ВЕСТНИК
ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
МОРСКОГО И РЕЧНОГО ФЛОТА ИМЕНИ АДМИРАЛА С. О. МАКАРОВА

DOI: 10.21821/2309-5180-2017-9-4-866-873

METHODS OF DETECTION OF NOISES AND OUTLIERS IN THE DATA STRUCTURE ON THE EXAMPLE OF SHIP HANDLING DURATION AT AN OIL TERMINAL

Yu. Yu. Rudnitckaia

Brno University of Technology, Czech Republic

Pre-processing of input information is essential part of any analysis. Data missing or the presence of extremes can significantly distort the results and lead to an erroneous solution of problems that are not in accordance with the real situation. Despite extensive research, finding and processing data anomalies is still an area of increased interest. For solving these issues, the results of using of statistical-mathematical and intelligent methods will be reflected in the paper.

This paper focuses on the time series, expressed in number of hours of ship handling on an oil terminal. Such insights as noise and outliers in data, and their causes will be considered. Moreover, it will be proposed detection and processing methods for outliers, and the best method will be chosen (based on comparative analysis of confidence intervals)

In practice, common and well-known methods, such as calculations based on mean value (expected value) or quartiles, cause the loss of valuable information, which further significantly affects the correctness of the overall results. The article points to the importance of preliminary estimation of the data structure. In the case of an asymmetric distribution, use calculations based on the median value is more effective. As an alternative for comparing, the result of data mining method (clustering) with its shortcomings and advantages is shown.

It should also be noted that further analysis of data outliers and the identification of existing dependencies between them lead to improvements of a whole process. Such objects should be considered as unique cases and their study can improve service delivery in the seaport.

Keywords: anomaly detection, descriptive statistics, data structure, data processing, outliers, port processes, ship handling, oil terminal.

For citation:
Rudnitckaia, Yuliya Yu. "Methods of detection of noises and outliers in the data structure on the example of ship handling duration at an oil terminal." *Vestnik Gosudarstvennogo universiteta morskogo i rechnogo flota imeni admirala S.O. Makarova* 9.4 (2017): 866–873. DOI: 10.21821/2309-5180-2017-9-4-866-873.

УДК 656.60.009.02

МЕТОДЫ ОПРЕДЕЛЕНИЯ ШУМОВ И ВЫБРОСОВ В СТРУКТУРЕ ИСПОЛЬЗУЕМЫХ ДАННЫХ НА ПРИМЕРЕ ДЛИТЕЛЬНОСТИ ОБРАБОТКИ СУДОЗАХОДА НА НЕФТЯНОМ ТЕРМИНАЛЕ

Ю. Ю. Рудницкая

Технический университет, г. Брно, Чешская Республика

Важным этапом любого анализа является предварительная обработка входной информации. Отсутствие данных или наличие экстремумов может значительно исказить результаты и привести к ошибочному решению проблем, несоответствующих реальной ситуации. Несмотря на многочисленные исследования, поиск и обработка аномалий в используемых данных до сих пор является областью повышенного интереса. Для решения этой проблемы в статье отражены результаты применения интеллектуальных и статистическо-математических методов.

Объектом для исследования является временной ряд, выраженный в количестве часов обработки судозахода на нефтяном терминале. Рассмотрены такие понятия, как шумы и выбросы в структуре исходных данных, и причины их возникновения. В статье предложены методы обнаружения и способы обработки выбросов, а также выбран оптимальный вариант для работы с данным временным рядом на основе

сравнительного анализа доверительных интервалов, полученных в результате применения рассматриваемых методов.

На практике универсальные и общеизвестные методы, такие как расчеты на основе среднего значения (математическое ожидание для генеральной совокупности) или квартилей, приводят к потере ценной информации, что в дальнейшем отражается на корректности результатов всего анализа. В статье рассмотрена проблема предварительной оценки структуры данных и необходимость использования расчетов на основе значения медианы в случае асимметричности ряда распределения величин. В качестве альтернативы показан результат метода анализа данных (кластеризации), его недостатки и преимущества.

Следует отметить, что определение выбросов в структуре используемых данных принесет большую пользу при их последующем анализе и выявлении существующих закономерностей между ними. Судозаходы, которые представляют собой выбросы, нужно рассматривать как уникальные случаи, удаление которых из общей выборки не приведет к дальнейшему улучшению процесса обработки танкерного флота.

Ключевые слова: временной ряд, анализ выбросов, описательная статистика, нефтяной терминал, портовые процессы, обработка судна.

Для цитирования:
*Рудницкая Ю. Ю.*Методы определения шумов и выбросов в данных на примере длительности обработки судозахода на нефтяном терминале / Ю. Ю. Рудницкая // Вестник Государственного университета морского и речного флота имени адмирала С. О. Макарова. — 2017. — Т. 9. — № 4. — С. 866–873. DOI: 10.21821/2309-5180-2017-9-4-866-873.

Введение

Использование современных алгоритмов, таких областей анализа как Data Mining, Machine learning и Process Mining, открывает ряд новых возможностей для улучшения количественных и качественных показателей портовых процессов. Исследование модели процесса «*как есть*» отслеживания потоков работ в режиме online, определение проблемных мест, перегруженных ресурсов, принятие решений и предсказание поведения системы — вот неполный перечень проблем, решаемых в данных областях [1]. Однако исследования должны базироваться на качественных входных данных. В противном случае ни один из существующих алгоритмов не справится с обработкой или приведет к бессмысленным результатам.

На сегодняшний день в России в портах введены информационные системы, отвечающие, в том числе, за проверку качества вводимых данных. Несмотря на это, ошибок в хранилищах данных не избежать, поэтому необходимо заранее определить рабочую группу анализируемых данных и оценить их возможные выбросы.

На основе визуального и сравнительного анализа определен эффективный метод определения шумов и выбросов в структуре данных. Основные работы, связанные с применением методов анализа данных, чаще всего относятся к области медицины и финансов [2] – [5]. Вопрос использования подобных средств анализа в области морской транспортной логистики, а именно ее составляющей части — обработки судна на терминале — является актуальным и до конца неизученным.

Методы и материалы

1. Временной ряд и аномалии в структуре данных

Объектом дальнейшего анализа выступает временной ряд, т. е. собранные в различные моменты времени значения каких-либо параметров исследуемого процесса. В данном случае параметром служит длительность обработки судозахода на нефтяном терминале. Временной ряд, как правило, состоит из следующих элементов: длительность обработки судна $T = \{t_1, t_2, \dots, t_n\}$ и числовое значение показателя (номер судозахода): $X = X(T) = \{x_1, x_2, \dots, x_n\}$ [6], где n — количество исследуемых судозаходов (в течение рассматриваемого периода зафиксировано 2657 судозаходов). Для того, чтобы визуально отразить качество применяемых далее методов, необходимо построить диаграмму зависимости количества судозаходов от длительности их обработки, выраженную в виде следующей формулы:

$$X(T) = \sum_{i=1}^{n} x_i(t), \quad (1)$$

где $X(T)$ — функция зависимости количества судозаходов от длительности их обработки; $\tau \in [5{,}5;\ 671{,}5]$ — длительность обработки судозахода (за интервал взят промежуток отминимального времени обработки до максимального включительно); x_i — отдельно взятый судозаход.

Таким образом, получаем диаграмму, отражающую количество судозаходов и длительность их обработки, необходимую для дальнейших расчетов (рис. 1 и 2). Для наглядности временные периоды указаны в соответствии с шагом 2,7 %, т. е. 18 ч:

$$\left(t_{\max} - t_{\min}\right) \cdot 0{,}027 = \left(671{,}5 - 5{,}5\right) \cdot 0{,}27 \approx 18 \text{ ч}. \quad (2)$$

Одним из основных факторов, оказывающих значительное влияние на результаты анализа данных, является *зашумленность*. Она может быть вызвана как ошибкой оператора и системы, так и ошибкой переноса данных на электронный носитель, погрешностью вычислений и даже причиной природного характера. В научной литературе определения понятий «шум» (noise) и «выброс» (outlier) различаются в зависимости от источника. Как правило, под шумом понимается ошибка в данных, определяемая как выражение «первоначальное значение ± сигнал». Самым тривиальным методом обработки шума служит удаление объекта, который является шумом, из общей выборки. Более разумные способы предлагают алгоритмы Data Mining, которые хорошо справляются с обработкой шумов — от замены шума на среднее или наиболее часто встречающееся значение до прогнозирования вероятного значения на основе остальных параметров объекта.

В свою очередь, выбросы представляют собой различные аномалии экстремумы в процессе, требующие дополнительного анализа. Изучение подобных выбросов может привести к значительному улучшению показателей эффективности всего процесса и повысить качество предоставляемых услуг. Важно отметить, что иногда объект можно отнести как к выбросу, так и к шуму одновременно. Например, судозаход с временем обработки 670 ч может являться частным случаем (выбросом) или ошибкой в базе данных (шумом). Для определения отклонений во временном ряде и выявления его структуры далее изложены основные математико-статистические методы анализа.

2. Статистические методы определение временных выбросов

Для определения выбросов временного ряда необходимо вычислить доверительный интервал, т. е. предельные значения статистической величины, которая с заданной доверительной вероятностью будет находиться в этом интервале при выборке большего объема. Соответственно все значения, лежащие вне интервала, будут относиться к выбросам. Были выбраны следующие методы определения интервала и их параметры (более подробно каждый метод описан в источниках [7] – [9]:

1. Подсчет математического ожидания(μ = 107,49) и стандартного отклонения (σ = 73,06). Эмпирическое правило при ассиметричном распределении ряда данных — правило Бьенамэ–Чебышева.

2. Использование медианы (Q_2= 90,67) как основы для расчетов, не подверженных влиянию временных экстремумов.

3. Использование перцентилей / квартилей и межквартильного интервала [10]. Вычислим межквартильный интервал по формуле

$$IQR = Q_3 - Q_1 = 141{,}67 - 55{,}67 = 86 \text{ ч}, \quad (3)$$

где IQR — межквартильный интервал; Q_1 — нижний квартиль, который делит ряд так, что имеется 25 % наблюдений, которые меньше его, и 75 % наблюдений, которые больше его; Q_3 — верхний квартиль.

4. Визуальное изучение графических представлений временных рядов. Данный метод используется при построении модели процессов в работе [11] и основан на интуитивном определении структуры данных.

Результаты

Для оценки эффективного метода определения выбросов нанесем на диаграмму временного ряда полученные интервалы. На рис. 1 изображен интервал, соответствующий методу квартилей. Межквартильный интервал *IQR* покрыл 54 % всех объектов, и при расчетах доверительный интервал принимает отрицательные значения, чего быть не должно (подтверждено табличными данными на с. 870).

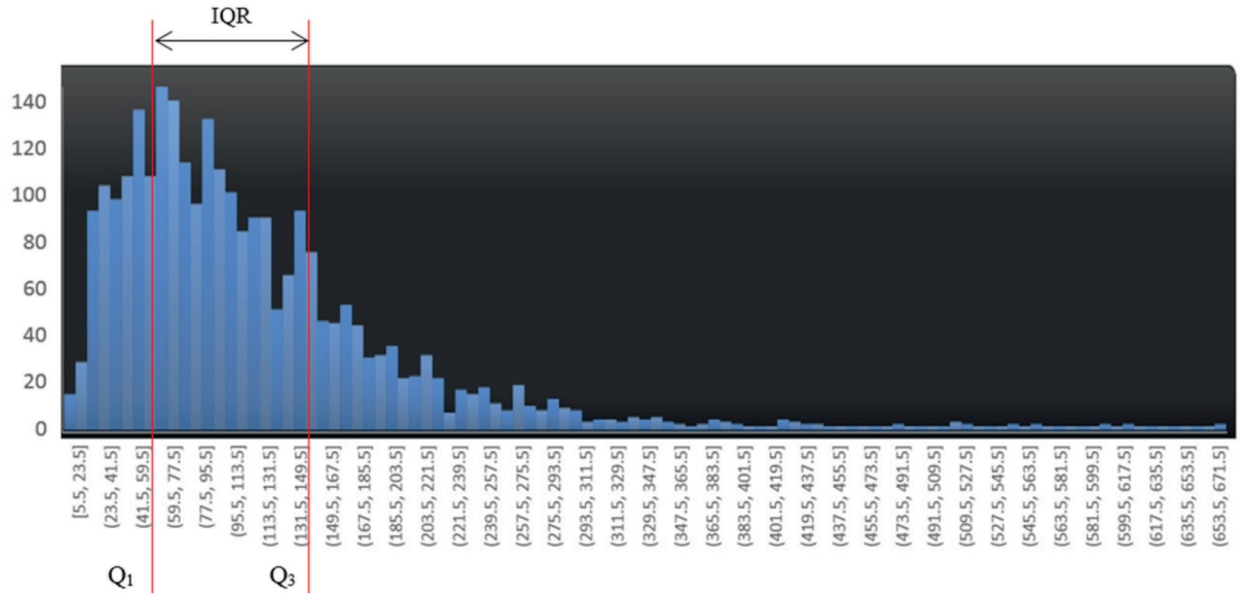

Рис. 1. Использование метода квартилей — межквартильного интервала

Далее, на рис. 2, указаны интервалы, полученные при использовании методов, основанных на математическом ожидании и медиане. Следует отметить, что интервал на основе математического ожидания дает лучшие результаты, чем квартили. Однако объекты, расположенные в области 23,5 ч, выбрасываются. При использовании эмпирического правила Бьенамэ–Чебышева [0], а именно двойного стандартного отклонения, значения снова дают отрицательный результат. Таким образом, наиболее эффективный и оптимальный метод определения выбросов показывает расчет на основе медианы.

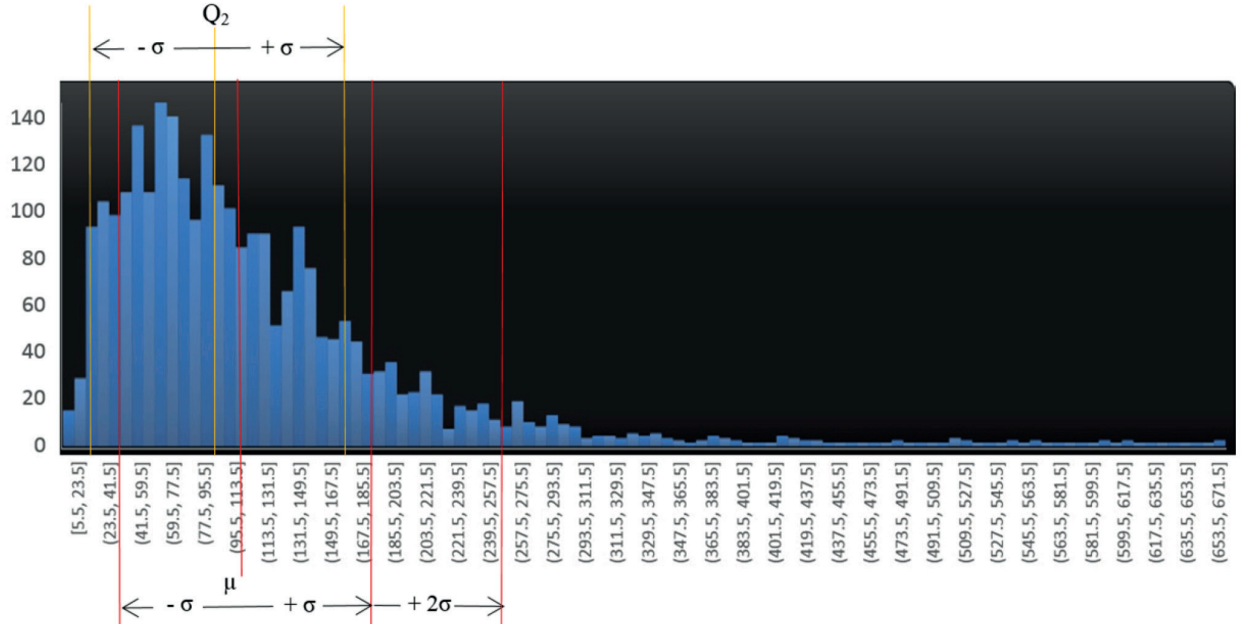

Рис. 2. Метод расчета доверительного интервала на основе медианы и математического ожидания

На основе выполненных расчетов была составлена сравнительная таблица, отражающая методы и их доверительные интервалы.

Сравнительная характеристика методов построения доверительного интервала

Метод построения интервала	Охват судозаходов, %	Доверительный интервал, ч
$\{Q_1 - 1{,}5IQR;\ Q_3 + 1{,}5IQR\}$	93	[–73,33; 270,67]
$\mu \pm \sigma$	77	[34,43; 180,55]
$\mu \pm 2\sigma$	95	[–38,63; 253,61]
$Q_2 \pm \sigma$	84	[17,61; 163,73]
Визуальное изучение диаграммы	92	[18; 216]

Обсуждения

Все статистические методы неспособны точно определить существующие выбросы в данных. Расчеты в большинстве случаев ведутся на основе математического ожидания, и выбросы могут повлиять на общий результат. Несмотря на то, что такие суммарные количественные характеристики, как первый, третий квартили и межквартильный размах не подвержены влиянию выбросов, их показатели также нельзя считать оптимальными в рассматриваемом примере. Теоретически основное количество наблюдений лежит в интервале плюс-минус (±) одно стандартное отклонение от среднего значения [7]. Однако на практике для асимметрично расположенных данных лучше использовать вместо среднего значения значение медианы (в источнике [12] также описываются преимущества использования медианы по сравнению со средним значением).

Использования ошибочных методов можно избежать, оценив изначально структуру данных или применив различные техники анализа данных. Для этого перед началом анализа можно провести проверку нормальности распределения ряда данных. С другой стороны, можно применить алгоритм для разбиения объекта по группам (кластеризация) и тем самым получить объекты с аномальным поведением. Эти варианты способствуют сокращению ошибочных результатов при использовании статистических методов. Их применение описано далее.

1. Проверка нормальности распределения временного ряда

Одно из требований классической регрессионной модели заключается в нормальности распределения случайной величины. Оно позволяет использовать стандартные процедуры и методы статистического анализа, поэтому проверка исследуемых переменных на нормальность распределения является важной составляющей частью анализа данных. Из приведенной таблицы видно, что некоторые доверительные интервалы попадают в область с отрицательными значениями, что для временного параметра является недопустимым. Следовательно, метод не может быть использован для анализа данных, поскольку дает некорректные результаты. Таким образом, необходимо искать другие робастные методы, для которых условие нормальности распределения данных некритично (например, линейная регрессия, метод главных компонент и т. д.).

Рис. 2 также демонстрирует, что медиана Q_2 принимает меньшее значение, чем математическое ожидание. Соответственно временный ряд имеет правую или положительную асимметрию, и использование медианы определяет вполне разумный временной интервал для проведения дальнейших исследований. Для того чтобы определить, возможно ли приведение ряда к нормальному распределению, следует прибегнуть к тесту Харки–Бера. В данном тесте проверяется нулевая гипотеза H_0: $S = 0$, $K = 3$ против гипотезы H_1: $S \neq 0$, $K \neq 3$, где S (Skewness) — коэффициент асимметрии; K (Kurtosis) — коэффициент эксцесса [13]. Чтобы вычислить основные показатели, был выбран язык программирования Python и его библиотеки для анализа данных. В результате получаются следующие показатели статистики: $S = 1{,}829$; $K = 9{,}161$; p-value = 0,000. При достаточно большом значении статистики p-value будет мало́, и тогда будет основание отвергнуть нулевую гипотезу. Статистики часто используют предельный порог для p-value $\geq 0{,}05$. Значения показате-

лей свидетельствуют о том, что временной ряд не имеет нормального распределения и, соответственно, однородности. Таким образом, предварительная проверка нормальности распределения временного ряда способна сократить время, затраченное на методы, которые приведут к ошибочным результатам.

2. Кластеризация

Все указанные ранее методы определения выбросов обладают существенным недостатком, состоящим в том, что будучи универсальными и обобщенными, они не учитывают специфику данных. Таким образом, шум и нетипичное поведение системы могут оказаться особым случаем, который также необходимо рассмотреть. Более приемлемой методикой в данном случае обработки судазахода на нефтяном терминале может быть *кластеризация*, т. е. разделение генеральной совокупности данных на группы / кластеры с последующим их анализом.

Задача кластеризации схожа с задачей классификации — группировкой объектов в кластеры / классы. Однако данный метод не учитывает особенности классической регрессионной модели, и число кластеров выбирается самим пользователем. В рассматриваемом случае кластеризация может быть использована для того, чтобы выделить нетипичные объекты, т. е. выявить отклонения в данных и в дальнейшем работать с ними. Для разбиения данных на кластеры используется программное средство RapidMiner (www.rapidminer.com) и алгоритм, определяющий расстояние между объектами.

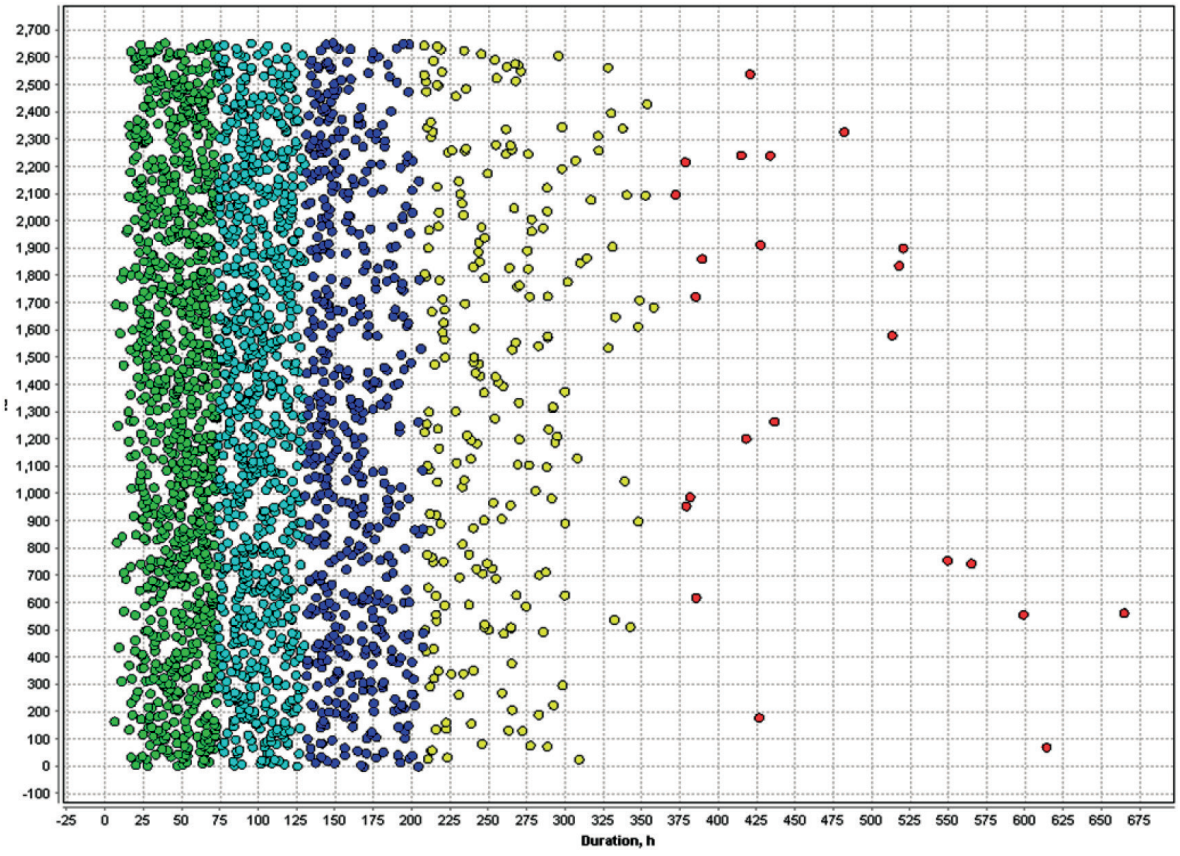

Рис. 3. Кластеризация объектов на пять кластеров

Как видно из рис. 3, судозаходы с длительной обработкой выделены в отдельный кластер и явно относятся к выбросам. Однако группа объектов с малым количеством времени обработки не была выделена. Соответственно при большом разбросе значений объектов применение кластеризации имеет смысл. Тем не менее, если значения расположены близко друг другу, то они не будут выделяться в отдельный кластер, как это произошло в данном случае (группа судозаходов с длительностью обработки в промежутке [5,5; 23,5]).

Заключение

На практике большинство аналитиков используют математическое ожидание для построения доверительного интервала. В машинном обучении также часто используется межквартильный интервал [9], не зависящий от существенного разброса значений анализируемого объекта. Тем не менее, ранее приведенный анализ показывает, что необходимо прежде всего ознакомиться со структурой данных и только уже затем применять известные статистические методы. Следует отметить, что при асимметричности ряда распределения величины расчеты, основанные на математическом ожидании и квартилях, приводят к потере важных данных. Кластеризация четко определяет экстремумы, точнее только судозаходы с длительной обработкой. Судозаходы, обработанные в течение короткого периода времени, не выделяются в отдельную группу. Таким образом, наиболее эффективной методикой при работе с ассиметричным распределением следует признать расчет доверительного интервала на основе медианы и стандартного отклонения.

Предварительная оценка структуры данных также играет большую роль в определении методов дальнейшего анализа, как, например, прогнозирование значения величины. Методы, используемые для нормального распределения, могут приводить к некорректным результатам при анализе асимметричного распределения.

Шумы и выбросы, определенные в данной работе, требуют, в свою очередь, дополнительного анализа. В дальнейшем можно изучить, какими именно признаками и свойствами объекта было вызвано подобное отклонение. Происхождение этих явлений не всегда вызвано ошибкой входных данных и, в свою очередь, может указывать на уникальность случаев. Изучение подобных объектов может повлиять на изменения в сфере оказания услуг в порту, сократить число объектов с аномальным поведением, а также улучшить в целом показатели процесса обработки судна на терминале.

Благодарности

Статья была написана при поддержке Министерства образования, молодежи и спорта Чешской Республики, Национальной программы NPU II; проект передовых технологий в области науки IT4Innovations-LQ1602.

СПИСОКЛИТЕРАТУРЫ

1. *Van der Aalst W.* Process mining: Overview and opportunities / W. Van der Aalst// ACM Transactions on Management Information Systems (TMIS). — 2012. — Vol. 3. — Is. 2. — Pp. 7. DOI:10.1145/2229156.2229157.

2. *Гржибовский А. М.* Типы данных, проверка распределения и описательная статистика / А. М. Гржибовский // Экология человека. — 2008. — № 1. — С. 52–60.

3. *Субботина А. В.* Описательная статистика и проверка нормальности распределения количественных данных / А. В. Субботина, А. М. Гржибовский // Экология человека. — 2014. — № 2. — С. 51–57.

4. *Кузовлев В. И.* Метод выявления аномалий в исходных данных при построении прогнозной модели решающего дерева в системах поддержки принятия решений / В. И. Кузовлев, А. О. Орлов // Наука и образование: науч. изд. МГТУ им. Н. Э. Баумана. — 2012. — № 9. — С. 16. DOI: 10.7463/0912.0483269.

5. *Chandolla V.* Anomalydetection: Asurvey / V. Chandolla, A. Banerjee, V. Kumar // ACM computingsurveys (CSUR). — 2009. — Vol. 41. — Is. 3. — Pp.15. DOI: 10.1145/1541880.1541882.

6. *Box G. E. P.*Time series analysis: forecasting and control /G. E. P. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung. — 5th edition. — John Wiley &Sons, 2015. — 712 p.

7. *Левин Д. М.* Статистика для менеджеров с использованием MicrosoftExcel / Д. М. Левин, Д. Стефан, Т. С. Кребиль, М. Л. Беренсон. — 4-е изд. — Пер. с англ. — М.: Вильямс, 2004. — 1312 с.

8. *Калинина В. Н.* Математическая статистика / В. Н. Калинина, В. Ф. Панкин. — М.: Высш. шк., 2002. — 336 с.

9. *Елисеева И. И.* Эконометрика / И. И. Елисеева. — М.: Изд-во «Юрайт», 2014. — 344 с.

10. *Mukherjee S.* Data Structures Using C: 1000 Problems and Solutions / S. Mukherjee. — McGrawHillEducation, 2007. — 556 p.

11. *Рудницкая Ю. Ю.* Исследование дескриптивной модели процесса обработки судна на нефтяном терминале на основе записей из таймшитов / Ю. Ю. Рудницкая // Вестник Астраханского государственного технического университета. Серия: Морская техника и технология. — 2017. — № 1. — С. 7–15.

12. *Дроботько Д. В.* Неавтономное обнаружение выбросов и очистка сигналов мониторинга внутричерепного давления /Д. В. Дроботько, А. И. Шевченко, В. Ф. Дроботько, И. В. Качур // Искусственный интеллект. — 2013. — № 3 (61). — С. 495–506.

13. *Damodar N. G.* Basic Econometrics / N. G. Damodar. — The McGraw-Hill Companies, 2004. — 1002 p.

REFERENCES

1. Van Der Aalst, Wil. “Process mining: Overview and opportunities.” *ACM Transactions on Management Information Systems (TMIS)* 3.2 (2012): 7. DOI:10.1145/2229156.2229157.

2. Grjibovski, A. M. “Data types, control of distribution and descriptive statistics.” *Human Ecology* 1 (2008): 52–60.

3. Subbotina, A. V., and A. M. Grjibovski. “Descriptive statistics and normality testing for quantitative data.” *Human Ecology* 2 (2014): 51–57.

4. Kuzovlev, V. I., and A. O. Orlov. “Method of detecting anomalies in the source data at constructing a prognostic model of a decision tree in decision support systems.” *Science and Education of Bauman MSTU* 9 (2012): 16. DOI: 10.7463/0912.0483269.

5. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. “Anomaly detection: A survey.” *ACM computing surveys (CSUR)* 41.3 (2009): 15. DOI: 10.1145/1541880.1541882.

6. Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time series analysis: forecasting and control.* 5th edition. John Wiley &Sons, 2015.

7. Levin, D. M., D. Stefan, T. S. Krebil’, and M. L. Berenson. *Statistika dlya menedzherov s ispol’zovaniem Microsoft Excel*. 4-e izd. M.: Vil’yams, 2004.

8. Kalinina, V. H., and V. F. Pankin. *Matematicheskaya statistika*. M.: Vyssh. shk., 2002.

9. Eliseeva, I. I. *Ekonometrika*. M.: Izdatel’stvo Yurait, 2014.

10. Mukherjee S. *Data Structures Using C: 1000 Problems and Solutions.* McGrawHillEducation, 2007.

11. Rudnitckaia, Julia Jurjevna. “Study of descriptive process model of ship handling at an oil terminal based on timesheets.” *Vestnik of Astrakhan State Technical University. Series: Marine Engineering and Technologies* 1 (2017): 7–15.

12. Drobotko, D. V., A. I. Shevchenko, V. F. Drobotko, and I. V. Kachur. “Online outliers detection and cleaning of intracranial pressure monitoring signals.” *Iskusstvennyi intellekt* 3(61) (2013): 495–506.

13. Damodar, N. G. *Basic Econometrics*. The McGraw-Hill Companies, 2004.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Рудницкая Юлия Юрьевна —
инженер, аспирант
Технический университет г. Брно
60190, Чешская Республика, Брно,
ул. Антонинска, 548/1
e-mail: *irudnickaia@fit.vutbr.cz*

INFORMATION ABOUT THE AUTHOR

Rudnitckaia, Yuliya Yu. —
Engineer, postgraduate
Brno University of Technology
548/1 Antonínská, Brno, 601 90,
Czech Republic
e-mail: *irudnickaia@fit.vutbr.cz*

Статья поступила в редакцию 24 апреля 2017 г.
Received: April 24, 2017.