УДК 519.23:681.5 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait/1995-5499/2023/3/121-133
Поступила в редакцию 06.12.2022
Подписана в печать 30.09.2023

ВЫЯВЛЕНИЕ ВЫБРОСОВ В ДАННЫХ ПРАКТИЧЕСКИ ОДНОРОДНЫХ ТЕХНИЧЕСКИХ СИСТЕМ

© 2023 А. С. Дулесов, А. В. Байшев✉

Хакасский государственный университет им. Н. Ф. Катанова
пр-т Ленина, 90, 655017 Абакан, Республика Хакасия, Российская Федерация

Аннотация. В статье представлен метод выявления выбросов в данных о том или ином параметре практически однородных технических систем(далее ПОТС) имеющих форму временных рядов путем их сравнительного анализа, отражающего случаи невозможности его проведения. Метод весьма актуален в силу распространенности применения временных рядов в различных современных технических системах. Для разработки метода был проведен неисчерпывающий обзор различных методов обнаружения дефектов в данных. Основное внимание в обзоре было приковано к обнаружению выбросов, так как такой дефект как пропуски в данных присутствуют в них практически в явном виде. Рассмотрены некоторые достоинства и недостатки методов обнаружения выбросов, с учетом которых была произведена разработка метода. Метод обладает широкими перспективами его дальнейшего применения, так как может помочь в выявлении выбросов в данных исследуемых систем, а также в случае сходства данных может дать дополнительную уверенность исследователю о том, что системы, данные которых исследуются находились в обозначенное время в исправном состоянии в практически равных условиях. В свою очередь, различие может говорить о наличии в данных выбросов, причинами наличия которых могут быть: неисправности той или иной из систем или в системе сбора и хранения данных о них, влиянием на ту или иную систем неучтенного фактора. Несмотря на некоторую субъективность, метод имеет существенный плюс в виде гибкости, кроме того, он не требует построения сложных моделей определяющих эталонное поведение исследуемого параметра ПОТС для проведения сравнительного анализа данных модели и исследуемых систем, что указывает на перспективность его применения для анализа данных даже сложных ПОТС.

Ключевые слова: система, системный анализ, временные ряды, закономерности, априорная информация, практически одинаковые технические системы, меры сходства, выбросы, данные, дефекты.

ВВЕДЕНИЕ

На фоне существующего множества различных технических систем, среди них можно выделить класс практически однородных технических систем (далее ПОТС) широко распространенных в современном мире.

В работе понимается, что ПОТС — это такие системы, о которых априорно известно, что, при условии исправности, они практически равны между собой по структуре и составу. То есть число их элементов, взаимосвязей между ними, как и значения параметров их характеризующих практически равны между собой. «Практическое равенство» здесь и далее указывает на то, что реально присутствующее различие является несущественным при их рассмотрении в контексте решения той или иной задачи.

Очевидно, что ПОТС должны характеризоваться практически одинаковыми данными в условиях априорного практического равенства внешних воздействующих факторов (далее априорно равных условиях) при их сравнительном анализе. Соответственно, фикса-

✉ Байшев Анатолий Викторович
e-mail: anatoly_bayshev@mail.ru

ция различий для таких случаев будет указывать на возможное наличие выбросов в них.

В работе под выбросами понимаются наблюдения, существенно отличающиеся от других наблюдений имеющегося набора данных [1].

В свою очередь, выбросы могут быть вызваны разными причинами, например, неисправностями в той или иной из систем или сопутствующих им системах сбора и хранения данных, либо действием неучтенных при отборе для сравнительного анализа данных факторов на ту или иную из систем, человеческим фактором и т. п.

Однако сравнительный анализ данных о том или ином параметре для нескольких ПОТС в ручном режиме, особенно на длительных временных интервалах эксплуатации, представляет собой трудную задачу, так как при этом приходится иметь дело с несколькими протяженными временными рядами, анализ которых сопряжен с рядом сложностей. Например, общеизвестно, что современные технические системы в большинстве своем оснащаются системами автоматического сбора данных отражающих динамику изменения массы их параметров фиксирующихся через короткие промежутки времени, что приводит к очень большой размерности временных рядов, то есть, по сути, приходится иметь дело с Big Data.

Кроме того, при практической работе помимо выбросов, в данных можно встретить такие дефекты как их пропуски от которых, на практике, часто избавляются путем применения различных способов интерполяции [1–4].

Однако, следует отметить, что информация как о выбросах, так и о их пропусках в случае технических систем несет определенную ценность, так как может указывать на их те или иные неисправности. Соответственно выявление дефектов в данных в виде выбросов и отсутствующих значений, а также представление фактов их фиксации в удобной для анализа форме и их анализ являют собой актуальные задачи в вопросах исследования систем.

Выявление фактов отсутствия данных не являет собой сложную задачу, так как они, в большинстве своем, присутствуют в любых данных практически в явном виде.

В свою очередь, обнаружение выбросов является более трудной задачей, так как требует определения того, действительно ли наблюдения несовместимые с остальными данными являются ими [1]. Однако степень несовместимости данных с остальными данными для выбросов, равно как и то какие данные должны быть взяты в качестве эталонных, определяется в тех или иных случаях по-разному.

Соответственно проблема выявления выбросов весьма сложна и не вызывает удивления тот факт, что в теории анализа данных до сих пор нет общепринятого универсального метода их обнаружения и(или) устранения [1].

На этом фоне сравнительный анализ данных о том или ином параметре ПОТС находящихся в априорно равных условиях представляет собой перспективное направление исследований, так как несходство данных между системами может указывать на наличие выбросов.

Соответственно, актуальной становится разработка алгоритмов, методов, моделей позволяющих в автоматизированном режиме производить сравнение данных о параметре, в том числе имеющих форму временных рядов, нескольких ПОТС находящихся в априорно равных условиях.

Актуален также и учет при их разработке фактов отсутствия данных, что в совокупности с результатами сравнительного анализа данных, может давать больше информации относительно данных и систем, которых они описывают.

Цель работы: разработка метода выявления выбросов путем сравнительного анализа данных о том или ином параметре ПОТС имеющих форму временных рядов, учитывающего факты отсутствия данных.

Для достижения цели работы обозначен ряд способствующих ее достижению задач:

– провести обзор методов обнаружения дефектов в данных временных рядов,

– определить метод сравнительного анализа данных ПОТС и провести с его помощью исследование данных о параметре нескольких ПОТС,

– сформулировать выводы на основании проведенного исследования.

1. МАТЕРИАЛЫ И МЕТОДЫ

1.1. Обзор методов обнаружения дефектов в данных временных рядов

Приведенный в работе обзор не является исчерпывающим, так как различных подходов обнаружения дефектов в данных существует очень много.

В работе внимание сосредоточено на таких дефектах данных как: выбросы и отсутствующие значения.

Обнаружение отсутствующих данных не являет собой трудную задачу, так как зачастую информация об их наличии присутствует в данных практически в явной форме, хотя в некоторых случаях она проявляется при восстановлении нормального временного распределения ряда.

С выбросами ситуация обстоит сложнее, так как общепринятого подхода к их определению нет [1]. По форме присутствия в данных они могут представлять собой как одно значение, так и являться их последовательностью.

Выбросы условно можно разделить на явно выраженные и неявные. Явно выраженные выбросы представляют собой значения параметра или их последовательность, определить которые можно путем простого установления границ на основе априорной информации о его нормальном поведении нахождение вне рамок которых для значения параметра будет говорить о том, что оно представляет собой выброс.

Однако существуют и неявные выбросы — это такие значения в данных, при которых они не нарушают границ нормального поведения, однако либо описывают недействительное состояние систем, либо систему при условиях ее неисправности.

Обнаружение выбросов является важной задачей, например, на этапе отбора данных для машинного обучения перед построением прогностических моделей, которыми часто предсказывают тот или иной параметр исправной системы. Соответственно, использование данных систем с выбросами, может существенно ухудшить результаты таких моделей, ведь недаром среди специалистов в машинном обучении известен принцип, согласно которому использование «мусора на входе» приведет к появлению «мусора на выходе».

Обнаружению выбросов в данных посвящено много различных методов. К базовыми методам относятся основанные на определении границ допустимого диапазона значений параметра рассчитываемого по модели объекта, выход за которые будет указывать на наличие выброса [4–5]. Однако построение моделей сложных объектов требует учета большого числа различных деталей, что делает их дорогостоящими и сложными в построении [5–6].

Для выявления последовательностей выбросов, а также для восстановления утраченных последовательностей в работе [4] предлагается использовать рекуррентные сети долгой краткосрочной памяти.

Множество исследований сосредоточено на подходах к выявлению выбросов с использованием кластеризации [7–8]. Кластеризация подразумевает разделение данных на группы (кластеры), внутри которых выделенные объекты имеют большее сходство между собой, чем с любыми прочими из другого кластера [9–10]. Так, в работе [7] указывается то, что кластеры с малым числом объектов потенциально содержат в себе выбросы. Также говорится о том, что такой метод хоть и способен выделить временной ряд с выбросами целиком, однако не позволяет точно определить его месторасположение в ряде [7].

Однако, многие исследования показывают, что применение кластерного анализа в отношении множества временных рядов является нетривиальной задачей обусловленной, в том числе, их большой размерностью [9, 11]. Кроме того, существует ряд проблем, которые в зависимости от выбранного алгоритма кластеризации, объема данных о тех или иных параметрах и их качества, могут проявляться в разной степени в процессе исследования: трудная интерпретация результатов, вычислительная сложность, необходимость задания количества кластеров, чувствительность к выбросам [10–11].

Следует отметить также существующую массу универсальных статистических методов обнаружения выбросов находящие свое применение в различных исследованиях: по правилу Томпсона, правилу Тьюки, с помощью теста Томпсона-Тау, на основе вычисления критериев Пирса, Смирнова-Граббса, Шовене, Титьена-Мура, Диксона [7, 12–18].

Как правило, статистические методы обнаружения выбросов основаны на предполагаемых распределениях данных, от параметров которых зависит количество ожидаемых выбросов и место их возможного расположения. Например, для нормального распределения существует широко известное правило трех стандартных отклонений от среднего, согласно которому превышающие его значения данных считают выбросами [8].

Основная проблема классических статистических подходов заключается в том, что для их применения требуется принятие допущения о конкретном распределении исследуемых данных, истинное распределение которых неизвестно. Кроме того, среднее значение, стандартное отклонение или ковариация используемые в статистических методах для обнаружения выбросов сами по себе не устойчивы к их наличию. Присутствие выбросов изначально не известно, однако если они имеются, то могут оказать существенное влияние на результаты их выявления статистическими методами [8].

Для решения проблемы влияния выбросов на результаты их обнаружения предлагают использовать различные методы, например, использующие присвоение разных весов данным. В работе [19] описан метод позволяющий применять правило трех стандартных отклонений для временных рядов с наличием тренда путем вычисления разностей исходного ряда с последующим применением правила. В исследовании [20] рассмотрены различные варианты устранения выбросов путём сглаживания исходного временного ряда, где указано, что использование некоторых алгоритмов сглаживания устраняет выбросы исходного ряда, но при этом может приводить к появлению новых выбросов, а также отмечено, что сглаживание на основе медианы в сравнении с другими рассмотренными методами(сглаживание с использованием среднего, 4253H, 4253H-twice, 3RSSH, sm, sm-twice) обладает преимуществами в устранении выбросов.

В работе [8] представлен весьма обширный обзор по существующим методам обнаружения выбросов на основе рассмотрения работ по этой теме за последние почти два десятка лет нынешнего века. Рассматриваются как статистические методы, так и методы, основанные на машинном обучении. Отмечены вероятностные, основанные на расстоянии, на основе кластеризации, теоретико-информационные методы, а также указаны сферы их применения.

Кроме того, в [8] даны различные подходы к классификации методов обнаружения выбросов. По объему анализируемых данных методы можно разделить на: глобальные, в которых учитывается вся база данных и локальные, с помощью которых исследуется только часть данных. Также методы можно разделить на: бинарные, при которых данные маркируются метками выброс(не выброс), а также оценочные, при которых дается оценка того, насколько данные являются выбросами в виде, например, вероятности.

Помимо перечисленных, методы обнаружения выбросов разделяют на: контролируемые, полуконтролируемые и неконтролируемые, а также на параметрические и непараметрические. Контролируемые и полуконтролируемые методы имеют недостаток в виде необходимости изначального наличия данных без выбросов для обучения модели [8].

Для выявления выбросов на примере анализа данных систем центрального теплоснабжения используют также формирование групп схожих систем, данные о параметре которых сравнивают между собой. Выбросами в такой ситуации считают данные той или иной системы, величина расстояния которых от группы схожих систем становится выше величины заданного порогового значения [4, 6].

В работе [6] обозначены методы обнаружения выбросов, предполагающие разделение массы исследуемых систем на схожие группы на основании сходства каких-либо их характеристик. Часто во многих методах

объединение в группы происходит на основе самих исследуемых данных. Однако при этом формирование групп часто являет собой нетривиальную задачу так как требует оценки адекватности выбора их объема, от которого напрямую зависит процесс выявления выбросов [6].

После формирования групп, за их данными наблюдают по изменению того или иного критерия и выявляют данные, которые с течением времени начинают существенно отличаться от групповых, то есть потенциально являются выбросами.

Однако такие методы сосредоточены на выявлении выбросов с целью определения неисправностей в технических системах (автопарк, системы отопления и т. д.) и часто более направлены на онлайн-мониторинг неисправностей, а не на анализ и выявление выбросов в уже накопленных данных, что, например, актуально для этапа предварительного анализа данных перед их использованием для построения каких-либо предиктивных моделей.

Также, они не подразумевают учета информации о возможных отсутствующих значениях данных, которые также несут в себе определенную полезную информацию.

Более того, небольшие пропуски в данных о системах, как правило, многими исследователями интерполируются какими-либо значениями, хотя в последствии такие данные могут не отражать действительности в должной мере [2, 6].

Кроме того, разделение систем на группы в методах, как правило происходит на основании самих исследуемых данных. В определенных случаях это приводит к тому, что некоторое количество систем не могут быть отнесены к той или иной группе, так как, например, они были неисправны изначально, либо являлись действительно сильно отличными по своим свойствам от других систем [6].

На взгляд авторов, анализ сходства данных ПОТС для выявления выбросов может быть более информативным при учете факторов времени, а также отсутствия данных, что может помочь в определении закономерностей относительно исследуемых систем. Кроме того, в условиях доступности информации об исследуемых системах, объединение их в группу схожих может осуществляться с привлечением эксперта, обладающим знаниями о них, а не только на основе их данных. Учет этих моментов даст возможность объединить в группу ПОТС системы, которые должны быть равны между собой в условиях равенства внешних воздействующих факторов, то есть в группу априорно практически равных систем. Это позволит обнаруживать выбросы в данных даже изначально отличных по той или иной причине (неисправность, действие неучтенного внешнего фактора) от остальных систем.

2. МЕТОД АНАЛИЗА ДАННЫХ ПОТС

Сравнительный анализ данных о том или ином параметре нескольких ПОТС при нахождении их в условиях практического равенства внешних воздействующих факторов является не простой задачей.

Сложности обусловливаются, в том числе тем, что данные о параметре нескольких ПОТС часто рассматривают на некотором временном интервале $T_{эксп}$, который может быть весьма протяженным и на котором системы могут подвергаться различным внешним воздействиям.

В свою очередь, сравнительный анализ данных о параметре тех или иных ПОТС стоит осуществлять при их нахождении в априорно равных условиях, оказывающих существенное влияние на исследуемый параметр. В такой ситуации, фиксация различий в данных систем будет указывать либо на наличие неисправностей в той или иной из них, либо на действие на них неучтенного фактора, либо на неисправности в аппаратуре, регистрирующей данные о параметре.

Соответственно требуется идентификация внешних воздействующих факторов, для которых в дальнейшем будет осуществляться извлечение и сравнение данных ПОТС, что в случае рассматриваемого протяженного непрерывного периода времени $T_{эксп}$ может представлять трудную задачу.

Большей эффективностью и информативностью в таком случае может обладать разделение протяженного промежутка времени $T_{эксп}$ на некоторое количество равных промежутков (сегментов) Δ_i: $T_{эксп} = \Delta_1, \Delta_2, \Delta_3, \ldots, \Delta_i, \ldots, \Delta_k$, $i = 1, 2, 3, \ldots, k$. Это позволит для каждого из сегментов Δ_i осуществить идентификацию априорно равных условий для исследуемого параметра ПОТС и выбрать данные о нем на каждом из сегментов Δ_i для каждой из систем. В случае отсутствия необходимых знаний у аналитика, резонным является привлечение эксперта, обладающего знаниями о исследуемых ПОТС для реализации этого этапа.

Важным условием для проведения сравнительного анализа данных о параметре ПОТС является необходимость в наличии на сегменте Δ_i данных о параметре как минимум двух из них при априорно равных условиях оказывающих существенное влияние на исследуемый параметр. Также к анализу должны извлекаться данные без отсутствующих значений. Однако в реальных приложениях встречаются и такие случаи, которые сами по себе несут определенную информацию, поэтому их следует выделять для последующего анализа. Представляют интерес случаи:

– когда на сегменте Δ_i невозможно извлечь данные о параметре, например из-за того, что системы работали в разном режиме. В этой ситуации данные невозможно сравнивать с данными других систем,

– когда данные о параметре всех систем на сегменте Δ_i отсутствуют полностью,

– когда извлекаемые данные о параметре как минимум одной из исследуемых систем на сегменте Δ_i содержат пропуски, отсутствуют полностью, либо не могут быть извлечены.

После реализации этапа отбора данных, становится возможным проведение сравнительного анализа данных ПОТС на сегментах Δ_i. Их сравнение можно осуществлять разным образом, например, визуально, что при большом количестве сегментов Δ_i трудоемко. Альтернативой может служить оценка сходства с помощью той или иной из мер (Евклидова, Манхэттэна и т. п.), с установлением допустимого максимального ее значения, превышение которого укажет на то, что данные систем существенно различаются. Преимуществом использования мер является то, что они, в отличие от массы статистических, более устойчивы к потенциальному наличию выбросов в данных [8]. Установить допустимое значение меры оценки сходства можно при помощи эксперта, обладающего знаниями о исследуемых ПОТС.

Отобразить информацию о проведенном анализе данных можно в виде диаграммы, примерный вид которой изображен на рис. 1, где не отмечаются только те сегменты Δ_i, на которых данные о параметре ПОТС обладали практическим сходством.

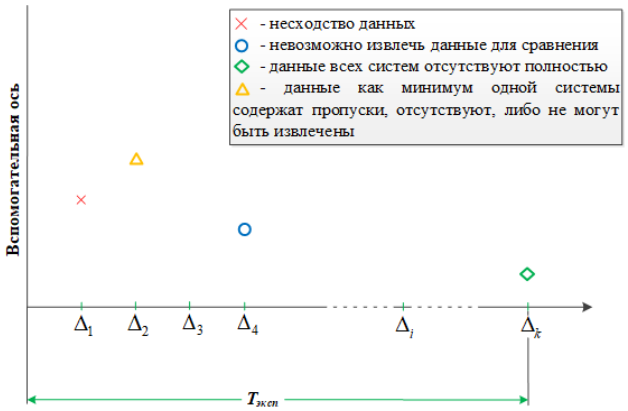

Рис. 1. Диаграмма с отмеченными сегментами Δ_i, на которых данные о параметре ПОТС имели различия или где сравнительный анализ невозможен
[Fig. 1. Chart with marked segments where data on the PHTS parameter had differences or where comparative analysis is not possible]

Диаграмма рис. 1. позволяет выявлять моменты времени, когда данные обладали несходством (потенциально содержали выбросы) при априорно равных условиях оказывающих существенное влияние на исследуемый параметр ПОТС, а также указывает на ситуации, с указанием причин, когда сравнительный анализ данных был невозможен, в том числе и по причине отсутствия данных.

3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

В работе демонстрация возможностей применения метода производится на данных

двух ПОТС. Они описывают изменение параметра среднечасовой температуры платы управления некоторым оборудованием в каждой из таких систем.

Об исследуемых системах известно, что при их исправности они всегда работают в априорно равных условиях, что говорит о том, что они должны характеризоваться практически одинаковыми данными при их сравнительном анализе в любое время.

В свою очередь, при реальном анализе в исследуемых данных о параметре может обнаруживаться несходство. Оно может являться следствием влияния на ту или иную из ПОТС какого-либо фактора, либо указывать на то, что в какой-то из систем или в регистрирующей данные аппаратуре есть неисправность.

Каждая из исследуемых ПОТС рассматривается на промежутке $T_{эксп} = 1636$ дней. Для дальнейшего выявления различий в исследуемых данных, рассматриваемый промежуток был разделен на $k = 1636$ сегментов Δ_i. Так как известно, что системы на каждом из сегментов Δ_i, как и в целом на промежутке $T_{эксп}$, находятся в априорно равных условиях, то на каждом из них можно выбирать для сравнительного анализа все имеющиеся данные об исследуемом параметре. Однако для сравнительного анализа оказались пригодны данные только на 1173 сегментах Δ_i, на остальных были обнаружены пропуски в данных, как частичные, так и полные. Вид таблиц с выбранными данными о параметре двух исследуемых ПОТС аналогичен и представлен на примере одной из них в табл. 1.

Таблица 1. Данные о параметре одной из систем
[Table 1. Data about the parameter of one of the systems]

№ сегмента	Время			
	00:00	...	22:00	23:00
0	31.55	...	32.05	30.71
...	...	...	...	...
1172	27.15	...	23.75	23.67

Отобранные данные по форме представляют собой временные ряды одинаковой длины. На этом фоне в качестве меры для их сравнительного анализа было выбрано стандартное расстояние Евклида, являющейся простым и интуитивно понятным вариантом оценки сходства между двумя временными x_i и y_i

$$d_{Евклидово} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \qquad (1)$$

Вид исследуемых данных при разных значениях меры сходства представлен на рис. 2.

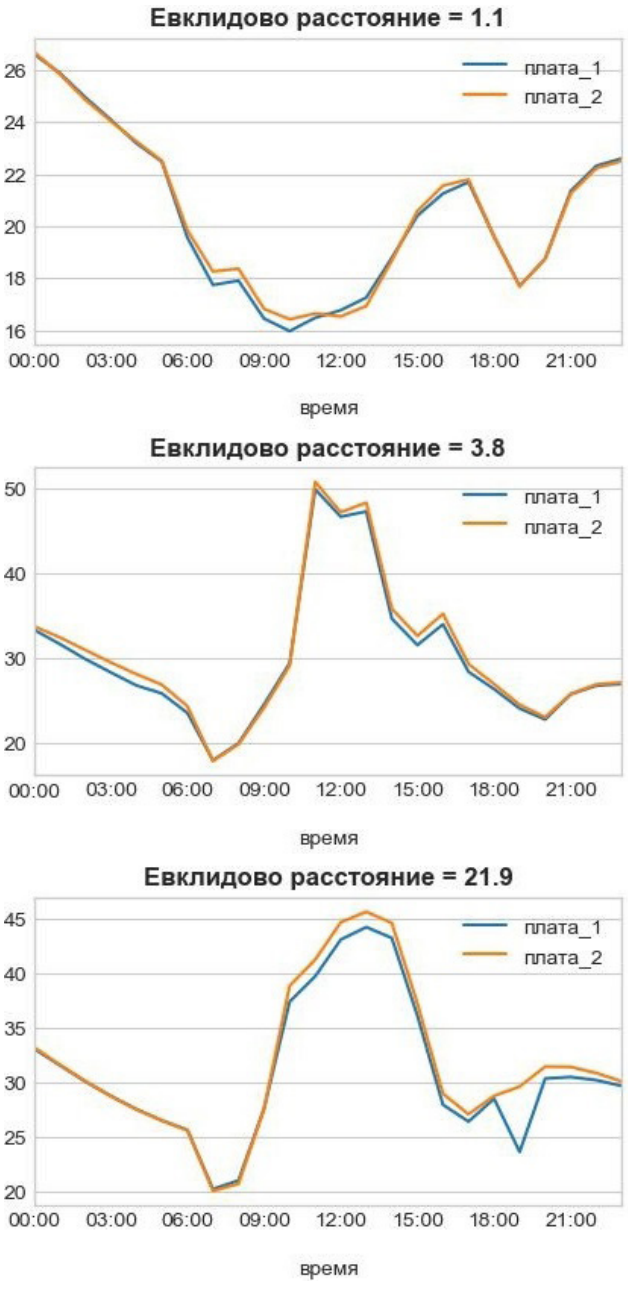

Рис. 2. Данные систем при различных значениях расстояния Евклида между ними
[Fig. 2. Data of systems at different values of the Euclidean distance between them]

Гистограмма расстояний для исследуемых данных на сегментах Δ_i представлена на рис. 3.

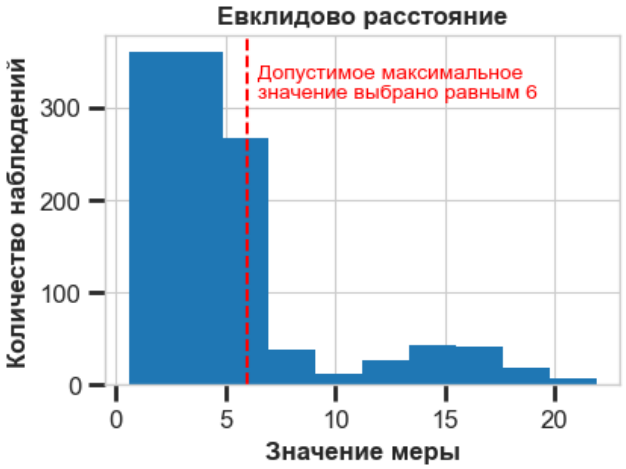

Рис. 3. Гистограмма значений расстояния Евклида для исследуемых данных
[Fig. 3. Histogram of Euclid distance values for the data under study]

В качестве допустимого максимального значения меры Евклидова расстояния было с привлечением эксперта выбрано значение равное «6». Таким образом, в работе для исследуемых данных принято, что любое превышающее число «6» значение Евклидова расстояния при сравнительном анализе данных об исследуемом параметре на сегменте Δ_i укажет на существование различия в данных о параметре и, напротив, при значениях расстояния равного или меньше «6» сравниваемые на сегменте Δ_i данные будут считаться практически одинаковыми.

Диаграмма, с отмеченными сегментами Δ_i, на которых данные о параметре ПОТС имели различия или где сравнительный анализ невозможен по той или иной причине представлена на рис. 4.

Анализируя диаграмму рис. 4. можно установить некоторые закономерности относительно исследуемых данных о параметре ПОТС. Например, на весьма протяженных участках времени сравнительный анализ невозможен ввиду полного отсутствия данных о ПОТС, таких сегментов Δ_i насчитывается 449шт., что составляет порядка 27 % всего числа промежутков(1636 шт.) рассматриваемого периода наблюдений. К сравнительному анализу на сегментах Δ_i также не были взяты данные с частичным отсутствием значений о параметре систем, которых было выявлено 14 шт. (0,008 % от общего числа) в самом начале рассматриваемого периода. Случаев, когда данные фиксировались только об одной системе на том или ином сегменте Δ_i не выявлено.

На диаграмме рис. 3. видно, что проблема с наличием полного отсутствия данных систем становится все более актуальной с течением времени. Кроме того, видно, что на 250 (порядка 15 % от общего числа) сегментах данные об исследуемом параметре обладают несходством, которое чаще фиксируется в холодное время года. Для теплого времени года также видна закономерность, связанная с ежегодным увеличением фактов фиксации несходства данных.

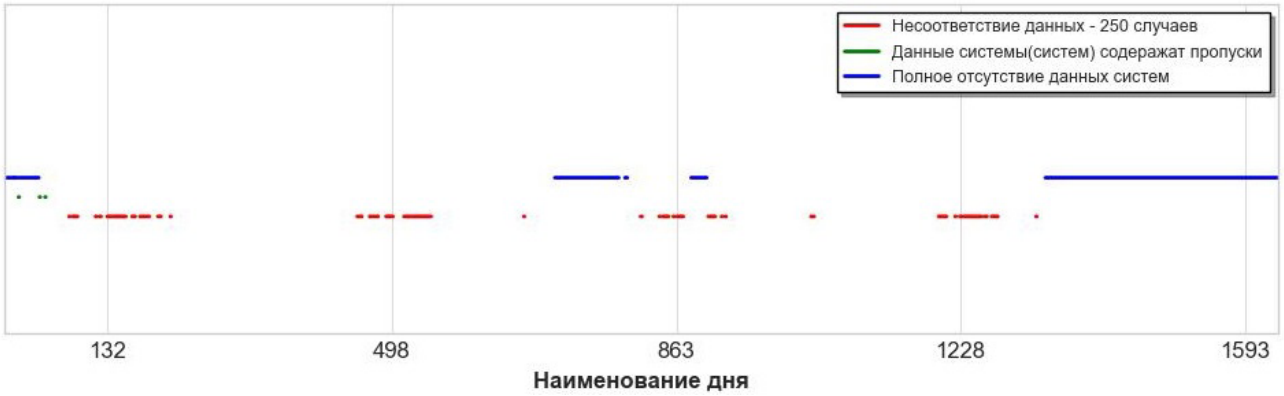

Рис. 4. Диаграмма, отображающая факты фиксации различий в данных о параметре ПОТС, а также их полное, либо частичное отсутствие. Значения [132, 498, 863, 1228, 1593] — календарные начала лет
[Fig. 4. A diagram showing the facts of fixing differences in the data on the PHTS parameter, as well as their complete or partial absence. Values [132, 498, 863, 1228, 1593] are calendar beginnings of years]

Таким образом, выяснение причин возникновения несходства данных, равно как и фактов фиксации их отсутствия, являет собой актуальную для исследуемых ПОТС задачу.

Относительно самого метода анализа данных о том или ином параметре ПОТС можно сказать о том, что он обладает как плюсами, так и минусами.

Для рассмотренной в примере ситуации, ПОТС, при их исправности постоянно находятся на всем периоде $T_{эксп}$ в априорно равных условиях. Это позволило легко определиться с данными о параметрах ПОТС, которые будут сравниваться между собой на каждом из сегментов Δ_i. Однако такие ПОТС встречаются далеко не всегда. Соответственно, выбор априорно равных условий и извлечение соответствующих им данных на каждом из сегментов Δ_i может являться сложной задачей.

Однако в случае неправильной формулировки априорно равных условиях или в случае ошибок при извлечении данных сравниваемые данные должны обладать существенными различиями, на что укажут дальнейшие результаты сравнительного анализа. Таким образом, анализируя факты фиксации различий в данных проблемы неправильной формулировки априорно равных условий и(или) ошибки в извлечении данных можно выявить и устранить.

Кроме того, нельзя не упомянуть и о субъективности метода. Число сегментов Δ_i, формулирование априорно равных условий, выбор метрики для сравнения данных на сегментах Δ_i между собой и ее величины, превышение которой будет указывать на существование различий в данных о параметре — все это определяется человеком.

Не дает метод ответа и на вопрос о том, данные какой из ПОТС, при фиксации их различий, стоит считать отражающими нормальное поведение систем на том или ином сегменте Δ_i, равно как и не говорит о причинах возникновения различия, что, однако, характерно для многих таких методов [6].

Кроме того, метод сосредотачивает внимание на сравнении данных систем в априорно равных условиях, то есть к сравнению на сегменте Δ_i могут подходить далеко не все данные о параметре. Также метод не обнаружит выбросы если они будут присутствовать в одно и то же время в сравниваемых данных всех исследуемых систем, однако, такие события на взгляд авторов маловероятны для большинства практических приложений.

Однако сходство сравниваемых данных дает дополнительную уверенность исследователю о том, что системы работают в нормальном режиме и их данные не содержат выбросов, что для решения задач отбора данных для машинного обучения перед построением прогностических моделей может быть весьма актуальным.

Следует отметить и то, что метод обладает гибкостью и может быть адаптирован к применению для анализа данных различных ПОТС. Он предполагает не только выявление фактов несходства данных в априорно равных условиях нахождения систем, но и отражает случаи, в которых сравнительный анализ данных на сегменте Δ_i был невозможен по той или иной причине. Эта информация может помочь в выявлении закономерности и неисправностей самих систем, так и сопутствующих им процессов сбора и хранения данных.

Кроме того, сравнение данных о параметре ПОТС не требует построения каких-либо моделей, которые определяли бы эталонное поведение данных о том или ином параметре ПОТС с которым будет проводиться сравнение данных систем. Это является существенным плюсом, так как построение адекватной модели сложной системы может являться крайне трудоемким и затратным процессом.

Достоинством метода также является и тот факт, что он не требует процедуры очистки исходных данных от выбросов. Напротив, с его помощью выбросы в данных обнаруживаются, при этом в отличие от массы статистических методов наличие выбросов не оказывает существенного влияния на их обнаружение, так как используются меры для оценки сходства данных.

ЗАКЛЮЧЕНИЕ

Обозначенный в работе метод сравнительного анализа данных о том или ином параметре ПОТС при их нахождении в априорно равных условиях является весьма перспективным для исследований данных имеющих форму временных рядов. Несмотря на субъективность, метод имеет существенный плюс в виде гибкости, кроме того, он не требует построения сложных моделей определяющих эталонное поведение исследуемого параметра ПОТС для проведения сравнительного анализа. Этот факт указывает на перспективность его применения для анализа данных даже сложных ПОТС.

Метод позволяет выявлять факты фиксации различий в данных об исследуемом параметре ПОТС с привязкой ко времени. При этом различия могут указывать на выбросы в данных, возникшие по разным причинам: неисправностям той или иной из систем, действием на них неучтенного фактора, нарушениями в процессах сбора и хранения данных. Наличие несходства также может говорить и о неправильном определении априорно равных условий нахождения систем, либо об ошибке в отборе данных для их сравнительного анализа.

Представление результатов сравнительного анализа в виде временной диаграммы с нанесением на нее также информации о случаях, в которых сравнительный анализ данных на сегменте Δ_i был невозможен по той или иной причине, в том числе из-за отсутствия данных, также несет определенную пользу. На основании ее анализа можно выявлять закономерности и строить гипотезы как относительно данных об исследуемом параметре, так и о самих ПОТС, которых они описывают. Кроме того, с ее помощью можно оценивать сложившийся на предприятии подход к сбору и хранению данных, а также принимать направленные на его совершенствование решения.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Золотова Т. В.* Методы интеллектуальной обработки данных для коррекции атипичных значений котировок акций / Т. В. Золотова, Д. А. Волкова // Статистика и экономика. – 2022. – Т. 19, № 2. – С. 4–13. DOI 10.21686/2500-3925-2022-2-4-13.

2. Problems in analyzing time series with gaps and their solution with the winabd software package / A. V. Desherevskii [et al.] // Izvestiya, Atmospheric and Oceanic Physics. – 2017. – Vol. 53, № 7, – P. 659–678. DOI 10.1134/s0001433817070027.

3. *Концевая Н. В.* Анализ методов заполнения пропусков во временных рядах показателей финансовых рынков / Н. В. Концевая // Вестник Воронеж. гос. ун-та. – 2012. – № 8. – С. 18–20.

4. Очистка сенсорных данных в интеллектуальных системах управления отоплением зданий / М. Л. Цымблер [и др.] // Вестник Южно-Уральского гос. ун-та. Сер. Вычислительная математика и информатика. – 2021. – № 3. – С. 16–36. DOI: 10.14529/cmse210302.

5. *Turner W.J.N.* Residential HVAC fault detection using a system identification approach / W.J.N. Turner, A. Staino, B. Basu // Energy and Buildings. – 2017. – Vol. 151. – P. 1–17. DOI 10.1016/j.enbuild.2017.06.008.

6. Large-scale monitoring of operationally diverse district heating substations: A reference-group based approach / S. Farouq [et al.] // Engineering Applications of Artificial Intelligence. – 2020. – Vol. 90. – 103492 DOI 10.1016/j.engappai.2020.103492.

7. *Генералов И. Г.* Аномалии в структуре временных рядов при оценке устойчивости производства зерна / И. Г. Генералов, О. Е. Завиваева, С. А. Суслов // Азимут научных исследований: экономика и управление. – 2019. – Т. 8, № 4(29). С. 351–354. DOI: 10.26140/anie-2019-0804-0080.

8. *Zimek A. and Filzmoser P.* There and back again: Outlier detection between statistical reasoning and data mining algorithms / A. Zimek, P. Filzmoser // WIREs Data Mining and Knowledge Discovery. – 2018. – Vol. 8, № 6 – 1280 DOI 10.1002/widm.1280.

9. Time Series K-means: A new K -means type smooth subspace clustering for time series data / X. Huang [et al.] // Information Sciences. – 2016. – Vol. 367–368, – P. 1–13 DOI 10.1016/j.ins.2016.05.040.

10. *Пестунов И. А.* Алгоритмы кластеризации в задачах сегментации спутниковых изображений / И. А. Пестунов, Ю. Н. Синявский // Вестник КемГУ. – 2012. –Т. 2, № 4(52). – С. 110–125.

11. *Зуева В. Н.* Регрессионные методы прогнозирования графика нагрузки электрооборудования / В. Н. Зуева // Научный журнал КубГАУ. – 2017. – №126(02). – С. 1–12. DOI: 10.21515/1990-4665-126-008.

12. Способ обработки результатов кавитационных испытаний насосов турбонасосных агрегатов с целью получения аппроксимирующей функции / А. С. Торгашин [и др.] // Сибирский аэрокосмический журнал. – 2022. – Т. 23, № 3. – С. 498–507. DOI: 10.31772/2712-8970-2022-23-3-498-507.

13. Development of bicarbonate buffer flow-through cell dissolution test and its application in prediction of in vivo performance of colon targeting tablets / S. Ikuta [et al.] // European Journal of Pharmaceutical Sciences. – 2023.– Vol. 180. – 106326 DOI 10.1016/j.ejps.2022.106326.

14. *Боярский М. В.* О выявлении и исключении аномальных результатов наблюдений в процессах деревообработки / М. В. Боярский // Известия ВУЗов. Лесной журнал. – 2003. – № 1. – С. 66–70.

15. *Попукайло В. С.* Построение математической модели эффективности работы банка в условиях малой выборки / В. С. Попукайло // Актуальные проблемы гуманитарных и естественных наук. – 2016. – № 3-2. – С. 1–6.

16. *Воданюк С. А.* Практика применения сравнительного подхода к оценке прав требования (дебиторской задолженности) / С. А. Воданюк // Имущественные отношения в РФ. – 2013. – №6 (141). – С. 42–53.

17. *Петровская Ю. А.* Методы исключения грубой погрешности / Ю. А Петровская, Е. А. Петровская, М. С. Эльберг // Актуальные проблемы авиации и космонавтики. – 2015. – №11. – С. 109–110.

18. A statistical and numerical modeling approach for spatiotemporal reconstruction of glaciations in the Central Asian Mountains / S. Saha [et al.] // MethodsX. – 2020.– Vol. 7, 100820 DOI 10.1016/j.mex.2020.100820.

19. *Муравьев, П. А.* Очистка временных рядов от выбросов / П. А. Муравьев, М. Н. Соловьев, М. В. Дементьев // Здоровье и образование в XXI веке. – 2011.– №3.– С. 328–329.

20. *Позолотин В. Е.* Применение алгоритмов преобразования данных при анализе временных рядов на предмет устранения выбросов / В. Е. Позолотин, Е. А. Султанова // Программные системы и вычислительные методы. – 2019. – №2. – С. 33–42. DOI: 10.7256/2454-0714.2019.2.28279

Дулесов Александр Сергеевич — д-р техн. наук, профессор кафедры «Цифровых технологий и дизайна» Хакасского государственного университета им. Н. Ф. Катанова.
E -mail: dulesov@khsu.ru
ORCID iD: https://orcid.org/0000-0001-6371-0171

Байшев Анатолий Викторович — аспирант 3-го года обучения кафедры «Цифровых технологий и дизайна» Хакасского государственного университета им. Н. Ф. Катанова.
E-mail: anatoly_bayshev@mail.ru
ORCID iD: https://orcid.org/0000-0002-3336-6928

DOI: https://doi.org/10.17308/sait/1995-5499/2023/3/121-133 ISSN 1995-5499
Received 06.12.2022
Accepted 30.09.2023

DETECTION OF OUTLIERS IN THE DATA IS PRACTICALLY HOMOGENEOUS TECHNICAL SYSTEMS

© 2023 A. S. Dulesov, A. V. Bayshev✉

Khakass State University N. F. Katanov
90, Lenin Avenue, Republic of Khakassia, Abakan 655017, Russian Federation

Annotation. The article presents a method for detecting outliers in data on a particular parameter of practically homogeneous technical systems (hereinafter referred to as PHTS) in the form of time series by their comparative analysis, reflecting cases of impossibility of its implementation. The method is very relevant due to the widespread use of time series in various modern technical systems. To develop the method, a non-exhaustive review of various methods for detecting defects in data was carried out. The main attention in the review was focused on the detection of outliers, since such a defect as omissions in the data is present in them almost explicitly. Some advantages and disadvantages of outlier detection methods are considered, taking into account which the method was developed. The method has broad prospects for its further application, as it can help in identifying outliers in the data of the systems under study, and in case of data similarity, it can give additional confidence to the researcher that the systems whose data are being studied were at the designated time in good condition and almost equal conditions. In turn, the difference may indicate the presence of outliers in the data, the causes of which may be: malfunctions of one or another of the systems or in the systems for collecting and storing data about them, the influence of an unaccounted factor on one or another system. Despite some subjectivity, the method has a significant plus in the form of flexibility, in addition, it does not require the construction of complex models that determine the reference behavior of the investigated parameter of the PHTS for a comparative analysis of the model data and the systems under study. This fact indicates the promise of its application for data analysis even of complex PHTS.
Keywords: system, system analysis, time series, patterns, a priori information, almost identical technical systems, similarity measures, outliers, data, defects.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Zolotova T. V. and Volkova D. A.* (2022) Intelligent data processing methods for the atypical values correction of stock quotes. *Statistics and Economics*. 19 (2). P. 4–13. DOI 21686/2500-3925-2022-2-4-13.

2. *Desherevskii A. V. [et al.]* (2017) Problems in analyzing time series with gaps and their solution with the WinABD software package. *Izvestiya, Atmospheric and Oceanic Physics*. 53(7). P. 659–678. DOI 10.1134/s0001433817070027.

3. *Kontsevaya N. V.* (2012) The analysis of methods of filling of admissions in temporary ranks of indicators of the financial markets. *Vestnik Voronezhskogo gosudarstvennogo universiteta*. (8). P. 18–20. (in Russian)

4. *Zymbler M. L. [et al.]* (2021) Cleaning sensor data in Intelligent Heating Control System. *Bulletin of the South Ural State University. Series "Computational Mathematics and Software Engineering"*. 10(3). P. 16–36. DOI 10.14529/cmse210302.

5. *Turner W. J. N., Staino A. and Basu B.* (2017) Residential HVAC fault detection using a system identification approach. *Energy and Buildings*. 151. P. 1–17. DOI 10.1016/j.enbuild.2017.06.008

6. *Farouq S. [et al.]* (2020) Large-scale monitoring of operationally diverse district heat-

✉ Байшев Анатолий Викторович
e-mail: anatoly_bayshev@mail.ru

ing substations: A reference-group based approach. *Engineering Applications of Artificial Intelligence*. 90. P. 103–492. DOI 10.1016/j.engappai.2020.103492.

7. Generalov I. G., Zavivaeva O. E. and Suslov S. A. (2019) Anomalies in the structure of time series when assessing the stability of grain production. *Azimuth of scientific research: economics and administration*. 8(29). P. 351–354. DOI 10.26140/anie-2019-0804-0080

8. *Zimek A. and Filzmoser P.* (2018) There and back again: Outlier detection be-tween statistical reasoning and data mining algorithms. *WIREs Data Mining and Knowledge Discovery*. 8(6). P. 1280. DOI 10.1002/widm.1280.

9. *Huang X. [et al.]* (2016) Time Series K-means: A new K-means type smooth subspace clustering for time series data. *Information Sciences*. 367(368). P. 1–13. DOI: 10.1016/j.ins.2016.05.040.

10. *Pestunov I. A. and Sinyavskiy Yu. N.* (2012) Clustering algorithms in satellite images segmentation tasks. *Vestnik Kemerovskogo gosudarstvennogo universiteta. Series "Mathematics"*. 4(52). P. 1–13. (in Russian)

11. *Zueva V. N.* (2017) Regressive methods of prognostication of the load-graph of electrical equipment. *Polythematic Online Scientific Journal of Kuban State Agrarian University*. 126(2). P. 1–12. DOI 10.21515/1990-4665-126-008.

12. *Torgashin A. S. [et al.]* (2022) Method for processing the results of cavitation tests of TNA pumps in order to obtain an approximating function. *Siberian Aerospace Journal*. 23(3). P. 498–507. DOI 10.31772/2712-8970-2022-23-3-498-507.

13. *Ikuta S. [et al.]* (2023) Development of bicarbonate buffer flow-through cell dissolution test and its application in prediction of in vivo performance of colon targeting tablets. *European Journal of Pharmaceutical Sciences*. (180), 106326. DOI: 10.1016/j.ejps.2022.106326

14. *Boyarsky M. V.* (2003) On the detection and elimination of anomalous results of observations in woodworking processes. *Izvestiya VUZov. Forest magazine*. 1. P. 66–70. (in Russian)

15. *Popukailo V. S.* (2016) Construction of a mathematical model for the efficiency of a bank in a small sample. *Actual problems of the humanities and natural sciences*. 3(2). P. 1–6. (in Russian)

16. *Vodanyuk S. A.* (2013) The practice of applying a comparative approach to the assessment of claims (receivables). *Property relations in the Russian Federation*. 6(141). P. 42–53. (in Russian)

17. *Petrovskaya Y. A.* (2015) Methods for Eliminating Gross Error. *Actual problems of aviation and cosmonautics*. (11). P. 109–110. (in Russian)

18. *Saha S. [et al.]* (2020) A statistical and numerical modeling approach for spatiotemporal reconstruction of glaciations in the Central Asian Mountains. *MethodsX*. (7). 100820 DOI: 10.1016/j.mex.2020.100820.

19. *Muravyov P. A. and Dementiev M. V.* (2011) Cleaning time series from outliers. *Health and education in the XXI century*. 3. P. 328–329. (in Russian)

20. *Pozolotin V. E. and Sultanova E. A.* (2019) Application of data transformation algorithms in the analysis of time series to eliminate outliers. *Software systems and computational methods*. 2(2). P. 33–42. DOI 10.7256/2454-0714.2019.2.28279.

Dulesov Alexander S. — Doctor of Engineering. Sci., Professor of the Department of Digital Technologies and Design, Khakass State University named after N. F. Katanov.
E-mail: dulesov@khsu.ru
ORCID iD: https://orcid.org/0000-0001-6371-0171

Bayshev Anatoly V. — 3-rd year postgraduate student of the Department of Digital Technologies and Design, Khakass State University named after N. F. Katanov.
E-mail: anatoly_bayshev@mail.ru
ORCID iD: https://orcid.org/0000-0002-3336-6928