ПРИБОРОСТРОЕНИЕ

DOI: 10.24412/2077-8481-2024-2-96-103

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)

ОСОБЕННОСТИ ПРОЦЕССА ОПРЕДЕЛЕНИЯ КОЛИЧЕСТВА АНОМАЛЬНЫХ ЗНАЧЕНИЙ ПРИ ОБРАБОТКЕ ИЗМЕРИТЕЛЬНОЙ ИНФОРМАЦИИ

Аннотация

В современных компьютерных пакетах инженерных расчетов имеется возможность реализовать алгоритмы обнаружения аномальных значений выборки. Целью данного исследования является применение алгоритма, определяющего количество аномальных значений в зависимости от объема выборки, а также выявление зависимости количества аномальных значений при использовании статистического подхода от числового множителя при стандартном среднеквадратическом отклонении. Полученные результаты позволяют проанализировать источники погрешностей в измерительном процессе и повысить достоверность получаемых приборами данных. Представлен результат удаления аномальных значений при анализе данных, полученных при измерении температуры. Построены зависимость количества аномальных значений от объема рассматриваемой выборки и график зависимости количества аномалий от параметра стандартного отклонения модели.

Ключевые слова:

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

Для цитирования:

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 (83). – С. 96–103.

Введение

Научные исследования, связанные с обработкой сигналов, полученных современными приборами, часто сопряжены с резким увеличением потока измерительной информации, требующей последующего анализа. Обработка таких баз данных без использования автоматизированных систем сбора и обработки результатов становится практически невозможной. При этом измерительная информация может содержать аномальные значения, связанные с погрешностями измерений, в том числе с так называемыми шумами. В отечественной и зарубежной литературе представлены различные способы их устранения [1, 2].

Анализ ряда литературных источников [3, 4] позволяет сделать вывод о том, что под аномальными значениями при измерениях можно понимать отклонение результатов измерения от ожидаемых значений, соответствующих паспортным данным измерительных приборов и измеряемым параметрам объекта. Исходя из физического смысла исследуемой величины могут рассматриваться как аномальные те значения,

которые не соответствуют монотонному характеру изменения величины при последовательных наблюдениях, а также значения, приращения которых превышают предельно возможную скорость изменения величины.

Известно, что погрешности данных, полученных приборами, могут быть классифицированы следующим образом [3]:

– постоянные погрешности, например, погрешность начального положения в рычажных механизмах, которая приводит к несимметричности градуировочной характеристики;

– прогрессивные погрешности, например, погрешности масштаба, погрешности, вызванные отклонением длин рычагов, отклонения значений электрических сопротивлений;

– нелинейные погрешности, характеризующиеся монотонным убыванием или возрастанием, но изменяющиеся нелинейно;

– периодические погрешности, например, кинематическая погрешность зубчатых колес;

– апериодические погрешности, например, суммарная конечная погрешность.

Часто наблюдаются сравнительно небольшие погрешности измерения прибора, которые возникают под действием многих факторов и обычно описываются нормальным законом распределения. Однако также могут иметь место и значительные погрешности, распределение которых обычно отличается от гауссовского закона распределения и, в общем случае, может быть неизвестно. Поэтому аномальные значения с учетом характера их проявления можно разделить на две группы:

1) систематические аномальные значения, которые остаются постоянными или закономерно меняющимися при повторных измерениях одной и той же физической величины;

2) грубые аномальные значения, которые возникают, например, при сбое в работе измерительных приборов и при других резких изменениях условий проведения измерений.

Шумы при измерениях по источнику возникновения можно разделить на три группы. Первая группа включает шумы, генерируемые объектом исследования, вторая – внешние воздействия, третья – шумы аппаратуры и носителей информации. Шумы могут содержать аномальные значения, принадлежащие обеим группам.

До сих пор остается актуальной задача анализа результатов эксперимента с помощью разных инструментов, с помощью одного и того же инструмента в разное время, либо с помощью некоторой комбинации этих ситуаций [5, 6]. Очевидна потребность анализа аномальных значений, в том числе их количества, чтобы иметь возможность обнаруживать источники возникновения погрешностей, в том числе шумов, и, соответственно, корректировать процесс измерений, полученных прибором, и алгоритм обработки измерительной информации. В инженерии своевременное обнаружение аномалий является основой предотвращения нештатных ситуаций и преждевременных отказов приборов.

Целью настоящего исследования является, во-первых, представление работы алгоритма, определяющего количество аномальных значений в зависимости от объема выборки, в том числе выявление систематических аномальных значений при использовании скользящего среднего, во-вторых, выявление зависимости количества аномальных значений при использовании статистического подхода от числового множителя при стандартном среднеквадратическом отклонении. Решение этих задач позволит повысить достоверность получаемых приборами данных и увеличит их информативность.

Исследование влияния числа измерений на количество аномальных значений при измерении температуры воздуха

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. Это ставит под угрозу выполнение гипотезы о нормальности распределения исходной выборки, происходит влияние на выводы для всей генеральной совокупности (о работоспособности или параметрах прибора). Часто результаты, полученные приборами, подчинены нормальному закону распределения. При анализе таких выборок с позиции статистических методов аномальными считаются те значения, которые выходят за пределы от первого до третьего квартилей. Однако такой подход не используется в случае выборки, которая не подчинена нормальному закону распределения.

Существуют разные подходы к тому, что делать с аномальными значениями, которые выявлены в выборке. Один из подходов основан на исключении значений из выборки. Второй подход основан на отсечении данных [6].

С целью исключения влияния систематических аномальных значений используется метод скользящего среднего. Метод заключается в замене исходных значений средними арифметическими нескольких ближайших к нему значений.

При простом сглаживании

$$\widetilde{X} = \frac{1}{n}\sum_{t=m}^{n+m} X_t , \qquad (1)$$

где t – текущий номер члена ряда; n – размер окна (период сглаживания); m – номер члена ряда, значение которого заменяется средним значением.

При взвешенном сглаживании исходные значения ряда заменяются на средние значения, вычисленные по окну, взятые с некоторыми весами, отражающими вклад члена ряда в представляемые им закономерности процесса:

$$X_k = \sum_{i=0}^{p} a_i t^i , \qquad (2)$$

где p – порядок аппроксимирующего полинома на интервале $(t - n, t + n)$; a – соответствующие числовые коэффициенты; t – параметр, по которому осуществляется разложение.

Заметим, что для сигнала, в котором присутствуют аномальные значения измерений, рекомендуется использовать сглаживание скользящей медианой. Скользящая медиана – это медиана определенного количества предыдущих периодов временного ряда.

Естественно, что такие подходы применяются часто для обработки сигналов, регистрируемых приборами в реальном времени, либо для обработки сигналов, у которых не прослеживается четкой структуры.

В настоящее время существует множество подходов для определения аномалий и подходов к автоматизации процесса их выявления. Одни требуют привлечения сложного математического аппарата (фильтрация аномальных значений при бинарном и многоальтернативном обнаружении), другие подходы достаточно непросты в реализации и должны учитывать особенности природы исходного анализируемого процесса (определение аномальных перепадов яркости пикселей изображения, при нахождении контуров на изображении). Для поиска аномальных значений авторами использовалась компьютерная система Wolfram Mathematica, содержащая встроенные команды, позволяющие найти аппроксимирующую функцию, которая достаточно хорошо описывает исходный набор данных. Опции систе-

мы позволяют пользователю создавать приложения для обработки данных различной физической природы, в том числе выявление аномальных значений выборок.

Наблюдаемый ряд результатов измерений можно рассматривать как выборку, представляющую собой сумму полезной и случайной составляющих. Случайная составляющая может включать в себя аномальные значения. Под ними в этом контексте подразумеваются грубые выбросы. Они могут возникать в каналах измерения, обработки и передачи данных. Заметим, что даже при небольшой частоте их появления, они могут вносить большие погрешности в результате восстановления сигнала или в итоговые оценки статистических характеристик генеральной совокупности.

Известно, что существует прямая зависимость объема выборки и количества аномальных значений [6].

Для апробации методики определения влияния количества измерений на число аномальных значений использовалось сглаживание скользящей медианой. Обработке были подвергнуты измерения температуры воздуха, производимые термопарой, подключенной к цифровому усилителю MAX31855. Данные с усилителя принимались микроконтроллером STM32F411. В результате эксперимента было получено 1000 значений температуры в течение трех минут.

До проведения исследования предполагалось, что обработанная выборка обладает меньшей дисперсией по сравнению с исходной выборкой.

На рис. 1 представлена зависимость значений скользящей медианы от номера обработанного измерения и зависимость значений температуры от номера обработанного измерения до обработки.

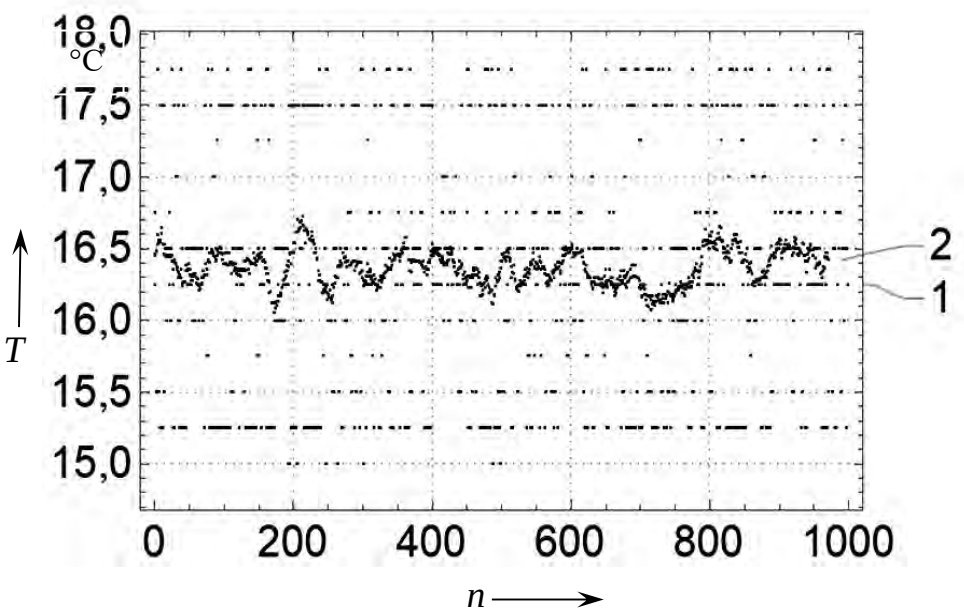

Рис. 1. Зависимость значений скользящей медианы от номера обработанного измерения (2), зависимость температуры от номера обработанного измерения до обработки (1)

Анализ полученных зависимостей позволяет сделать вывод, что процесс не является стационарным, на это может влиять шум аппаратуры либо внешних воздействий. Шумы, генерируемые объектом исследования, можно исключить,

т. к. в процессе эксперимента в течение измерений отсутствовали высокочастотные скачки температуры воздуха.

Кроме того, оказалось, что для данной выборки значение дисперсии снизилось на 73 % при использовании скользящего окна размером 10 измерений и на 78 % при использовании скользящего окна размером 15 измерений.

Если такие аномальные значения учитывать при последующем расчете, то при многократном и достаточно точном измерении они порождают рассеяние результатов.

Исследование влияния объема рассматриваемой выборки на количества аномальных значений

На рис. 2 представлен график зависимости количества аномальных значений от объема рассматриваемой исходной выборки. Этот график получен в результате работы алгоритма определения количества аномальных значений, реализованного в компьютерной системе *Wolfram Mathematica*. Он основан на использовании многомерного нормального распределения и принципов автоматического машинного обучения в системе.

Заметим, что при увеличении объема выборки количество выявленных аномальных значений снижается. Это связано с тем, что происходит процесс машинного обучения и те значения, которые при малом объеме выборки считались аномальными, при большем значении объема выборки считаются нормальными.

Для сглаженного ряда зависимость представлена на рис. 3 и имеет более монотонный характер изменения количества аномальных значений.

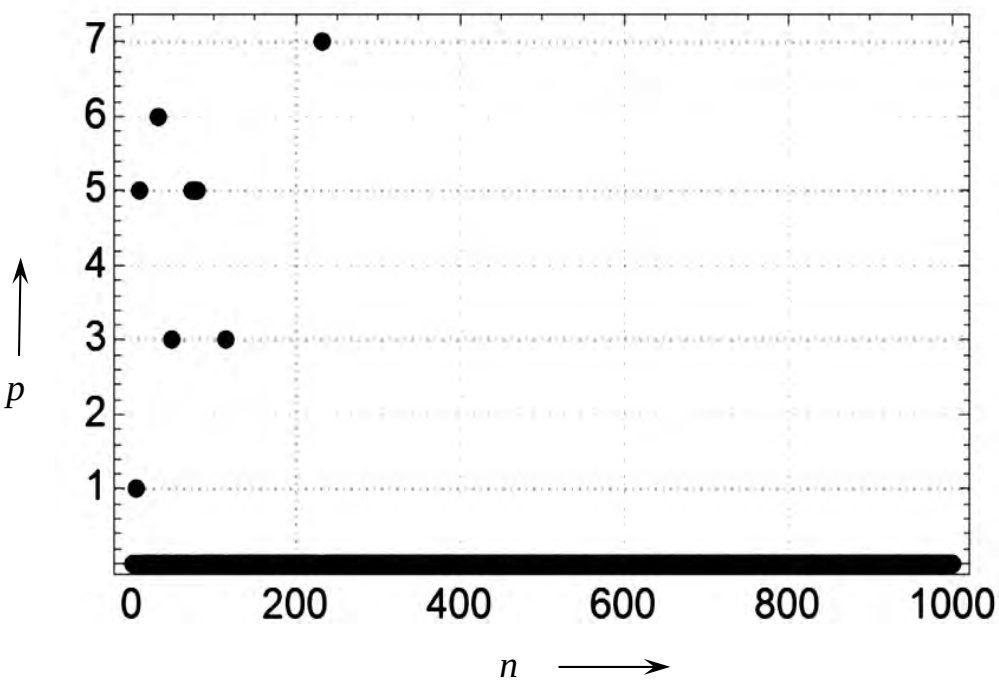

Рис. 2. Зависимость количества аномальных значений *p* от объема *n* рассматриваемой выборки

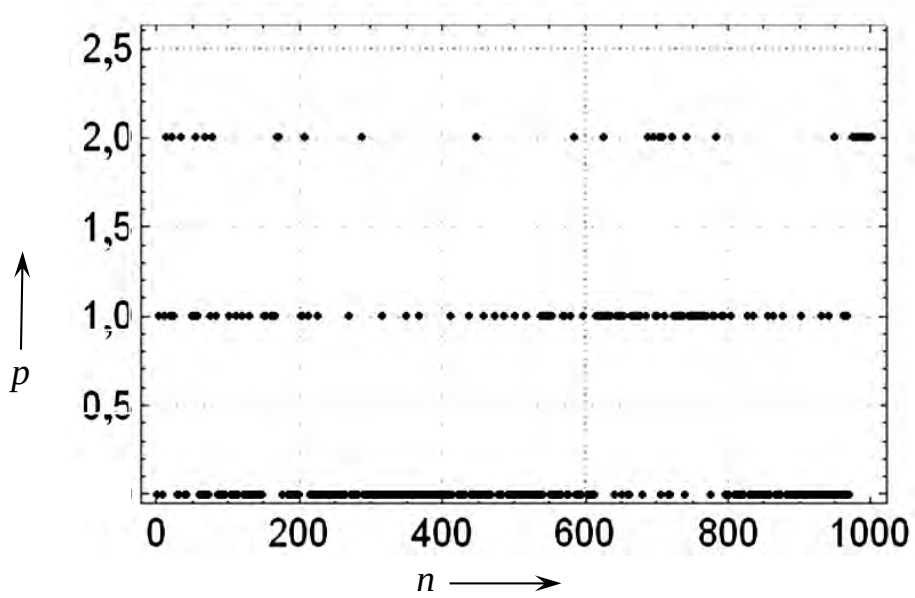

Рис. 3. Зависимость количества аномальных значений от объема рассматриваемой сглаженной выборки

Небольшие колебания в количестве определенных аномалий связаны с тем, что при выявлении аномальных значений было проведено сглаживание данных.

Исследование зависимости количества грубых и случайных аномальных значений от порога определения аномалий

Одной из групп методов выявления аномальных значений является группа статистических методов [7–9].

Рассмотрим некоторые особенности реализации метода поиска аномалий с помощью правила выбора коэффициента k, определяющего порог определения аномалий. С помощью этого метода можно осуществлять контроль нахождения параметра в допустимых границах, что удобно в производственных процессах.

Анализ выбросов в данных позволяет определить аномальные значения в нестационарных рядах с распределением, близким к нормальному закону распределения. Основу данного метода анализа составляет расчет среднего значения ряда и среднеквадратичного отклонения.

Формула для вычисления среднего значения ряда:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \qquad (3)$$

где n – количество элементов выборки; x_i – i-й элемент выборки.

Формула для вычисления среднеквадратичного отклонения:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} . \qquad (4)$$

Суть данного метода сводится к тому, что любые значения ряда, отличающиеся от среднего больше, чем на $k\sigma$, являются потенциальными аномалиями. Порог определения аномалий задаётся формулой

$$T = \bar{x} \pm k\sigma . \qquad (5)$$

Очевидно, что ширина порога зависит от выбранного коэффициента k при среднеквадратичном отклонении.

Рассмотрим график зависимости количества аномалий от значения параметра k (рис. 4).

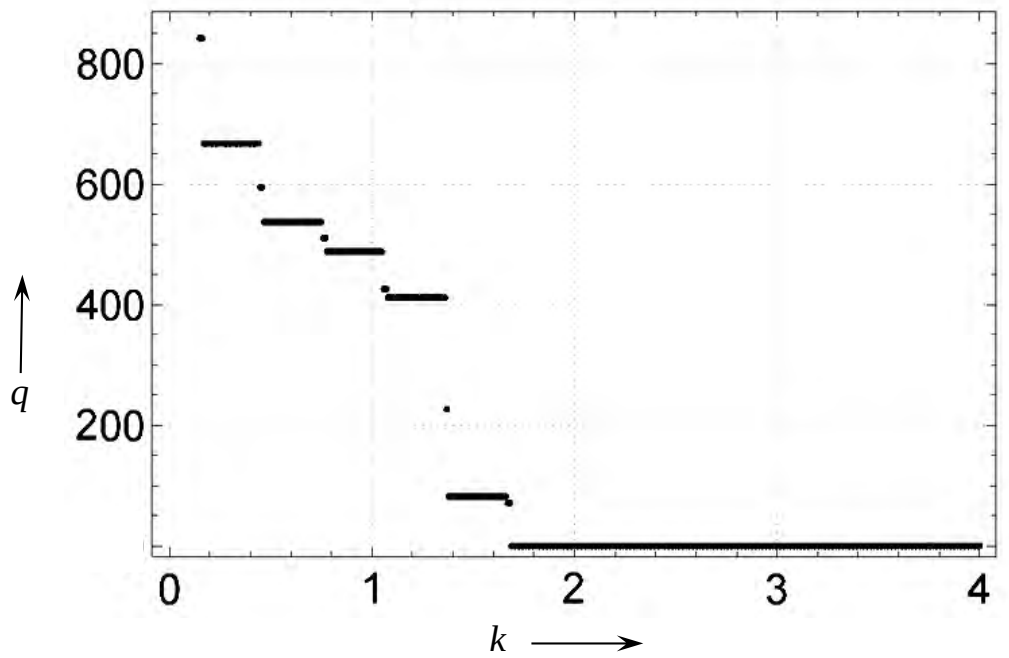

Рис. 4. График зависимости количества аномалий q от параметра k

Очевидно, что для проведенного эксперимента по измерению температуры при расширении порога T за счет уменьшения значений k более 2, параметр k теряет свою информативность, что соответствует теории [9]. Аналогичные результаты получены другими авторами при использовании автоматических алгоритмов [10].

Заключение

Применение алгоритма сглаживания скользящей медианой позволило обнаружить нестационарность процесса измерения температуры и снизить значение дисперсии на 73 % при использовании скользящего окна размером 10 измерений и на 78 % при использовании скользящего окна размером 15 измерений.

Количество аномальных значений зависит от числа измерений после сглаживания в меньшей степени, чем до сглаживания.

Описана процедура получения зависимости количества аномальных значений от объема выборки. При увеличении значения коэффициента k при среднеквадратичном отклонении более 2 для данной выборки аномальные значения не наблюдаются, что позволяет сделать следующий вывод: при выборе модели для данного процесса измерения необходимо останавливать свой выбор для k из диапазона от (0,2].

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. **Подстригаев, А. С.** Классификация и способы устранения аномальных ошибок измерения частотно-временных параметров сигналов в широкополосных приемниках / А. С. Подстригаев // Журн. Сиб. федер. ун-та. Сер. Техника и технологии. – 2022. – Т. 15, вып. 2. – С. 223–237.

2. **Dunning, T.** Practical Machine Learning: A New Look at Anomaly Detection / T. Dunning, E. Friedman. – Sebastopol (California): O`Reilly, 2014. – 66 p.

3. **Коротков, В. П.** Основы метрологии и теории точности измерительных устройств / В. П. Коротков, В. П. Тайц. – Москва: Изд-во стандартов, 1978. – 352 с.

4. **Шкодырев, В. П.** Обзор методов обнаружения аномалий в потоках данных / В. П. Шкодырев // Second Conference on Software Engineering and Information Management. – 2017. – С. 50.

5. **Marchuk, V. I.** Methods for detecting anomalous values based on the method of multiplying estimates of the investigated implementation of a non-stationary random process / V. I. Marchuk, S. V. Tokareva // Informatics, telecommunications and management. – 2009. – № 1. – P. 64–68.

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

7. **Линник, Ю. В.** Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений / Ю. В. Линник. – Москва: Физматгиз, 1962. – 352 с.

8. **Смирнов, Н. В.** Курс теории вероятностей и математической статистики для технических приложений / Н. В. Смирнов, И. В. Дунин-Барковский. – Москва: Наука, 1965. – 552 с.

9. **Akoglu, L.** Graph based anomaly detection and description: a survey / L. Akoglu, H. Tong, D. Koutra // Data mining and knowledge discovery. – 2015. – Vol. 29. – P. 626–688.

10. **Samuelsson, M.** Anomaly Detection In Time Series Data a practical implementation for pulp and paper industry / M. Samuelsson. – Gothenburg: Chalmers University of Technology, 2016. – 27 p.

Статья сдана в редакцию 17 января 2024 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

FEATURES OF THE PROCESS OF DETERMINING THE NUMBER OF ANOMALOUS VALUES WHEN PROCESSING MEASUREMENT INFORMATION

Abstract

In modern computer packages for engineering calculations, it is possible to implement algorithms for detecting anomalous sample values. The purpose of this research is to use an algorithm that determines the number of anomalous values depending on the sample size, as well as when using a statistical approach, to identify the dependence of the number of anomalous values on the numerical multiplier with the standard deviation. The results obtained make it possible to analyze the sources of errors in the measurement process and increase the reliability of the data obtained by the instruments. The result of removing anomalous values, when analyzing data obtained by temperature measurements, is presented. The dependence of the number of anomalous values on the volume of the sample under consideration and a graph of the dependence of the number of anomalies on the standard deviation parameter of the model were constructed.

Keywords:

anomalous values, sampling, measurement error, deviation, device.

For citation:

Hundzina, M. A. Features of the process of determining the number of anomalous values when processing measurement information / M. A. Hundzina, P. S. Bohdan, V. V. Yukhnouskaya // Belarusian-Russian University Bulletin. – 2024. – № 2 (83). – P. 96–103.