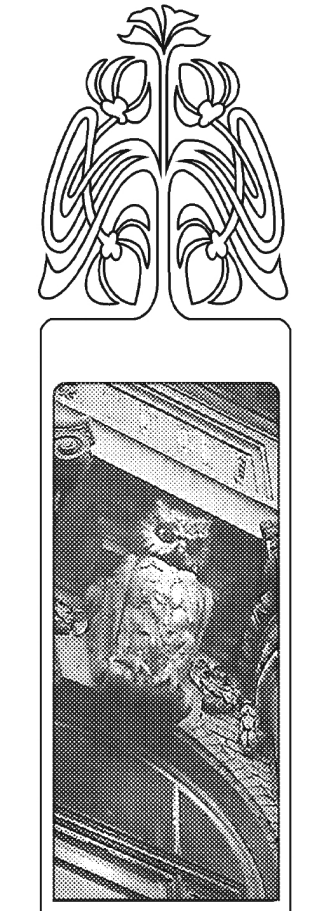


Научный отдел

ИНФОРМАТИКА

Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2025. Т. 25, вып. 4. С. 566–577
Izvestiya of Saratov University. Mathematics. Mechanics. Informatics, 2025, vol. 25, iss. 4, pp. 566–577
https://mmi.sgu.ru
DOI: https://doi.org/10.18500/1816-9791-2025-25-4-566-577
EDN: https://elibrary.ru/TMDBOY

Научная статья
УДК 519.254

Алгоритм обнаружения выбросов в нестационарных временных рядах натурных измерений

В. С. Петракова

Институт вычислительного моделирования СО РАН, Россия, 660036, г. Красноярск, ул. Академгородок, д. 50/44

Петракова Виктория Сергеевна, кандидат физико-математических наук, научный сотрудник, vika-svetlakova@yandex.ru, https://orcid.org/0000-0003-1126-2148, SPIN: 3099-4941, AuthorID: 1182060

Аннотация. Работа посвящена поиску эффективного алгоритма обнаружения выбросов в нестационарных одномерных временных рядах, представляющих собой натурные измерения. Так, нестационарность ряда характеризуется наличием изменчивого тренда в данных, а также гетероскедастичностью — непостоянством дисперсии для отдельно взятых подпоследовательностей временного ряда. Неучет этих особенностей приводит к тому, что выбросы, связанные с поломками или неточностью аппаратуры, фиксирующей натурные измерения, могут быть классифицированы как регулярные значения. Это делает большинство существующих методов обнаружения выбросов во временных рядах неэффективными. В работе описаны реальные данные, представляющие собой наблюдения за температурой и концентрацией загрязнителя в пограничном слое атмосферы г. Красноярска, которые обладают заданными свойствами. Приведен краткий обзор существующих методов, показаны их преимущества и недостатки в применении к имеющимся данным. Предложен авторский подход к обнаружению выбросов в рядах описываемого типа. Представленный в работе метод направлен на коррекцию и объединение существующих подходов и разделен на два этапа: локализация точек, подозрительных на выброс, и регрессия по локализованному участку с адаптивным порогом отсечения точек. Предложенный алгоритм протестирован на имеющихся данных. Проведено сравнение с существующими подходами.

Ключевые слова: анализ временных рядов, обнаружение выбросов, z-балл, нестационарность, регрессия, разработка алгоритма

Благодарности: Работа выполнена при финансовой поддержке Российского научного фонда (проект № 24-71-10022).

Для цитирования: *Петракова В. С.* Алгоритм обнаружения выбросов в нестационарных временных рядах натурных измерений // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2025. Т. 25, вып. 4. С. 566–577. DOI: https://doi.org/10.18500/1816-9791-2025-25-4-566-577, EDN: TMDBOY

Article

Algorithm for searching for outliers in non-stationary time series of field measurements

V. S. Petrakova

Institute of Computational Modelling SB RAS, 55/44 Academgorodok St., Krasnoyarsk 660036, Russia

Viktoriya S. Petrakova, vika-svetlakova@yandex.ru, https://orcid.org/0000-0003-1126-2148, SPIN: 3099-4941, AuthorID: 1182060

Abstract. The paper is devoted to finding an efficient algorithm for detecting outliers in non-stationary one-dimensional time series representing field measurements. Thus, the non-stationarity of a series is characterized by the presence of a variable trend in the data, as well as heteroscedasticity which is the inconstancy of variance for individual subsequences of the time series. Failure to take these features into account leads to the fact that outliers associated with breakdowns or inaccuracies of the equipment recording field measurements can be classified as regular values. This makes most existing methods for detecting outliers in time series ineffective. The paper describes real data representing observations of temperature and pollutant concentration in the boundary layer of the atmosphere in Krasnoyarsk, which have specified properties. A brief overview of existing methods is given, their advantages and disadvantages in application to the available data are shown. The author's approach to detecting outliers in the series of the described type is proposed. The method proposed in the paper is aimed at correcting and combining existing approaches and is divided into two stages: localization of points suspected of being outliers and regression on the localized section with an adaptive threshold for cutting off points. The proposed algorithm was tested on the available data. A comparison with existing approaches was made.

Keywords: time series analysis, outlier detection, z-score, non-stationarity, regression, algorithm development

Acknowledgements: This work was supported by the Russian Science Foundation (project No. 24-71-10022).

For citation: Petrakova V. S. Algorithm for searching for outliers in non-stationary time series of field measurements. *Izvestiya of Saratov University. Mathematics. Mechanics. Informatics*, 2025, vol. 25, iss. 4, pp. 566–577 (in Russian). DOI: https://doi.org/10.18500/1816-9791-2025-25-4-566-577, EDN: TMDBOY

Введение

Временные ряды представляют собой последовательность наблюдений за изменением некоторой переменной x во времени, которые могут быть использованы в разных приложениях, в том числе для понимания и прогнозирования поведения некоторой системы, описываемой переменной. Особым видом временных рядов являются ряды, представляющие собой натурные измерения. Во-первых, достижения в области измерительной техники позволяют собирать большие объемы данных с заданной точностью [1]. Во-вторых, такие ряды, как правило, являются существенно нестационарными, т. е. не сохраняют статистические характеристики подпоследовательностей временного ряда со временем, не имеют единого тренда, а также явно выраженной сезонности. Большое количество наблюдений за переменной (т. е. большой размер выборки) усложняет анализ временных рядов, делая

ряд методов неэффективными [2]. Вместе с тем в том числе из-за требований к точности измерений обнаружение выбросов в таких данных является первым шагом в разведочном анализе данных.

С классической точки зрения широко используемое определение понятия «выброс» было дано в работе [3]. Согласно [3, с. 1], выброс — это «наблюдение, которое настолько отличается от других наблюдений, что вызывает подозрения, что оно было вызвано другим механизмом». Этим определением покрывается большой круг процессов — выброс может быть как просто шумом, так и неожиданной сменой поведения системы (аномалией) или ошибкой измерения, связанной с поломкой оборудования. Методы обнаружения выбросов в данных широко развиваются в настоящее время (см. обзорные работы [4–6]), однако лишь небольшое число из них применимо к анализу временных рядов. Это связано с тем, что временные ряды отличаются от выборок тем, что не могут быть перемешаны, так как позиция каждого измерения во временном ряду имеет значение. В обзорных работах [7, 8], посвященных анализу существующих методов для обнаружения выбросов во временных рядах, выделено порядка 15 методов, сгруппированных по нескольким направлениям. Однако выделенные авторами методы имеют ограничения по применимости к длинным нестационарным временным рядам натурных измерений. Таким образом, разработка новых подходов к поиску выбросов во временных рядах, учитывающих контекст, в котором локализуется выброс, является актуальной задачей.

Статья организована следующим образом. В разд. 1 приведено описание данных, которые используются на протяжении статьи для тестирования и обоснования методов. Раздел 2 посвящен краткому обзору существующих методов поиска выбросов во временных рядах и их применению к имеющимся данным. В разд. 3 приведено описание авторской методики и обсуждение результатов ее применения к имеющимся данным. Наконец, заключительный раздел посвящен описанию результатов исследования.

1. Описание имеющихся данных

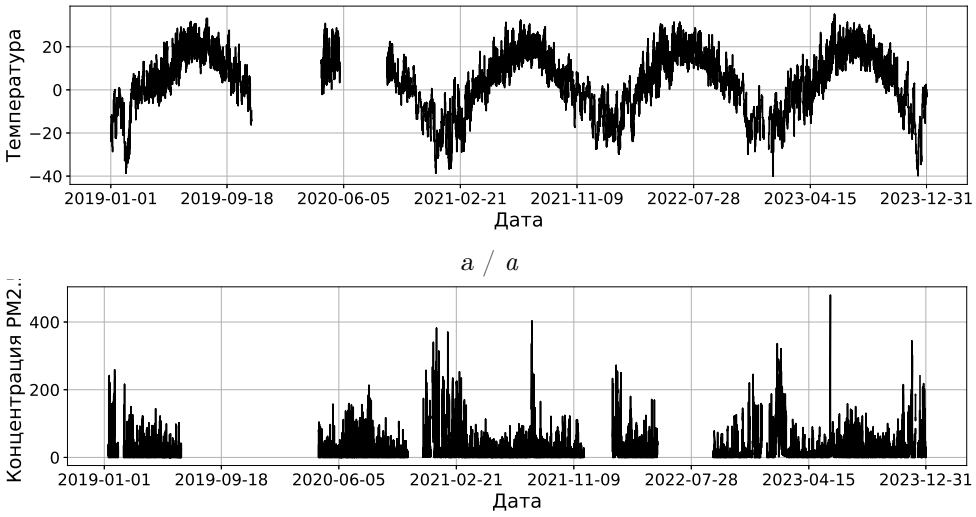
Сложноустроенные нестационарные временные ряды встречаются, в частности, при накоплении наблюдений за окружающей средой, например за качеством атмосферного воздуха. Твердые частицы диаметром 2.5 мк и менее (PM2.5) являются одним из наиболее вредных загрязнителей воздуха в приземном слое атмосферы современных городов и широко признанным маркером качества воздуха. Анализ таких данных позволяет описать динамику смен периодов повышенных концентраций PM2.5 в пограничном слое атмосферы и прогнозировать развитие таких периодов по известным сценариям. Значительное количество текущих исследований сосредоточено на изучении данных о концентрациях PM2.5, собранных сетью пространственно распределенных датчиков, и последующем прогнозировании развития загрязнений (см., например, [9–12]).

Красноярск входит в число российских городов, где качество атмосферного воздуха контролируется на стационарных постах наблюдения по всему городу. В Министерстве экологии и рационального природопользования Красноярского края действует региональная ведомственная информационно-аналитическая система данных о состоянии окружающей среды Красноярского края (КИАС), имеющая девять автоматизированных наблюдательных пунктов (АНП) в черте города. Каждые 20 мин. замеряются температура, влажность, давление и концентрация PM2.5, которые автоматически записываются в набор данных. Собранные данные отражены на геопортале «Система мониторинга воздуха г. Красноярска» (https://air.krasn.ru/map.html?2=). Для контроля концентраций PM2.5 в КИАС используются анализаторы пыли модели E-BAM[1] (Met One Instruments Inc., США), принцип действия которых основан на измерении поглощения β-излучения частицами пыли, осажденными на фильтровальной ленте. Данный метод сертифицирован Агентством по

[1]E-BAM particulate monitor operation manual. Electronic resource. URL: https://metone.com/wp-content/uploads/2022/06/E-BAM-9805-Manual-Rev-G.pdf (дата обращения: 09.01.2025).

охране окружающей среды США[2]. Анализаторы этого класса рекомендованы для измерения содержания фракций PM10 и PM2.5 в атмосфере, сертифицированы и аккредитованы во многих странах мира, в том числе в России (№ 57884-14 в Государственном реестре средств измерений). Описанные датчики имеют особенность — при низких концентрациях PM2.5 дисперсия измерений существенно выше, чем при высоких. Это, помимо разных трендов на каждом участке (рис. 1), приводит к гетероскедастичности полученных числовых наборов значений.

На рис. 1 представлен полный набор данных по температуре и PM2.5 за период с 1 января 2019 г. по 21 декабря 2023 г. с датчика поста «Ветлужанка».



а / *a*

б / *b*

Рис. 1. Измерения температуры (*а*) и концентрации PM2.5 (*б*) в г. Красноярске за период с 1 января 2019 г. по 31 декабря 2023 г.

Fig. 1. Measurements of temperature (*a*) and PM2.5 concentration (*b*) in Krasnoyarsk from January 1, 2019, to December 31, 2023

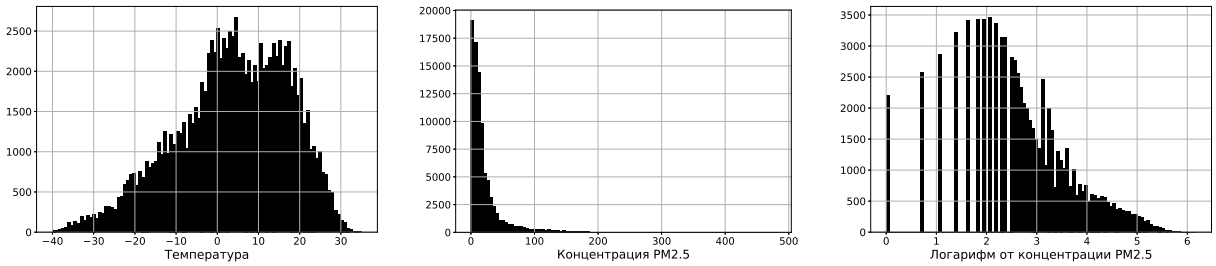
Как видно, данные, представленные на рис. 1, содержат большие пропуски. Также в значениях большое количество недлинных пропусков, которые не видны при визуальном анализе всего набора данных. Также показания по концентрации фиксируют несколько выбросов (значения, превышающие 400), которые могут быть идентифицированы визуально. В работе рассматриваются и данные по температуре, которые имеют понятный квадратичный тренд, сезонность и одинаковую дисперсию для всего набора. Показания по концентрации PM2.5 представляют больший интерес для исследования, но, как видно из рис. 1, *б*, не имеют явно выраженного тренда и сезонности, а также имеют разную дисперсию при низких и высоких концентрациях PM2.5.

2. Обзор существующих методов обнаружения выбросов во временных рядах

Существующие методы обнаружения выбросов во временных рядах могут быть условно разделены на следующие подгруппы: основанные на статистическом описании ряда или его части; на оценке близости значения к соседним значениям; на оценке плотности распределения данных; на кластеризации; на регрессионных методах. Наиболее исчерпывающее сравнение работы разных методов на синтетических данных приведено в работе [7]. Здесь будет обсуждаться применимость существующих методов к описанным в предыдущем разделе данным.

[2]Environmental Technology Verification Report. URL: https://archive.epa.gov/nrmrl/archive-etv/web/pdf/01_vr_metone_bam1020.pdf (дата обращения: 18.08.2023).

Статистические методы основаны на предположении, что выброс — это значение, которое не вписывается в статистическое описание всего временного ряда или его подпоследовательности. Здесь основным подходом является вычисление (модифицированной) z-оценки, квартилей и межквартильного размаха [13]. Квартили и межквартильный размах показывают, какие значения наиболее распространены в выборке, в предположении, что данные распределены нормально. Это требование не выполняется для имеющихся данных, даже если рассматривать отдельно взятые подпоследовательности. Гистограммы распределения временных рядов для измерений температуры и концентрации PM2.5 приведены на рис. 2.



а / *a* *б* / *b* *в* / *c*

Рис. 2. Гистограммы распределения данных измерений: *а* — температура; *б* — концентрация PM2.5; *в* — десятичный логарифм от измерений концентрации PM2.5

Fig. 2. Histograms of measurement data distribution: *a* is temperature; *b* is PM2.5 concentration; *c* is decimal logarithm of PM2.5 concentration measurements

Таким образом, данные не имеют нормального распределения, более того, гистограммы распределений исследуемых временных рядов мультимодальны. Это делает критерий, основанный на расчете квартилей, непригодным для анализа имеющихся данных.

Расчет z-оценки [14] здесь является более подходящим инструментом. Под z-оценкой понимают меру относительного разброса наблюдаемого значения, которая показывает, насколько отличается разброс значений от выбранной меры центральной тенденции. Различают стандартную ($Z_{mean}(x_i)$) и модифицированную ($Z_{med}(x_i)$) z-оценку для значения x_i временного ряда (или его подпоследовательности) $X_t = \{x_1, x_2, ..., x_N\}$:

$$Z_{mean}(x_i) = \left|\frac{x_i - \mu}{\sigma}\right|, \quad Z_{med}(x_i) = \left|\frac{x_i - \text{median}(X_t)}{1.486 \cdot MAD}\right|,$$

где для X_t определены: μ — выборочное среднее, σ — среднеквадратичное отклонение, $\text{median}(X_t)$ — медиана, MAD — медианное абсолютное отклонение. Модифицированная z-оценка более устойчива в сравнении со стандартной z-оценкой. Значение x_i считается выбросом, если z-оценка больше некоторого заданного порога, чаще всего выбираемого равным 3 (для нормально распределенных данных). Для сложноустроенных временных рядов не имеет смысла строить z-оценку для всего ряда X_t, а только для его подпоследовательности заданной длины. Такой подход называется MZS (*Moving z-score*). Здесь для принятия решения по точке x_i вычисляется z-оценка по последовательности $\{x_{i-w}, ..., w_{i-w+1}, ..., x_{i-1}\}$, где w — заранее заданная величина окна. Результаты применения стандартной и модифицированной z-оценки для некоторых подпоследовательностей исследуемых данных с $w = 10$ приведены на рис. 3, 4. Здесь и далее использование разных цветов для выделения точек, определяемых как выброс, не является характеристикой метода, а используется только для визуального контрастирования результата применения разных методов.

Так, z-оценка хорошо справляется с определением единичных, явно выбивающихся из последовательностей значений (рис. 3, *д*–*з*), при этом модифицированная оценка определяет гораздо больше точек, подозрительных на выброс. Часть из этих точек действительно является пропущенной, как, например, точки в районе 9–11 утра на рис. 3, *в*, *г*, а часть из них — как на рис. 3, *д*, *е*, являются ложно принимаемыми.

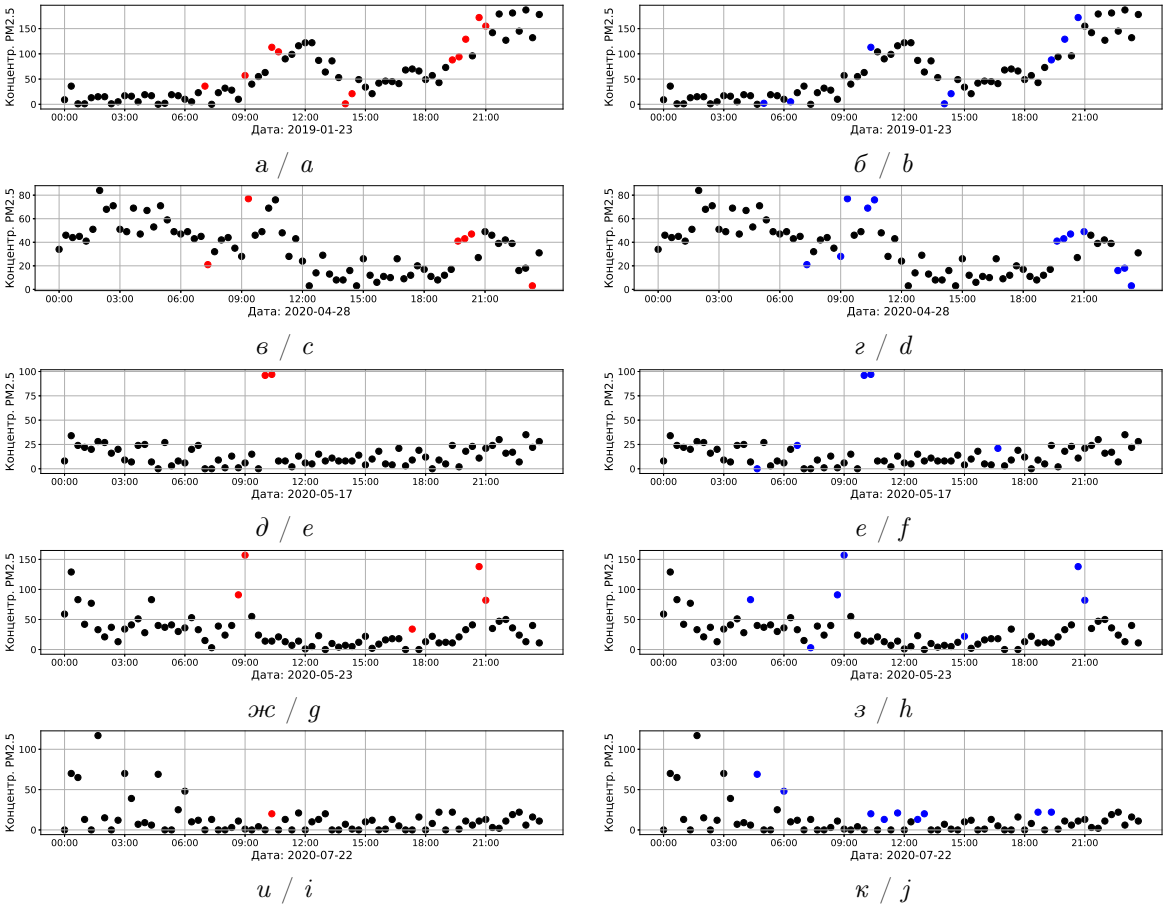


Рис. 3. Результат применения z-оценки к некоторым подпоследовательностям временного ряда по показаниям концентрации PM2.5: *а*, *в*, *д*, *ж*, *и* — результаты применения стандартной z-оценки; *б*, *г*, *е*, *з*, *к* — результаты применения модифицированной z-оценки (цвет онлайн)

Fig. 3. The result of applying the z-score to some subsequences of the time series based on the PM2.5 concentration readings: *a*, *c*, *e*, *g*, *i* are results of applying the standard z-score; *b*, *d*, *f*, *h*, *j* are results of applying the modified z-score (color online)

Результат получается хуже, если данные обладают меньшей дисперсией, но частыми сменами тренда, как данные по температуре (см. рис. 4).

Поскольку статистические характеристики подпоследовательности меняются при смене тренда, то точки разладки (точки смены тренда) определяются как выбросы. Все рисунки для каждого набора данных доступны по ссылке: https://colab.research.google.com/drive/1IIrw6Jp3X3QSwcCey_NumvRxBaWr7i9O?usp=sharing

Методы обнаружения выбросов, основанные на оценки близости значения к ближайшим соседям [15], основаны на правиле, что расстояние между x_i и группой ее ближайших соседей $x_{i-w}, x_{i-1}, x_{i+1}, , x_{i+w}$ значений меньше заданного порога d. Проблема такого класса методов заключается в том, что если в данных есть быстрорастущий тренд, то такие методы становятся неэффективными.

Методы, основанные на плотности распределения данных [16], заключаются в том, что для временного ряда (или его подпоследовательности) X_t строится гистограмма, а значения, определяемые как выбросы, находятся в хвосте гистограммы. Такая группа методов не подходит для исследуемых данных, поскольку данные имеют мультимодальное распределение; т. е. сначала должны быть разделены на классы, соответствующие каждой моде. Эта задача является сложной и практически не описана в литературе, за исключением небольшого числа работ (см., например, [17,18]), а также требует предварительного знания о распределении в каждом классе.

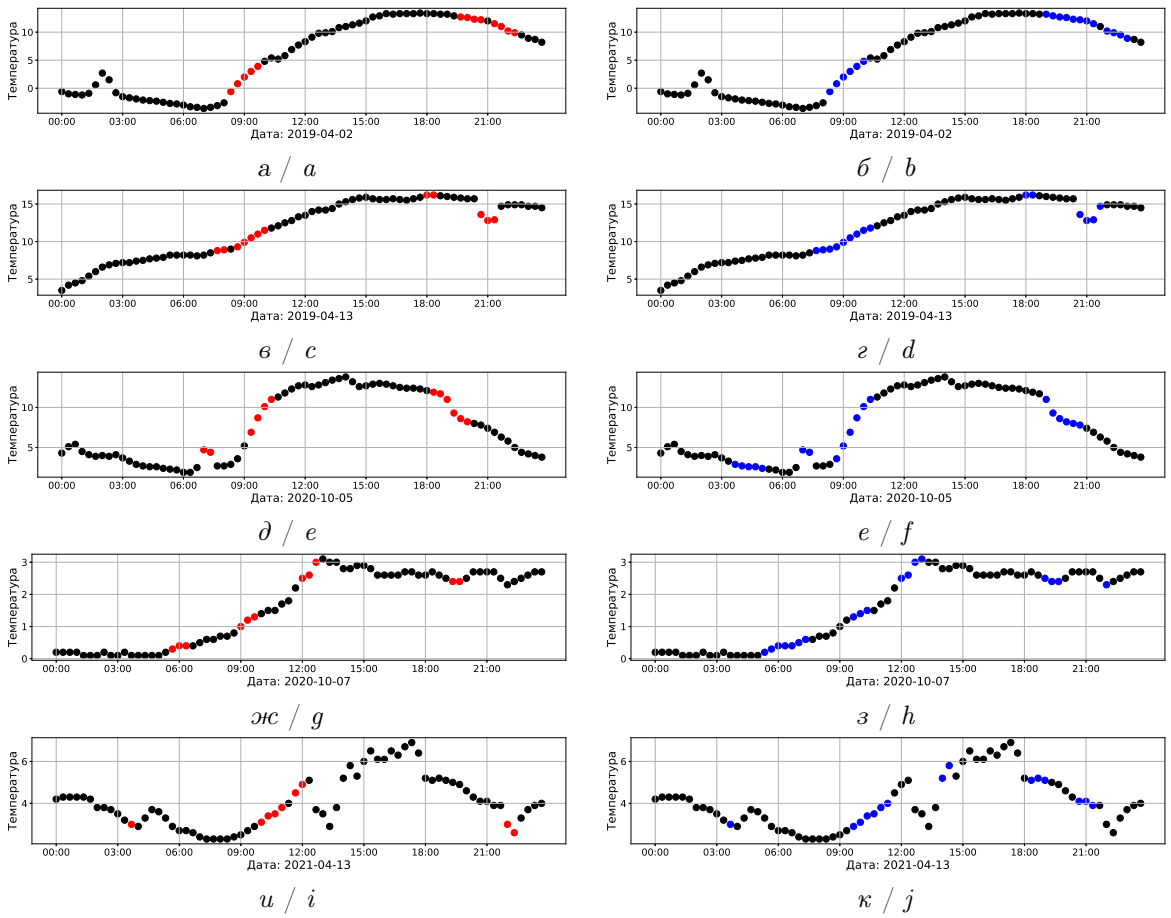


Рис. 4. Результат применения z-оценки к некоторым подпоследовательностям временного ряда по показаниям температуры: а, в, д, ж, и — результаты применения стандартной z-оценки; б, г, е, з, к — результаты применения модифицированной z-оценки (цвет онлайн)

Fig. 4. The result of applying the z-score to some subsequences of the time series of temperature readings: a, c, e, g, i are results of applying the standard z-score; b, d, f, h, j are results of applying the modified z-score (color online)

Отметим, что метод, как и в предыдущих случаях, может быть применен к отдельно взятым подпоследовательностям. Но здесь подразумевается, что подпоследовательности должны быть большого размера, а значит, также отражают общую мультимодальность данных.

Последняя группа регрессионных методов является наиболее перспективной в применении к исследуемым данным. Регрессионные методы основываются на определении выброса как точки, которая значительно отклоняется от своего ожидаемого значения. Таким образом, при наличии одномерного временного ряда точка в i-й момент времени может быть объявлена выбросом, если расстояние до ее ожидаемого значения превышает предопределенный порог τ. Так, если ожидаемое значение для x_i определяется как $\hat{x}_i$ и $|x_i - \hat{x}_i| > \tau$, то (t_i, x_i) является выбросом.

Ожидаемое значение $\hat{x}_i$ определяется по регрессионной модели, построенной по подпоследовательности временного ряда, включающей точку x_i. Основные трудности связаны с необходимостью предварительного знания о структуре тренда, определением значения порога τ и вычислениями для больших наборов данных. Проблема автоматического поиска структуры тренда может быть решена выбором полиномиальной аппроксимации высокого порядка для небольшого числа точек. Проблема выбора порога τ — более сложная, поскольку при гетероскедастичных данных значение порога должно определяться на каждом наборе точек самостоятельно, так как одна регрессионная кривая может объединить участки

с разной дисперсией. На рис. 5 представлен результат работы регрессионного алгоритма для тех же подпоследовательностей временного ряда, что и на рис. 3, 4. В качестве порога были выбраны значения $\tau = 0.7$ для температуры и $\tau = 20$ для концентрации PM2.5. Регрессионная модель строилась по 12 точкам.

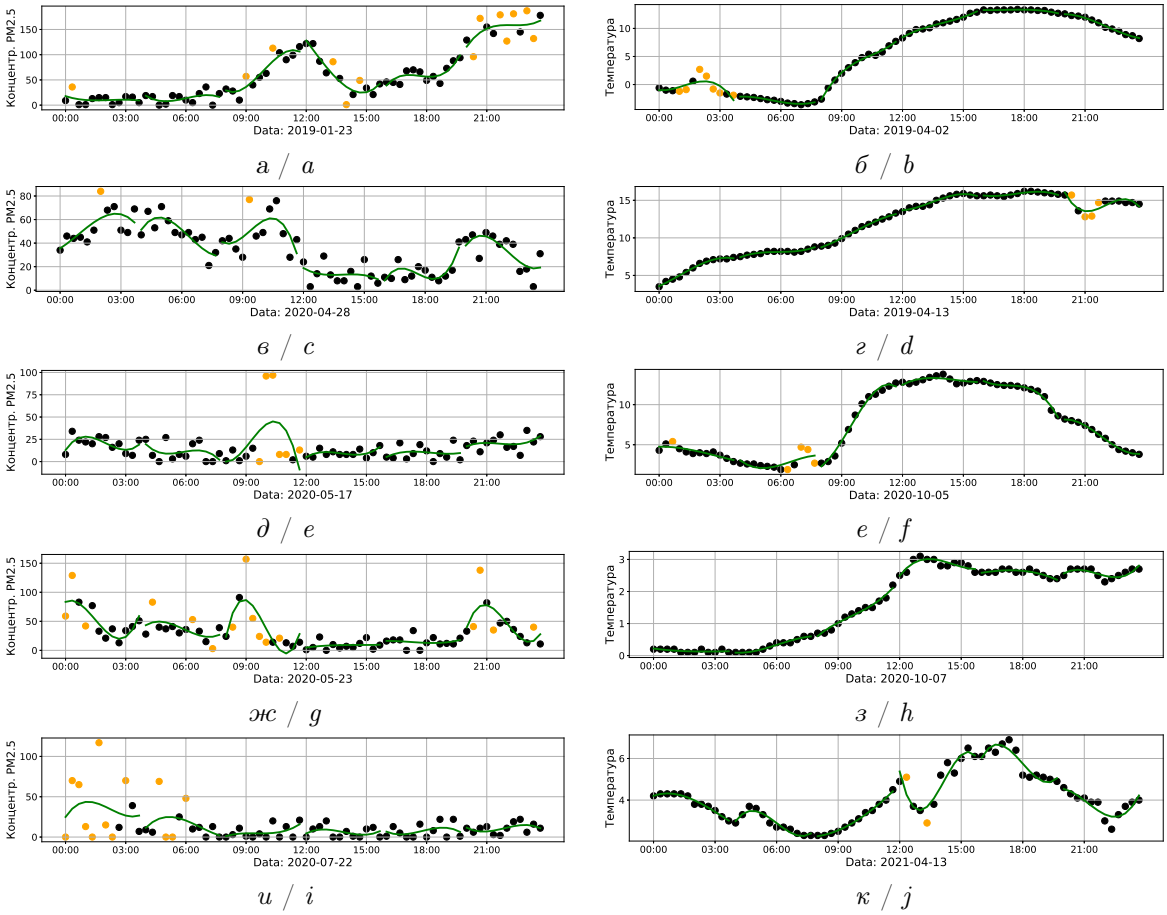


Рис. 5. Результат применения регрессионной кубической полиномиальной модели для оценки выбросов: *а*, *в*, *д*, *ж*, *и* — в данных по концентрации PM; *б*, *г*, *е*, *з*, *к* — в данных измерений температуры (цвет онлайн)

Fig. 5. The result of applying a regression cubic polynomial model to estimate outliers: *a*, *c*, *e*, *g*, *i* are in the PM concentration data; *b*, *d*, *f*, *h*, *j* are in the temperature measurement data (color online)

Таким образом, регрессионный подход более точен в определении аномальных значений, чем описанные выше, относительно исследуемых данных. Однако, как было указано выше, такой метод может выдавать ошибки в случае гетероскедастичных данных (см., рис. 5, *д*, *ж*, *и*). Метод не точен, если совпадают точки смены регрессионной модели и тренда в данных, тогда крайние точки подпоследовательности определяются как выброс (см. рис. 5, *к*), необходимость строить модель высокого порядка на каждом небольшом наборе точек делает подход вычислительно- и времязатратным.

3. Частично регрессионный метод поиска выбросов в сложноустроенных временных рядах

Предлагаемый здесь подход направлен на устранение недостатков методов, основанных на построении регрессии и оценке расстояния до ближайших соседей. Методы, основанные на расчете расстояний к рядом лежащим точкам, позволяют учесть основное свойство ряда — связанность его значений между собой, а регрессионые — оценить отклонение значения от наблюдаемого в данных тренда. Предлагаемый метод заключается в том, чтобы сначала

найти места скопления точек, подозрительных на выброс, а затем их обработать регрессионной моделью. Наличие предобработки позволит снизить вычислительную сложность регрессионного метода и устранить влияние гетероскедастичности, наблюдаемой в данных, на результат отнесения точки к выбросу. Таким образом, предлагаемый алгоритм состоит из двух этапов:

1) предобработка;

2) регрессионная модель с адаптивным условием определения выброса.

3.1. Первый этап. Предобработка

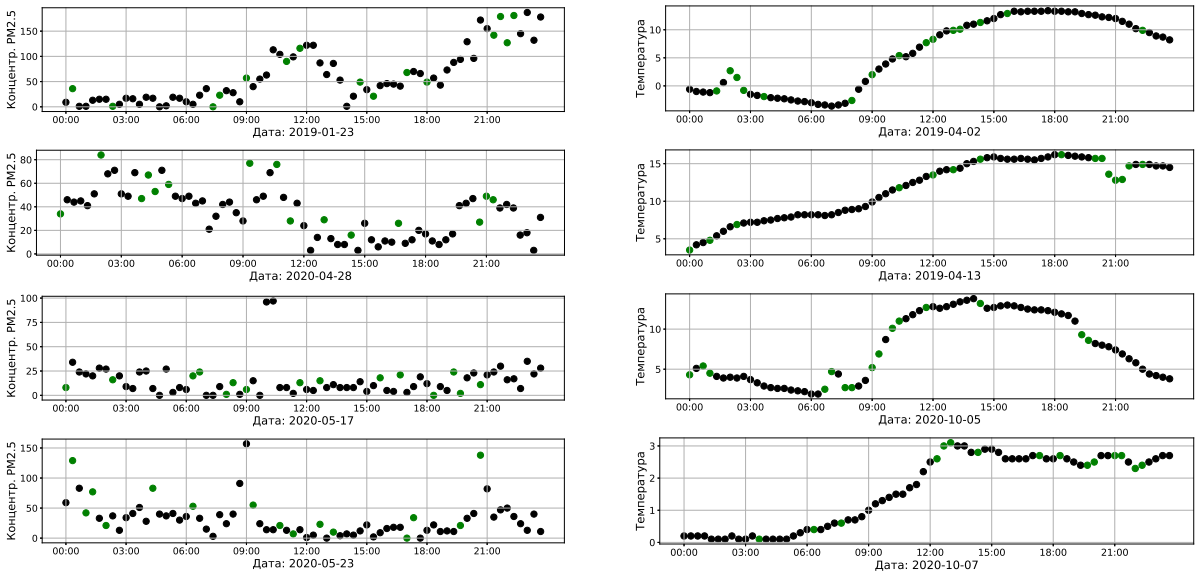
Предобработка заключается в том, чтобы найти точки, значение которых являются подозрительными на выброс. С этой целью для каждой точки x_i проводится две прямых. Положим, что через точки (t_i, x_i), (t_{i+2}, x_{i+2}) проходит прямая, заданная уравнением

$$y_{i,1} = k_{i,1}x + b_{i,1}, \tag{1}$$

а через точки (t_{i+1}, x_{i+1}), (t_{i+3}, x_{i+3}) — прямая

$$y_{i,2} = k_{i,2}x + b_{i,2}. \tag{2}$$

После этого для точек (t_i, x_i), (t_{i+2}, x_{i+2}) считаются расстояния d_i, d_{i+2} до прямой, заданной уравнением (2), а для точек (t_{i+1}, x_{i+1}), (t_{i+3}, x_{i+3}) — до прямой, заданной уравнением (1). Таким образом, за один проход по временному ряду X_t вычислением 8 констант получен набор расстояний $D = d_i$. Эти расстояния позволяют оценить, насколько значение в точке отличается от ее ближайших соседей, при этом учитывая факт, что в данных может быть непостоянный тренд. Теперь для отсечения набора времен t_i, рядом с которыми есть значения, подозрительные на выброс или соседние с ними, необходимо выбрать пороговое значение d_T, такое, что если $d_i > d_T$, то рядом с t_i может находиться выброс. В качестве D_T можно выбрать некоторое значение процентили для набора D. Так, на рис. 6 выделены значения, которые являются подозрительными на выброс или соседствуют с ними для исследуемых данных по температуре и концентрации PM2.5. В качестве порога отсечения использовалось значение d_T, равное 80-й процентили (q_{80}) набора D.



а / *a* *б* / *b*

Рис. 6. Результат поиска точек, подозрительных на выброс, в данных по концентрации PM (a) и измерений температуры ($б$) (цвет онлайн)

Fig. 6. The result of searching for outlier points in PM concentration data (a) and temperature measurements (b) (color online)

3.2. Второй этап. Регрессия и отсечение

Второй этап заключается в том, что для каждого скопления точек, подозрительных на выброс, строится регрессионная модель третьего порядка для заданного набора значений и ближайших к скоплению точек. Так, положим, точка (t_j, x_j) определена на предыдущем этапе как подозрительная на выброс. Если в диапазоне (t_{j-w}, t_{j+w}) находятся еще точки, подозрительные на выброс, то рассчитывается величина j' — медианное положение точки, подозрительной на выброс. Далее по $2w$ точкам, лежащим в диапазоне $(t_{j'-w}, t_{j'+w})$, строится полиномиальная регрессия третьего порядка:

$$y_{j',k}(x) = a_{j',k}x^3 + b_{j',k}x^2 + c_{j',k}x + d_{j',k}.$$

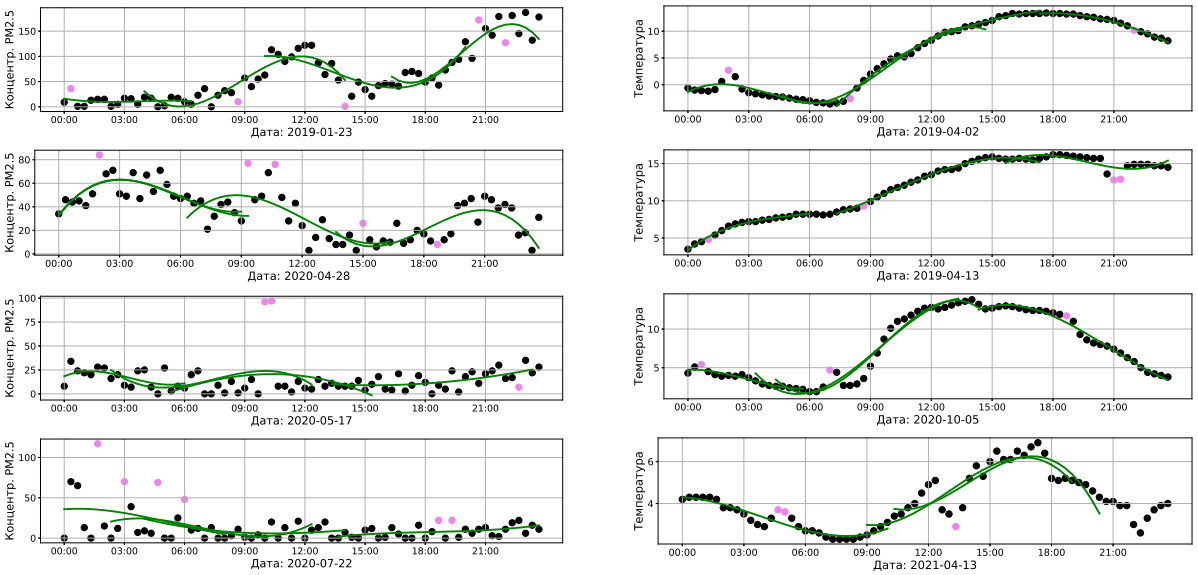
Величина $w \in Z$ является гиперпараметром алгоритма и задается пользователем. Решение, является ли точка (t_j, x_j), где $t_j \in (t_{j'-w}, t_{j'+w})$, выбросом, принимается при выполнении одного из двух (или их совокупности) условий:

$$Z_{med}(|y_{j',k}(t_j) - x_j|) > Z_T, \tag{3}$$

$$Z_{med}(dist((t_j, x_j), y_{j',k}(x))) > Z_T. \tag{4}$$

Так, условие (3) отсекает те точки (t_j, x_j), которые далеки от прогнозируемого значения, полученного по построенной регрессии, а условие (4) — точки, которые далеки от построенной регрессионной кривой. Использование z-оценки в качестве условия позволяет снизить чувствительность метода к гетероскедастичности данных.

Условие (3) является более простым в реализации, а (4) увеличивает точность, но требует решения задачи минимизации. Результаты применения алгоритма к исследуемым данным с гиперпараметрами $d_T = q_{80}$, $w = 15$, $z_T = 1.5$ представлены на рис. 7.



а / *a* *б* / *b*

Рис. 7. Результат применения предложенного алгоритма для оценки выбросов в данных по концентрации PM (*а*) и измерений температуры (*б*) (цвет онлайн)
Fig. 7. The result of applying the proposed algorithm to estimate emissions in PM concentration data (*a*) and temperature measurements (*b*) (color online)

Таким образом, предложенный алгоритм делает меньшее количество ложно положительных классификаций по сравнению с описанными выше алгоритмами для гетероскедастичных данных (см., например, рис. 7, *а*). Однако, как и другие методы, возможно ложно

отрицательное срабатывание (см., например, последний рисунок из рис. 7, *а*: крайние левые точки не были классифицированы как выбросы). Эта проблема может быть решена увеличением чувствительности алгоритма к выбросам (уменьшении значения Z_T), однако это приведет к увеличению ложно положительных срабатываний.

Выводы

Работа посвящена поиску эффективных алгоритмов обнаружения выбросов в длинных сложно устроенных временных рядах с частой сменой тренда и проявлением гетероскедастичности в данных. Рассмотрены классы существующих алгоритмов, для которых определены их преимущества и недостатки для имеющихся реальных временных рядов натурных измерений. Показано, что методы, основанные на статистическом описании временного ряда, распознают точки разладки (точки смены тренда) как выбросы, что является ложно положительным срабатыванием. Методы, основанные на анализе гистограмм, не подходят для мультимодально распределенных данных, более того, требуют предварительного знания или предположения о распределении временного ряда. Методы, основанные на сравнении расстояний между ближайшими точками, плохо работают для рядов с быстро растущим трендом. Регрессионные методы вычислительно сложны и допускают большее число ложно положительных и ложно отрицательных срабатываний для гетероскедастичных данных.

Предложенный в работе метод направлен на коррекцию и объединение известных подходов. Метод разделен на два этапа: сначала происходит предобработка, в результате которой локализуются элементы ряда, поблизости с которыми могут быть выбросы, затем локально применяется регрессионная модель. Порог определения точек на выброс является адаптивным и зависит от локализации точки, что позволяет ему учитывать смены тренда и гетероскедастичность.

Недостатки предложенного алгоритма заключаются, в первую очередь, в большом количестве гиперпараметров (три параметра), к определению одного из которых (параметр Z_T) алгоритм крайне чувствителен. Также алгоритм является вычислительно затратным при большом числе точек, подозрительных на выброс. Так, если количество таких точек больше, чем N/w, то алгоритм за счет препроцессинга является более вычислительно- и времязатратным, чем регрессионный.

Полный набор результатов для всего набора реальных данных, описанных в разд. 1, можно найти по ссылке: https://colab.research.google.com/drive/1IIrw6Jp3X3QSwcCey_NumvRxBaWr7i9O?usp=sharing.

Список литературы / References

1. Bezmenov I. V., Drozdov A. E., Pasynok S. L. A strategy for finding outliers in noisy data series including an unknown trend. *Measurement Techniques*, 2022, vol. 65, iss. 5, pp. 339–345. DOI: https://doi.org/10.1007/s11018-022-02085-6
2. Fan J., Han F., Liu H. Challenges of big data analysis. *National Science Review*, 2014, vol. 1, iss. 2, pp. 293–314. DOI: https://doi.org/10.1093/nsr/nwt032
3. Hawkins D. M. *Identification of outliers*. Monographs on Statistics and Applied Probability. New York, Springer Netherlands, 1980. 188 p. DOI: https://doi.org/10.1007/978-94-015-3994-4
4. Kiani R., Jin W., Sheng V. S. Survey on extreme learning machines for outlier detection. *Machine Learning*, 2024, vol. 113, pp. 5495–5531. DOI: https://doi.org/10.1007/s10994-023-06375-0
5. Rhyu J., Bozinovski D., Dubs A. B., Mohan N., Cummings Bende E. M., Maloney A. J., Nieves M., Sangerman J., Lu A. E., Hong M. S., Artamonova A., Ou R. W., Barone P. W., Leung J. C., Wolfrum J. M., Sinskey A. J., Springs S. L., Braatz R. D. Automated outlier detection and estimation of missing data. *Computers & Chemical Engineering*, 2024, vol. 180, art. 108448. DOI: https://doi.org/10.1016/j.compchemeng.2023.108448
6. Hu R., Chen L., Wang Y. *An efficient outlier detection algorithm for data streaming*. arXiv:2501.01061 [stat] January 2, 2025. 12 p. DOI: https://doi.org/10.48550/arXiv.2501.01061
7. Alimohammadi H., Shengnan N. Ch. Performance evaluation of outlier detection techniques in

production timeseries: A systematic review and meta-analysis. *Expert Systems with Applications*, 2022, vol. 191, art. 116371. DOI: https://doi.org/10.1016/j.eswa.2021.116371
8. Blázquez-García A., Conde A., Mori U., Lozano J. A. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 2021, vol. 54, iss. 3, pp. 1–33. DOI: https://doi.org/10.1145/3444690
9. Zhao N., Liu Y., Vanos J. K., Cao G. Day-of-week and seasonal patterns of PM2.5 concentrations over the United States: Time-series analyses using the Prophet procedure. *Atmospheric Environment*, 2018, vol. 192, pp. 116–127. DOI: https://doi.org/10.1016/j.atmosenv.2018.08.050
10. Zhai B., Chen J., Yin W., Huang Z. Relevance analysis on the variety characteristics of PM2.5 concentrations in Beijing, China. *Sustainability*, 2018, vol. 10, iss. 9, art. 3228. DOI: https://doi.org/10.3390/su10093228
11. Liu B., Yan S., Li J., Li Y., Lang J., Qu G. A spatiotemporal recurrent neural network for prediction of atmospheric PM2.5: A case study of Beijing. *IEEE Transactions on Computational Social Systems*, 2021, vol. 8, iss. 3, pp. 578–588. DOI: https://doi.org/10.1109/TCSS.2021.3056410
12. Wang P., Zhang H., Qin Z., Zhang G. A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting. *Atmospheric Pollution Research*, 2017, vol. 8, iss. 5, pp. 850–860. DOI: https://doi.org/10.1016/j.apr.2017.01.003
13. Senthamarai Kannan K., Manoj S. K., Arumugam S. Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, 2015, vol. 10, iss. 2, pp. 231–238.
14. Mare D. S., Moreira F., Rossi R. Nonstationary Z-Score measures. *European Journal of Operational Research*, 2017, vol. 260, iss. 1, pp. 348–358. DOI: https://doi.org/10.1016/j.ejor.2016.12.001
15. Wang H., Bah M. J., Hammad M. Progress in outlier detection techniques: A survey. *IEEE Access*, 2019, vol. 7, pp. 107964–108000. DOI: https://doi.org/10.1109/ACCESS.2019.2932769
16. Tang B., He H. A local density-based approach for outlier detection. *Neurocomputing*, 2017, vol. 241, pp. 171–180. DOI: https://doi.org/10.1016/j.neucom.2017.02.039
17. Boulmerka A., Allili M. S., Ait-Aoudia S. A generalized multiclass histogram thresholding approach based on mixture modelling. *Pattern Recognition*, 2014, vol. 47, iss. 3, pp. 1330–1348. DOI: https://doi.org/10.1016/j.patcog.2013.09.004
18. Karpatne A., Khandelwal A., Kumar V. Ensemble learning methods for binary classification with multi-modality within the classes. *Proceedings of the 2015 SIAM International Conference on Data Mining*, 2015, pp. 730–738. DOI: https://doi.org/10.1137/1.9781611974010.82

Поступила в редакцию / Received 09.01.2025
Принята к публикации / Accepted 22.04.2025
Опубликована / Published 28.11.2025