И. Л.** Обнаружение и интерпретация ошибочных данных при статистическом анализе потребления энергоресурсов / И. Л. Макарова, А. М. Игнатенко, А. С. Копырин // Программные системы и вычислительные методы. – 2021. – № 3. – С. 40–51.

3. **Гундина, М. А.** Особенности процесса определения количества аномальных значений при обработке измерительной информации / М. А. Гундина, П. С. Богдан, О. В. Юхновская // Вестник Белорусско-Российского университета. – 2024. – № 2 (77). – С. 96–103.

4. **Гундина, М. А.** Выявление аномальных кластеров выборки в компьютерной системе Wolfram Mathematica / М. А. Гундина // Вестник Белорусско-Российского университета. – 2022. – № 4 (77). – С. 75–83.

5. **Miot, H. A.** Anomalous values and missing data in clinical and experimental studies / H. A. Miot. – Botucatu : Jornal Vascular Brasileiro, 2019. – Vol. 18. – 7 p.

6. **Miot, H. A.** Analise de concordancia em estudos clinicos e experimentais / H. A. Miot. – Botucatu : Jornal Vascular Brasileiro, 2016. – Vol. 15 (2). – P. 89–92.

7. SPSS Finding Outliers in a Dataset. – URL: www.statistichero.com/en/spss-finding-outliers-in-a-dataset (date of access: 08.12.2025).

8. **Hoaglin, D. C.** Performance of Some Resistant Rules for Outlier Labeling / D. C. Hoaglin, B. Iglewicz, J. Tukey // Journal of the American Statistical Association. – 1986. – Vol. 81 (396). – P. 991–999.

Статья сдана в редакцию 20 декабря 2025 года

Контакты:
hundzina@bntu.by (Гундина Мария Анатольевна);
pbogdan@bntu.by (Богдан Павел Сергеевич);
juhnovskaja@bntu.by (Юхновская Ольга Витальевна).

M. A. HUNDZINA, P. S. BOHDAN, V. V. YUKHNOUSKAYA

IDENTIFICATION OF NON-EXTREME ANOMALOUS VALUES

Abstract

The objective of the research is to present the performance results of algorithms that analyze the visual interpretation of data obtained by measuring instruments for identifying non-extreme anomalous values. The Wolfram Mathematica computational system was used to achieve this objective. The research presents the results of processing data obtained with a digital optical pulse oximeter sensor of the reflectometric type.

Keywords:

anomalous values, sampling, measurement error, deviation, measuring instrument.

For citation:

Hundzina, M. A. Identification of non-extreme anomalous values / M. A. Hundzina, P. S. Bohdan, V. V. Yukhnouskaya // Belarusian-Russian University Bulletin. – 2026. – № 1 (90). – P. 87–96.