УДК 004.4 DOI: 10.26102/2310-6018/2021.34.3.001

Выявление аномалий в многомерных временных рядах с помощью пакета на языке R

Э.С. Раюшкин, М.В. Щербаков, И.Д. Казаков, В.О. Колесникова
Волгоградский государственный технический университет, Волгоград, Российская Федерация

Резюме. Задача поиска аномалий в данных встречается при реализации систем предиктивной аналитики, ставшей очень популярной за последние несколько лет. Предиктивная аналитика – это возможность организаций предсказывать данные на небольшой период времени, тем самым заранее угадывая возможные кризисы или непредвиденные случаи в работе систем на основе уже существующих и поступающих данных. Но предиктивная аналитика достаточно сложна, и поэтому ее реализация также сопряжена с трудностями. Когда компании применяют традиционный подход к предиктивной аналитике (то есть относятся к ней как к любому другому типу аналитики), они часто сталкиваются с препятствиями. Именно поэтому данная область нуждается в инструментах выявления аномалий в данных. Эти инструменты должны помогать выявлять выделяющиеся значения для того, чтобы проводить зависимости с факторами их возникновения и выявлять их в будущем. В данной статье описан пакет на языке R (совокупность R функций, данных и документации к ним, собранных в единое целое), разработанный для выявления аномалий в многомерных временных рядах. Данный пакет способен выявлять аномалии с помощью трех различных методов: метода n-сигм, CUSUM-метода и метода центральных моментов 4 порядка. Также данный пакет производит поиск комплексных аномалий, которые являются прямым показателем ошибки в системе благодаря тому, что аномалии обнаружены в многомерных данных.

Ключевые слова: аномалия, выброс, временные ряды, правило трех сигм, язык R.

Для цитирования: Раюшкин Э.С., Щербаков М.В., Казаков И.Д., Колесникова В.О. Выявление аномалий в многомерных временных рядах с помощью пакета на языке R. *Моделирование, оптимизация и информационные технологии.* 2021;9(3). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=948 DOI: 10.26102/2310-6018/2021.34.3.001

Detecting anomalies in multidimensional time series using the R package

E.S. Rayushkin, M.V. Shcherbakov, I.D. Kazakov, V.O.Kolesnikova
Volgograd State Technical University, Volgograd, Russian Federation

Abstract: The task of finding anomalies in data can be found when implementing predictive analytics systems. Predictive analytics has become very popular over the past few years. It helps banks approve loans or identify suspicious account activity, email providers filter spam, and retailers predict the likelihood of buying to attract customers. But predictive analytics is quite complex, and therefore its implementation is also fraught with difficulties. When companies take the traditional approach to predictive analytics (that is, treat it like any other type of analytics), they often face obstacles. That is why this area needs tools to detect anomalies in the data. These tools should help identify outstanding values to draw dependencies with the factors of their occurrence and identify them in the future. This article describes a package in the R language that is anomalies in multidimensional time series. This package enables detecting anomalies using three different methods: the n-sigma method, the CUSUM method, and the 4th order central moment method. Also, this package searches for complex anomalies, which are a direct indicator of errors in the system since anomalies are discovered in multidimensional data.

Keywords: anomaly, outlier, time Series, three Sigma Rule, R Language

For citation: Rayushkin E.S., Shcherbakov M.V., Kazakov I.D., Kolesnikova V.O. Detecting anomalies in multidimensional time series using the R package. *Modeling, Optimization and Information Technology*. 2021;9(3). Available from: https://moitvivt.ru/ru/journal/pdf?id=948 DOI: 10.26102/2310-6018/2021.34.3.001 (In Russ).

Введение

Данная статья посвящена актуальной проблеме поиска аномалий в многомерных данных. Такая задача встречается при реализации систем предиктивной аналитики. Она помогает провайдерам электронной почты - фильтровать спам, ритейлерам - прогнозировать вероятность того, что необходимо закупать для привлечения клиентов, а банкам утверждать кредит или выявлять подозрительные действия со счетами [1,2].

Выявление ошибок или критических отказов в оборудовании, рост или спад продаж в магазине, а также другие различные ситуации, в которых происходит резкое изменение данных относительно уже сложившийся тенденции являются острой проблемой на сегодняшний день. Возможность предсказания поломки какого-либо механизма, детали механизма или любую другую вещь, которая способна кардинально снизить убытки на производстве, является актуальной задачей. Для решения данной задачи подходит статистический язык программирования R. Основой данного языка являются пакеты. Пакет языка R — это совокупность R функций, данных и документации к ним, собранных в единое целое.

Целью данного исследования является разработка программного продукта, который позволил бы выявлять аномалии в многомерных временных данных, являющиеся показателем ошибок в работе системы.

Методы и алгоритмы

Аномалии и многомерные временные ряды

Временной ряд – это собранный в различные моменты времени материал, содержащий статистические данные какого или каких-либо параметров за определённый промежуток времени. Иначе говоря, временной ряд – это упорядоченная по времени последовательность значений какого-либо показателя. Такие данные довольно легко собираются и имеют высокую значимость для определения состояний системы, так как они могут быть получены в реальном времени [3]. Пример временного ряда представлен на Рисунке 1.

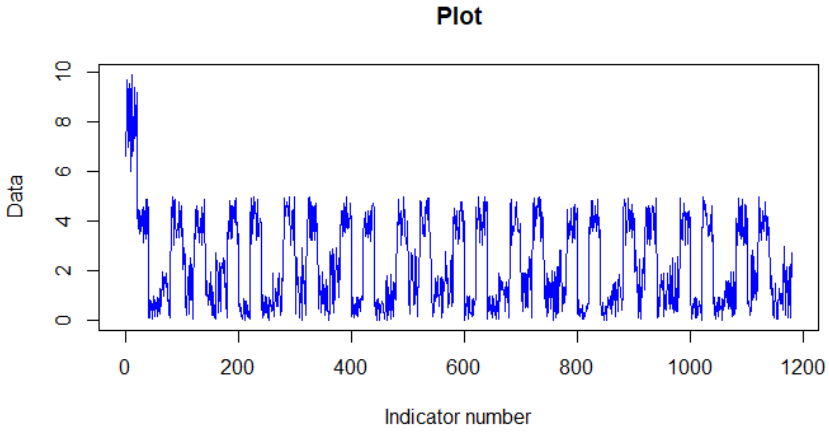

Рисунок 1 – Временной ряд
Figure 1 - Time series

Данный пример является одномерным временным рядом, поскольку в каждый момент времени измеряется и известна только одна числовая характеристика изучаемого явления. Однако на практике многие явления характеризуются не одной, а целым набором (совокупностью) характеристик, возможно взаимосвязанных между собой (изменение одной характеристики влечет изменение других). В этом случае говорят о многомерных временных рядах.

Многомерные временные ряды – это одна из основных областей применения многомерного статистического анализа, временные ряды, построенные на основе многочисленных показателей, характеризующих наблюдаемые экономические процессы на каждом этапе наблюдения [4]. Пример многомерного временного ряда представлен на Рисунке 2. На данном рисунке изображены многомерные временные ряды, характеризующие поведение всех датчиков случайной выборки 10 двигателей, которые были исследованы аспирантом ВолгГТУ Сай Ван Квонгом в статье «Метод прогнозирования остаточного ресурса на основе обработки данных многообъектных сложных систем» [5].

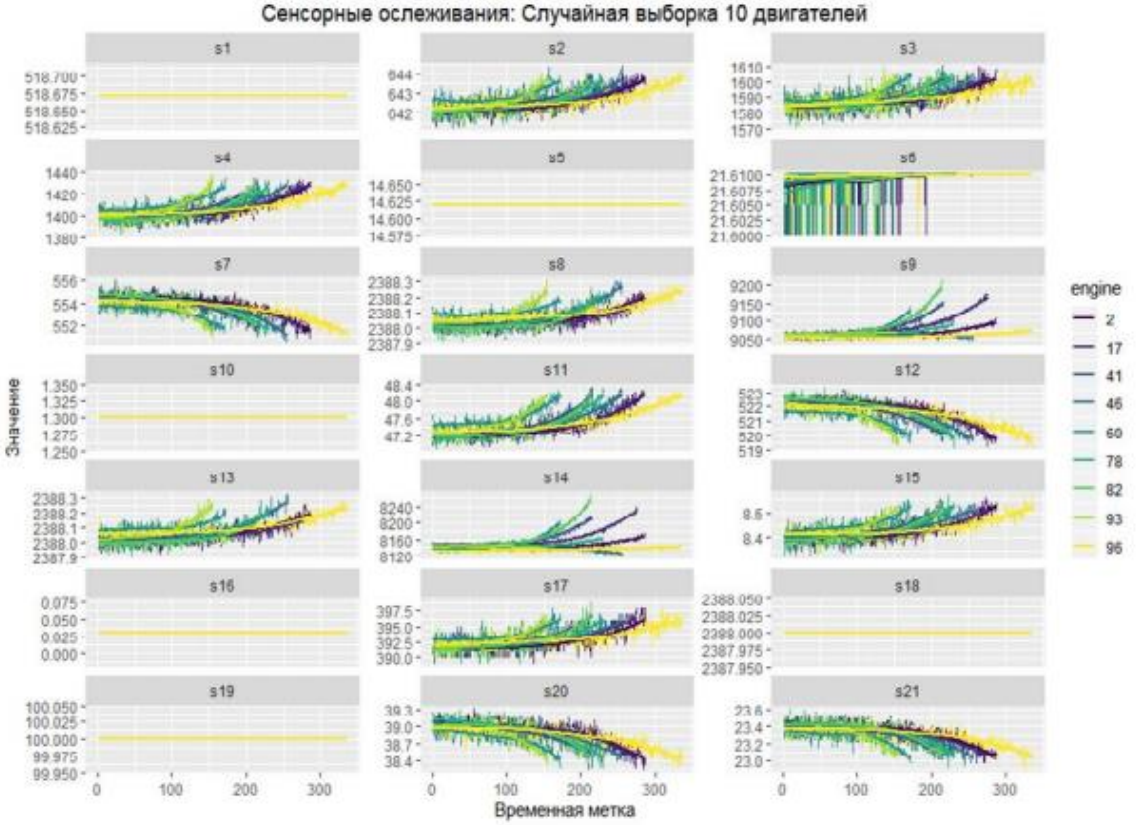

Рисунок 2 – Случайная выборка 10 двигателей, состоящая из многомерных временных рядов
Figure 2 - A random sample of 10 engines, consisting of multivariate time series

Аномалии — это данные, не соответствующие четко определенному понятию нормального поведения [6]. В данных временных рядов аномалией или выбросом называют точку данных, которая не следует общему коллективному тренду, сезонному или циклическому шаблону всех данных и значительно отличается от остальных данных. Под значительным большинство специалистов по данным подразумевают статистическую значимость, которая, по порядку слов, означает, что статистические свойства точки данных не совпадают с остальной частью ряда [7].

В анализе данных есть направление, которое занимается поиском аномалий: детектирование выбросов (Outlier Detection). Выброс — это объект, который отличается по своим свойствам от объектов выборки [8].

Выбросы являются следствием:

- ошибок в данных (неточности измерения, округления, неверной записи и т.п.);
- наличия шумовых объектов (неверно классифицированных объектов);
- присутствия объектов «других» выборок (например, показаниями сломавшегося датчика).

Пример выбросов на временном ряду представлен на Рисунке 3.

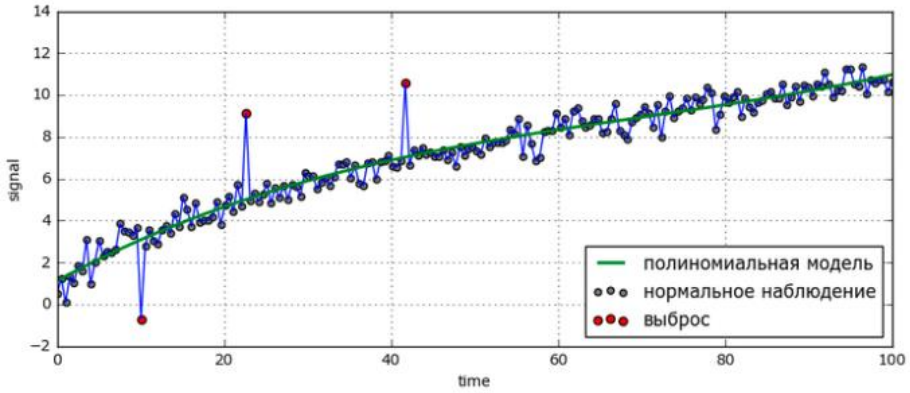

Рисунок 3 – Пример выбросов на временном ряду
Figure 3 - Example of outliers on a time series

Важность выявления аномалий в данных обусловлена тем, что они всегда содержат важную информацию необходимую для различных областей применения. Большой трафик в сети может, например, означать, что компьютер был взломан и в данный момент отправляет данные стороннему лицу. А аномалии на аппарате МРТ могут указать на наличие опухолей [9].

Для более точного поиска аномалий на многомерном наборе данных, необходимо производить поиск аномалий по всем многомерным данным. Данные аномалии, которые случаются в один и тот же момент времени на нескольких показателях, называются комплексными аномалиями.

Результаты

Входные и выходные данные

Имеется набор синтетических данных. Синтетические данные – искусственные данные, сгенерированные компьютером, но выглядящие аналогично реальным. Имеющийся набор синтетических данных, сгенерирован, похожим на процесс работы оборудования [10]. Данные содержаться в файле формата csv, и загружаются оттуда в датафрейм. Данные состоят из заданного количества измерений. Началом измерений является 1, концом измерений является 1180 (последний) номер выборки данных. Количество измерений - 1180, параметров измерений - 1. Структура данных представлена в Таблице 1.

Таблица 1 - Перечень измеряемых параметров
Table 1 - List of measured parameters

Идентификатор параметра	Обозначение	Единица измерения
Number	Номер измерения	numeric
FirstData	Первое значение	numeric
SecondData	Второе значение	numeric

Идентификатор параметра	Обозначение	Единица измерения
ThirdData	Третье значение	numeric
FourthData	Четвертое значение	numeric

В каждом столбце данного датафрейма содержатся числовые значения. Минимум и максимум числового значения, а также индекс элемента им соответствующий приведены в Таблице 2.

Таблица 2 – Максимумы и минимумы столбцов
Table 2 – Column's highs and lows

Название столбца	Индекс минимального элемента	Индекс максимального элемента	Минимальное значение	Максимальное значение
Number	1	1180	1	1180
FirstData	211	11	0	9,92
SecondData	68	30	0	9,91
ThirdData	103	30	0	9,91
FourthData	181	19	0	9,99

Выходными данными является график зависимости данных от индекса, с выделенными аномальными значениями, которые были отобраны по одному из трех методов или методом поиска комплексных аномалий. Пример выходных данных изображен на Рисунке 4.

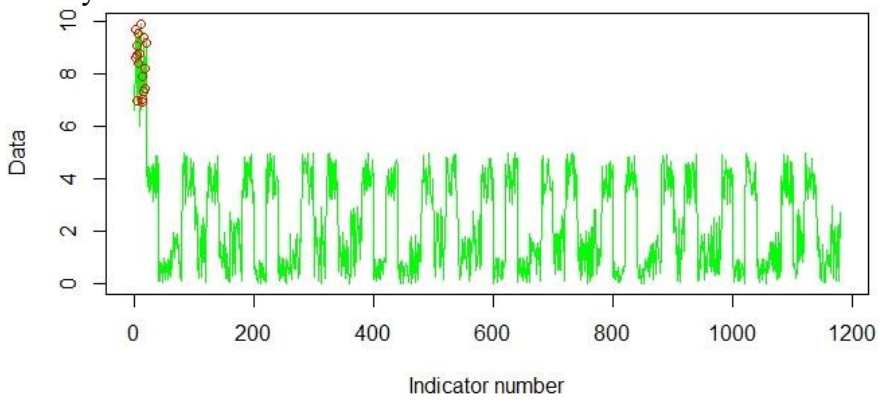

Рисунок 4 – Пример выходных данных работы программы
Figure 4 - An example of the output data of the program

Разработанный пакет на языке R для выявления аномалий в многомерных временных рядах состоит из четырех функций:

- поиск аномалий временного ряда;
- поиск комплексных аномалий многомерных временных рядов;
- построение графика аномалий на временных рядах;
- построение графика комплексных аномалий на временных рядах.

Функции поиска аномалий временного ряда, производит поиск аномалий по трем различным методам: метод n-сигм, CUSUM метод и метод центральных моментов четвертого порядка. Для каждого из методов задается коэффициент поиска, чтобы

отбирать необходимые аномальные значения. Данная функция получает на вход столбец датафрейма и возвращает его же, но с отобранными аномальными значениями.

Функция поиска комплексных аномалий многомерных временных рядов, производит отбор комплексных аномалий среди всех измерений многомерного временного ряда. Функция получает на вход датафрейм с аномальными значениями и возвращает его, добавляя столбец, в котором хранится количество аномальных значений.

Функция построения графика аномалий на временных рядах строит график зависимости данных от метрики (даты, времени или индекса), и отмечает на данном графике аномальные значения. Функция получает столбец, содержащий метрику, столбец значений, в котором проводился поиск аномальных значений, и столбец, в котором произведен поиск аномальных значений и возвращает построенный график.

Функция построения графика комплексных аномалий строит гистограмму зависимости комплексных аномалий от метрики. Функция получает столбец, содержащий комплексные аномалии и столбец метрики, и возвращает построенный график.

Для каждого из разработанных методов была проверена работоспособность пакета. Были проведены опыты с данными, для выявления точности отбора аномальных значений. В столбцах с данными было сгенерировано 20 аномальных значений с 1 по 20 элементы датафрейма.

Проверим работоспособность метода поиска аномалий методом n-сигм. Возьмем для проверки значения n = 3. График, полученный в результате отбора аномалий представлен на Рисунке 5.

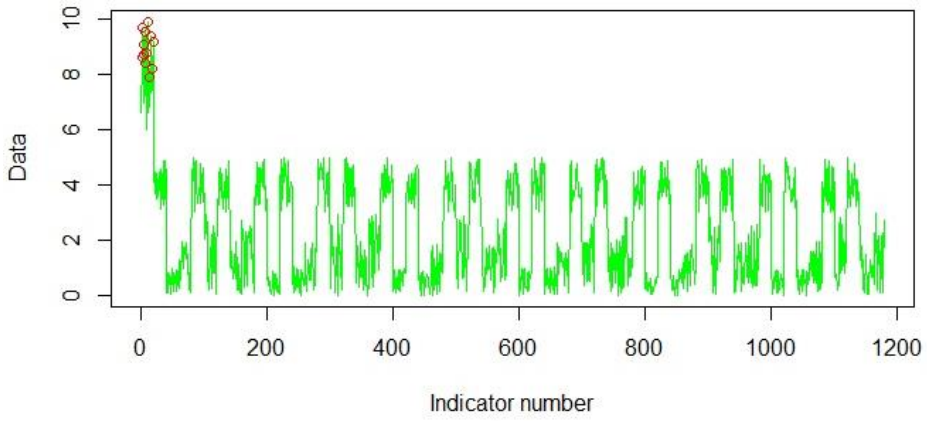

Рисунок 5 - Поиск аномалий по методу n-сигм, с параметром n равным 3
Figure 5 - Search for anomalies using the n-sigma method, with parameter n equal to 3

При применении данного алгоритма были рассчитаны 12 аномальных значений. При 20 действительно аномальных значений, 12 аномальных значений составляют 60% от всех существующих аномалий. Данный метод не отобрал 8 аномальных значений. 60% показатель поиска аномалий считаем удовлетворительным результатом, однако проверим другие методы, и определим, существует ли возможность улучшить данный результат.

Было выдвинуто предположение о том, что для более точного поиска аномалий необходимо использовать метод центральных моментов четвертого порядка, в котором необходимо изменить поиск коэффициента эксцесса с единичного значения на промежуток. Данный тип поиска должен увеличить качество поиска аномальных значений. График результата работы данного метода представлен на Рисунке 6.

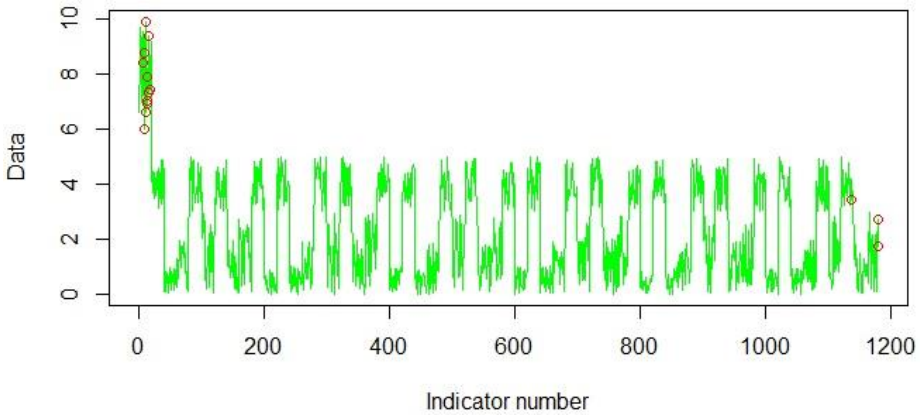

Рисунок 6 – Поиск аномалий по методу центральных моментов четвертого порядка с промежутком равным 110
Figure 6 - Search for anomalies using the method of central moments of the fourth order with an interval of 110

При применении алгоритма с промежутком, равным 110 были рассчитаны 14 аномальных значений. При 20 действительно аномальных значений, 14 аномальных значений составляют 70% от всех существующих аномалий. Данный метод отобрал 11 реально существующих аномалий и 3 не аномальных значения. Соотношение количества действительных аномалий к общему числу аномалий составило 55%. Можно сделать вывод, что для более точного определения аномалий необходимо подобрать такой промежуток, в котором аномальные значения будут наиболее точно выражены. Данный метод не является удовлетворительным, по сравнению с методом n-сигм. Однако данный результат отбирает больше 50% аномалий верно, что является неплохим показателем работы данного метода.

Второе предположение заключалось в том, что, для более точного поиска аномалий необходимо использовать CUSUM метод. В зависимости от задаваемых пороговых значений поиска, количество аномалий будет изменяться. График результата работы данного метода представлен на Рисунке 7.

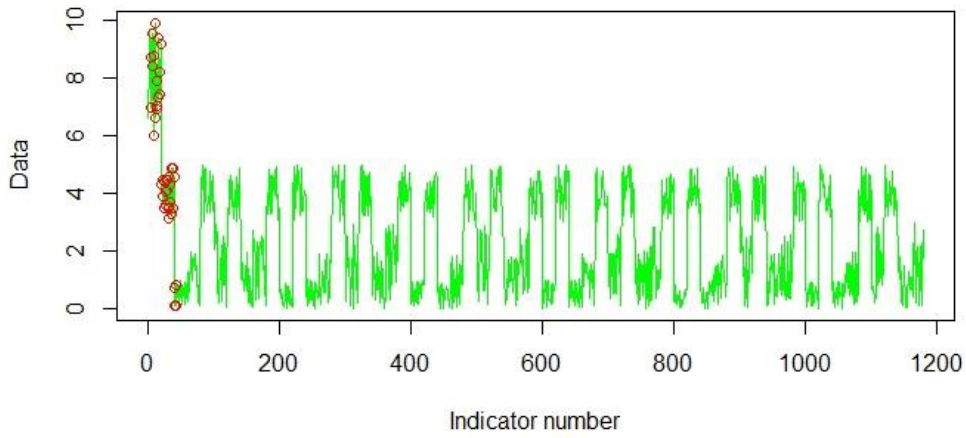

Рисунок 7 - Поиск аномалий методом кумулятивной суммы с пороговым значением 5
Figure 7 - Search for anomalies using the cumulative sum method with a threshold value of 5

При применении алгоритма с пороговым значением 5 были рассчитаны 40 аномальных значений. При 20 действительно аномальных значений, 40 аномальных значений составляют 200% от всех существующих аномалий. Данный метод отобрал 16

реально существующих аномалий и 24 не аномальных значения. Соотношение количества действительных аномалий к общему числу аномалий составило около 40%.

При выборе различного порогового значения повышается точность алгоритма. Однако для повышения эффективности определения аномалий необходимо подобрать пороговое значение. Данный алгоритм считает аномальными и те значения, которые возвращаются к нормальным после аномальных. Поэтому эффективность данного алгоритма ниже, чем у предыдущих.

Самым эффективным из разработанных алгоритмов, оказался алгоритм на основе метода n-сигм. Однако, для более точного поиска аномалий на наборе данных, необходимо производить поиск аномалий по всем столбцам данных. Данные аномалии называются комплексными аномалиями. Для их поиска необходимо отобрать все аномалии, в многомерном временном ряду и выделить их в отдельный набор данных. После чего необходимо произвести подсчет количества аномалий из всех столбцов с одним и тем же индексом. Значения, которые будут выше, чем количество столбцов минус n (где n-задаваемое число) будут являться аномалиями. Отбор данных проведем по методу n-сигм, с коэффициентом n равным 3, поскольку данный метод показал себя как наиболее эффективный. Гистограмма, отображающая комплексные аномалии представлена на Рисунке 8.

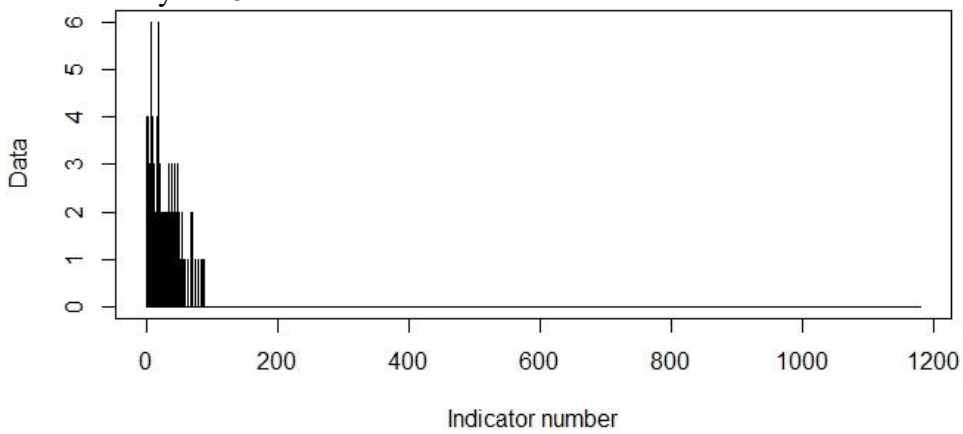

Рисунок 8 - Гистограмма комплексных аномалий
Figure 8 - Histogram of complex anomalies

Комплексные аномалии отражают точнее отражают ошибки в работе системы, поскольку определяют только те значения, при которых аномалии были замечены на нескольких показателях. Это говорит о том, что при единичном изучении аномалии, аномальное значение может являться выбросом или случайной аномалией, тогда как комплексные аномалии показывают полноценный подход к поискам ошибок в работе системы.

Заключение

Областью применения, разработанного программного продукта является: обработка данных для поиска аномальных значений во временных рядах. Данный пакет позволяет пользователю загрузить данные в виде многомерного временного ряда и отследить в какой момент времени значения того или иного показателя являлись аномальными. Также данный пакет производит поиск комплексных аномалий, которые являются прямым показателем ошибки в системе благодаря тому, что аномалии обнаружены в многомерных данных. Данный показатель способен помочь обнаруживать сбои в работе той или иной системы, что позволит предсказать такие

вещи как: выход из строя сложного технического оборудования или, например, несанкционированные операции с банковским счетом.

ЛИТЕРАТУРА

1. Антоненко С.В. Эффективность банковских систем безопасности, основанных на машинном обучении. *Материалы II международной научно-практической конференции «Тенденции и перспективы развития банковской системы в современных экономических условиях» Брянск, 17-18 декабря 2019 г. - Брянский государственный университет имени академика И.Г. Петровского*. 2020; 97-102.
2. Временной ряд (Time series data) *Логином*. Доступно по: https://wiki.loginom.ru/articles/time-series.html (дата обращения: 12.03.2021).
3. Рафикул И., Назим С., Мохаммад Али М., Прохоллад С., Бушра Р. Комплексное исследование обнаружения аномалий во временных рядах данных социальных сетей в Интернете. *Международный журнал компьютерных приложений*. 2017;180(3):13-22.
4. Экономико-математический словарь - Многомерные временные ряды *Академик.* Доступно по: https://economic_mathematics.academic.ru/2615/Многомерные_временные_ряды (дата обращения: 12.03.2021).
5. Ван Квонг С., Щербаков М. В. Метод управления данными, для прогнозирования оставшегося срока полезного использования многокомпонентных систем. *Каспийский журнал: Контроль и высокие технологии*. 2019;1:33–44.
6. Чандола В., Банерджи А., Кумар В. Обнаружение аномалий: обзор. *ACM Computing Surveys.* 2009;41(3).
7. Эффективные подходы к обнаружению аномалий временных рядов Towards data science. Доступно по: https://towardsdatascience.com/effective-approaches-for-time-series-anomaly-detection-9485b40077f1 (дата обращения: 12.03.2021).
8. Поиск аномалий (Anomaly Detection) *Академик*. Доступно по: https://dyakonov.org/2017/04/19/поиск-аномалий-anomaly-detection/ (дата обращения: 12.03.2021).
9. Хе К., Чжэн Ю. Дж., Чжан К.Л., Ван Х. Ю., "MTAD-TF: Обнаружение аномалий многомерного временного ряда с использованием комбинации временного шаблона и шаблона признаков". *Complexity*; 2020;2020. DOI: https://doi.org/10.1155/2020/8846608.
10. Ефимов А.И. Методы применения нейронных сетей для оценки и повышения фотореалистичности виртуальной реальности. *ИВД.* 2019;3(54). Доступно по: https://cyberleninka.ru/article/n/metody-primeneniya-neyronnyh-setey-dlya-otsenki-i-povysheniya-fotorealistichnosti-virtualnoy-realnosti (дата обращения: 12.03.2021).

REFERENCES

1. Antonenko S.V. The effectiveness of machine learning-based banking security systems. *Materials of the II International Scientific and Practical Conference "Trends and Prospects for the Development of the Banking System in Modern Economic Conditions" Bryansk, December 17-18, 2019 - Bryansk State University named after Academician I.G. Petrovsky*, 2020; 97-102.
2. Time series (Time series data). *Loginom*. Available at: https://wiki.loginom.ru/articles/time-series.html (accessed 12.03.2021).
3. Rafiqul I., Naznin S., Mohammad Ali M., Prohollad S., Bushra R. A Comprehensive Survey of Time Series Anomaly Detection in Online Social Network Data. *International Journal of Computer Applications*. 2017;180(3):13-22.

4. Dictionary of Economics and Mathematics - Multidimensional time series . *Academic*. Available at: https://economic_mathematics.academic.ru/2615/Многомерные_временные_ряды (accessed 12.03.2021).
5. Van Cuong S., Shcherbakov M.V. A data-driven method for remaining useful life prediction of multiple-component systems. *Caspian J.: Control and High Technologies.* 2019;1:33–44.
6. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. *ACM Computing Surveys.* 2009;41(3).
7. Effective Approaches for Time Series Anomaly Detection . Towards data science. – Mode of access: https://towardsdatascience.com/effective-approaches-for-time-series-anomaly-detection-9485b40077f1 (date of access 12.03.2021).
8. Anomaly Detection . *Academic*. Available at: https://dyakonov.org/2017/04/19/поиск-аномалий-anomaly-detection/ (accessed 12.03.2021).
9. He Q., Zheng Y. J., Zhang C.L., Wang H. Y., "MTAD-TF: Multivariate Time Series Anomaly Detection Using the Combination of Temporal Pattern and Feature Pattern". *Complexity*; 2020;2020. DOI: https://doi.org/10.1155/2020/8846608.
10. Ephimov A.I. Methods for using neural networks to assess and enhance the photorealism of virtual reality. *IVD*. 2019;3(54). Available at: https://cyberleninka.ru/article/n/metody-primeneniya-neyronnyh-setey-dlya-otsenki-i-povysheniya-fotorealistichnosti-virtualnoy-realnosti (accessed 12.03.2021).

ИНОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Раюшкин Эдуард Сергеевич, студент кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: J.Rayushkin@gmail.com
ORCID: 0000-0002-3895-0330

Eduard S. Rayushkin, Student of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation

Максим Владимирович Щербаков, д-р техн.наук, профессор, заведующий кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: maxim.shcherbakov@gmail.com
ORCID: 0000-0001-7173-4499

Maxim V. Shcherbakov, Doctor of Tech.Sciences, Professor, Head of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation

Игорь Дмитриевич Казаков, студент кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: igorkazakov1997@gmail.com
ORCID: 0000-0003-4810-600X

Igor D. Kazakov, Student of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation

Колесникова Вероника Олеговна, студентка кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация
e-mail: verona.7@yandex.ru

Veronika O. Kolesnikova, Student of the CAD Department Volgograd State Technical University, Volgograd, Russian Federation