

УДК 004.942
DOI 10.17513/snt.40397

МЕТОДИКА АНАЛИЗА И ОБРАБОТКИ ВРЕМЕННЫХ РЯДОВ МЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ

^{1,2}Юмашев Е.А.

¹ФГБУН «Институт динамики систем и теории управления им. В.М. Матросова СО РАН»,
Иркутск, e-mail: yumashevgeny@mail.ru;

²ФГБОУ ВО «Иркутский государственный университет путей сообщения», Иркутск

В статье рассмотрены вопросы анализа и обработки больших массивов метеорологических данных. Целью исследования является обмен опытом в выявлении и корректировке пропущенных и аномальных значений временных рядов метеорологических данных. В рамках исследования разработаны алгоритмы нормализации данных, выявления их индивидуальных и контекстных аномалий, а также корректировки пропущенных и аномальных значений. Особенности разработанных алгоритмов являются использование модели машинного обучения, основанной на применении деревьев решений, для изучения рядов данных при выявлении пропущенных значений, а также анализ временных и сезонных закономерностей при выявлении индивидуальных и контекстных аномалий на основе специализированных библиотек на языке программирования Python. Разработанные алгоритмы включены в программные модули сервис-ориентированного приложения для получения, обработки и анализа климатических данных при решении сложных мультидисциплинарных научных и прикладных задач экологического мониторинга Байкальской природной территории. Предложенная методика применена для анализа и обработки нескольких временных рядов метеорологических данных, полученных с метеостанций, которые расположены на Байкальской природной территории. Результаты применения методики показали существенное повышение качества обработанных данных, использованных в дальнейшем на практике в задачах моделирования работы автономных энергетических систем инфраструктурных объектов Байкальской природной территории.

Ключевые слова: метеорологические данные, пропущенные и аномальные значения, обработка данных, машинное обучение

Исследование проведено при поддержке Министерства науки и высшего образования Российской Федерации, проект № FWEW-2021-0005 «Технологии разработки и анализа предметно-ориентированных интеллектуальных систем группового управления в недетерминированных распределенных средах» (рег. № 121032400051-9).

METHOD FOR ANALYZING AND PROCESSING TIME SERIES OF METEOROLOGICAL DATA

^{1,2}Yumashev E.A.

¹Matrosov Institute for System Dynamics and Control Theory of SB RAS,
Irkutsk, e-mail: yumashevgeny@mail.ru;

²Irkutsk State Transport University, Irkutsk

The study investigates the problems of analysis and processing of large datasets of meteorological data. The goal of the research is to exchange experience in identifying and correcting missing and anomalous values in meteorological time series data. Within the scope of the study, algorithms for data normalization, detection of individual and contextual anomalies, and correction of missing and anomalous values, were developed. A distinctive feature of the developed algorithms is the use of a machine learning model based on decision trees for the analysis of data series for the identification of values, as well as for the analysis of temporal and seasonal patterns in the detection of individual and contextual anomalies, using specialized libraries in the Python programming language. The developed algorithms are integrated into software modules of a service-oriented application for obtaining, processing, and analyzing climate data in solving complex multidisciplinary scientific and applied tasks related to ecological monitoring of the Baikal Natural Territory. The proposed method was applied to the analysis and processing of several meteorological time series obtained from meteorological stations located in the Baikal Natural Territory. The results of the method application demonstrated a significant improvement in the quality of processed data, which was later used in practical tasks such as modeling the operation of autonomous energy systems for infrastructure facilities in the Baikal Natural Territory.

Keywords: meteorological data, missing and abnormal values, data processing, machine learning

The study was supported by the Ministry of Science and Higher Education of the Russian Federation, project No. FWEW-2021-0005 “Technologies for the development and analysis of domain-oriented intelligent group control systems in non-deterministic distributed environments” (reg. No. 121032400051-9).

Введение

Экологический мониторинг играет важную роль в изучении состояния окружающей среды и выявлении факторов, влияющих на ее изменения. Важным элементом такого мониторинга являются метеорологические наблюдения, которые предоставляют временные ряды данных о климатических и погодных условиях, воздействующих на исследуемые экосистемы. Метеорологические данные широко используются на практике для анализа тенденций изменения климата, оценки воздействия антропогенных факторов и прогнозирования экологических изменений. Надежность и точность таких данных имеют решающее значение для разработки стратегий управления природными ресурсами и адаптации к изменениям окружающей среды [1]. Выявление пропущенных [2] и аномальных [3] значений метеорологических временных рядов данных, а также их устранение и корректировка являются важной проблемой в области метеорологии [4]. В частности, климатические данные часто содержат пропуски, ошибки, абнормальные отклонения значений и другие погрешности, которые могут существенно влиять на результаты последующего анализа, прогнозирования или принятия решений. Примером таких данных являются архивы метеорологических наблюдений, в которых часто содержатся искаженные данные, что обусловлено техническими сбоями оборудования, ошибками измерений или экстремальными природными явлениями.

В связи с этим **целями данного исследования** являются разработка методики, а также обмен опытом в выявлении и корректировке пропущенных и аномальных значений в метеорологических временных рядах данных. В рамках методики предложены алгоритмы выявления погрешности данных и их корректировки. Приведен пример применения предложенной методики и подчеркнуты ее особенности и преимущества.

Материалы и методы исследования

Для анализа и обработки временных рядов метеорологических данных предложена методика, включающая следующие основные этапы:

- выявление пропусков значений данных и их заполнение:

- перевод категориальных данных в числовые данные;

- обучение модели прогнозирования пропущенных значений данных временного ряда;

- анализ временного ряда с целью выявления пропусков значений данных;

- заполнение пропусков значений данных временного ряда на основе обученной модели;

- выявление аномалий значений данных временного ряда:

- нормализация данных;

- анализ временного ряда метеорологических данных с целью выявления индивидуальных аномалий значений данных;

- смягчение (коррекция) индивидуальных аномалий значений данных временного ряда;

- анализ временного ряда метеорологических данных с целью выявления контекстных аномалий значений данных;

- смягчение (коррекция) контекстных аномалий значений данных временного ряда.

Общая схема применения данной методики продемонстрирована применительно к задаче построения моделей типичного, оптимистического и пессимистического годов, широко используемых в дальнейшем при решении различных научных и прикладных задач [5, 6], она приведена на рисунке. Исследование в рамках данной статьи направлено, в частности, на повышение точности вышеупомянутых моделей. Этапы применения предложенной методики выделены оливковым цветом.

Временной ряд метеорологических данных представляет собой последовательность наблюдений за определенными показателями, зафиксированными через равные промежутки времени. Такие данные важны для выявления закономерностей, трендов и предсказания будущих значений. Процесс выявления и заполнения пропусков значений является неотъемлемой частью анализа и обработки метеорологических данных. В рамках данного процесса использована регрессионная модель, представленная деревом решений [7]. Эта модель позволяет эффективно моделировать нелинейные зависимости между признаками, учитывать временные и контекстуальные особенности данных, а также обеспечивает высокую интерпретируемость результатов. Так как дерево решений не поддерживает работу с категориальными данными [8], они преобразуются в числовой формат.

Для построения дерева решений использованы временные признаки, такие как год, месяц, день и час. Эти признаки выбраны как независимые переменные, поскольку временные аспекты, такие как сезонные колебания и время суток, часто оказывают значительное влияние на метеорологические данные.

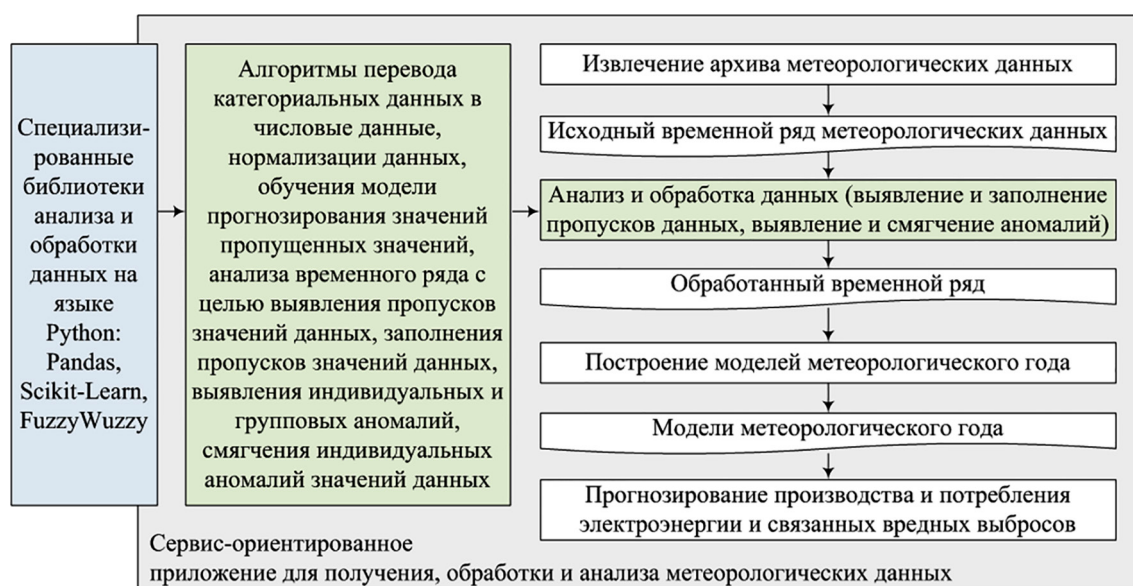


Схема применения методики в задаче построения моделей метеорологического года

Источник: составлено автором

Включение временных характеристик позволяет учесть изменения, связанные с различными временными периодами, что делает данную модель более чувствительной к закономерностям, присущим конкретным временным интервалам. Обученная модель применена для прогнозирования пропущенных значений в исследуемых временных рядах. Спрогнозированные значения использованы для заполнения пропущенных значений в соответствующих столбцах ряда. Это позволило восстановить недостающие данные без необходимости удаления строк с пропусками и обеспечить сохранение целостности и структуры данных. При этом категориальные данные, преобразованные в числовой формат, переводятся в исходный текстовый формат.

Процесс выявления аномалий в метеорологических данных включает выявление индивидуальных и групповых аномалий, а также их коррекцию. В рамках данного процесса важным этапом предобработки данных, использующих разные шкалы измерения, является их нормализация. Она обеспечивает приведение всех данных к единой шкале. Это особенно важно для методов, чувствительных к масштабу данных. В рамках исследования для выявления индивидуальных аномалий выбран метод ближайших соседей [9], поскольку результаты его работы легко интерпретируются, он обладает гибкостью в настройке, позволяет изменять число соседей и выбирать метрику расстояний. Это дает возможность адаптировать данный метод под раз-

личные характеристики данных и улучшить точность выявления аномальных значений в зависимости от структуры временного ряда. Метод ближайших соседей основан на предположении, что аномалии – это точки данных, которые имеют меньшую плотность соседей по сравнению с остальными точками данных временного ряда. Порог аномалии устанавливается на основе Z-оценки расстояний до ближайших соседей [10]. Такой подход позволяет выявлять выбросы на основе статистического анализа расстояний между точками в многомерном пространстве.

В отличие от индивидуальных аномалий, характеризующихся отдельными отклонениями значений данных, групповые аномалии проявляются в конкретном временном контексте рассмотрения данных, например в течение суток, месяца, квартала и т.д. [11]. В случае поиска групповой аномалии для ее выявления используется информация о заданном периоде времени. С этой целью вычисляются средние значения и стандартные отклонения для каждого периода времени. Если значение того или иного параметра временного ряда имеет существенное отклонение в рамках заданного временного контекста рассмотрения данных, то оно считается аномалией. После определения групповых аномалий данных выполняется коррекция их значений. Для этого используется регрессионная модель, представленная деревом решений и рассмотренная выше.

Алгоритмы выявления пропусков значений данных и их заполнения, а также об-

наружения индивидуальных и групповых аномалий значений данных реализованы на языке Python. Он включает широкий спектр программных библиотек, существенно упрощающих разработку алгоритмов машинного обучения, анализа и обработки данных. Для реализации алгоритмов выявления и заполнения пропусков используются свободно распространяемые библиотеки Pandas, Scikit-Learn и FuzzyWuzzy [12]. На первом этапе с помощью библиотеки Pandas загружаются данные временного ряда из Excel-файла. Затем из столбца ряда данных с датой извлекаются временные признаки (год, месяц, день и час), которые используются для обучения модели. Категориальные данные, включая информацию об облачности, высоте облаков, направлении ветра и других параметрах, преобразуются в числовые значения с использованием заранее подготовленных словарей, содержащих все возможные категории. Для заполнения пропущенных значений применяется модель дерева решений DecisionTreeRegressor из библиотеки Scikit-Learn, которая обучается на временных признаках. После обучения модель прогнозирует пропущенные значения, которые затем заменяют отсутствующие данные в соответствующих столбцах. Категориальные данные, преобразованные в числовой формат, впоследствии возвращаются в текстовую форму.

Алгоритмы выявления аномалий и их корректировки реализованы с использованием библиотек Scikit-Learn и Pandas. Выполняется нормализация данных [13] с помощью инструмента StandardScaler библиотеки Scikit-Learn. Все данные приводятся к единой шкале. Это обеспечивает корректную обработку данных алгоритмами машинного обучения. Для выявления индивидуальных аномалий используется метод ближайших соседей, реализованный в библиотеке Scikit-Learn с помощью инструментов KNeighborsClassifier и KNeighborsRegressor. Порог аномалии устанавливается на основе Z-оценки, вычисляемой через расстояния до ближайших соседей. Z-оценка служит индикатором того, на-

сколько сильно значение отклоняется от среднего. Если ее величина превышала установленный порог, то точка данных считалась аномальной. При выявлении групповых аномалий для каждого месяца вычисляются средние значения и стандартные отклонения для метеорологических показателей с использованием библиотеки Pandas. На основе этой статистической информации для каждого наблюдения оценивается, насколько его значение отклоняется от норм, характерных для конкретного месяца. Если отклонение является значительным, то данные считаются аномалией в рамках заданного временного контекста. После обнаружения аномалий применяется коррекция значений с использованием модели дерева решений DecisionTreeRegressor из библиотеки Scikit-Learn.

Результаты исследования и их обсуждение

Предложенная методика применена для анализа и обработки временных рядов данных, используемых в задаче прогнозирования экстремальных метеорологических условий мультиагентной системой моделирования взаимодействия микросетей [14]. Результаты ее применения показаны на модельном примере. В качестве ретроспективного временного ряда взят один из рядов данных, полученных с метеостанции, расположенной вблизи взаимодействующих микросетей. На его основе получен зашумленный ряд (ts_1) путем внесения пропусков данных и аномальных значений для следующих параметров: температура воздуха, атмосферное давление, скорость ветра и относительная влажность. Затем получен обработанный ряд (ts_2) с помощью алгоритмов выявления пропусков значений данных и их заполнения, а также выявления индивидуальных и групповых аномалий значений данных и их коррекции. Среднеквадратические отклонения данных зашумленного и обработанного (путем выявления и заполнения пропусков данных) рядов относительно данных ретроспективного временного ряда для вышеупомянутых параметров представлены в таблице 1.

Таблица 1

Результаты выявления и заполнения пропусков значений

Ряд	Температура воздуха	Атмосферное давление	Скорость ветра	Относительная влажность
ts_1	4,5958	4,5644	1,4706	16,1919
ts_2	4,5578	4,5592	1,4629	16,1883

Источник: составлено автором.

Таблица 2

Результаты выявления аномалий и их коррекции

Ряд	Температура воздуха	Атмосферное давление	Скорость ветра	Относительная влажность
ts_1	8,0610	59,5603	5,1360	11,5824
ts_2	4,4575	4,0105	1,5493	9,4202

Источник: составлено автором.

Среднеквадратические отклонения данных зашумленного и обработанного (путем выявления и коррекции аномалии данных) рядов относительно данных ретроспективного временного ряда для тех же параметров представлены в таблице 2.

В процессе работы алгоритмов выявления и заполнения пропущенных значений достигнуто существенное уменьшение среднеквадратического отклонения обработанных данных от реальных данных для тех же параметров составило 3,6035, 55,5498, 3,5867 и 2,1622 соответственно. Эти результаты подтверждают эффективность предложенной методики и ее практическую значимость с точки зрения возможности улучшения качества временных рядов метеорологических данных. Дальнейшее направление исследований связано с развитием предложенной методики. В частности, предполагается использование более сложных моделей машинного обучения и методов глубокой аналитики, представленных, например, в работе [15], с целью более точного обнаружения и смягчения аномальных значений, а также восстановления пропущенных значений.

Заключение

В рамках исследования предложена методика выявления и корректировки пропущенных и аномальных значений в метеорологических временных рядах данных с использованием специализированных библиотек программ на языке Python. Особенностью предложенной методики является совместное согласованное применение нормализации данных, выявления одиночных и контекстных аномалий на основе сравнения со значениями ближайших соседей во временном ряде, а также корректировки

пропущенных и аномальных значений с использованием дерева решений. Разработаны алгоритмы, обеспечивающие эффективное устранение пропущенных и аномальных значений метеорологических данных с целью смягчения влияния некорректных значений на результаты последующего моделирования в процессе решения научных и прикладных задач. Алгоритмы реализованы с использованием библиотек Pandas, Scikit-Learn и FuzzyWuzzy. Предложенная методика успешно применена для улучшения качества метеорологических данных, используемых для построения моделей метеорологического года.

Список литературы

1. Feoktistov A., Gorsky S., Kostromin R., Fedorov R., Bychkov I. Integration of Web Processing Services with Workflow-Based Scientific Applications for Solving Environmental Monitoring Problems // ISPRS International Journal of Geo-Information. 2022. Vol. 11. № 1. P. 8. DOI: 10.3390/ijgi11010008.
2. Макушкин М.С., Лапшин В.А. Обработка пропусков в рыночных данных на примере задачи оценки кривой доходностей облигаций // Финансы: теория и практика. 2023. Т. 27. № 6. С. 44–53. DOI: 10.26794/2587-5671-2023-27-6-44-53.
3. Darban Z.Z., Webb G.I., Pan S., Aggarwal C., Salehi M. Deep Learning for Time Series Anomaly Detection: A Survey // ACM Computing Surveys. 2024. Vol. 57. № 1. P. 15. DOI: 10.1145/3691338.
4. Jijo B.T., Abdulazeez A.M. Classification Based on Decision Tree Algorithm for Machine Learning // Standard Journal Issues. 2021. Vol. 2. № 1. P. 20–28. DOI: 10.38094/jastt20165.
5. Бычков И.В., Феоктистов А.Г., Горский С.А., Костромин Р.О., Федоров Р.К. Автоматизация интеграции сервисов веб-обработки данных экологического мониторинга с распределенными научными приложениями // Автометрия. 2022. Т. 58. № 4. С. 67–75. DOI: 10.15372/AUT20220400.
6. Карамов Д.Н., Еделев А.В., Феоктистов А.Г. Моделирование энергоснабжения объектов охраняемой Байкальской природной территории на основе возобновляемых источников энергии // Международный технико-экономический журнал. 2020. № 5. С. 7–24. DOI: 10.34286/1995-4646-2020-74-5-7-24.
7. Liu Y., Yang S. Application of Decision Tree-Based Classification Algorithm on Content Marketing // Journal of Mathematics. 2022. Vol. 2022. P. 1–10. DOI: 10.1155/2022/6469054.
8. Dinh D., Huynh V., Sriboonchitta S. Clustering mixed numerical and categorical data with missing values // Information Sciences. 2021. Vol. 571. P. 418–442. DOI: 10.1016/j.ins.2021.04.076.

9. Li R., Li S. Multimedia Image Data Analysis Based on KNN Algorithm // *Computational Intelligence and Neuroscience*. 2022. Vol. 2022. P. 1–28. DOI: 10.1155/2022/7963603.
10. Azim M., Sharif M.J. Usability of Z Score: A Case Study on Peoples Leasing and Financial Services Limited & Bangladesh Industrial Finance Company Limited // *International Journal of Management and Accounting*. 2020. Vol. 2. № 3. P. 38–46. DOI: 10.34104/ijma.020.038046.
11. Pasini K., Khoudja M., Samé A., Trépanier M., Oukhellou L. Contextual anomaly detection on time series: a case study of metro ridership analysis // *Neural Computing and Applications*. 2021. Vol. 34. P. 1483–1507. DOI: 10.1007/s00521-021-06455-z.
12. Afriyanto A., Wibawa Y.E. Development of Chatbot Services for Ordering Media Support Using Fuzzy Logic Algorithm: Case Study of PT. Pemuda Cari Cuan (Mangkokku) // *Journal of Electrical Engineering and Computer Sciences*. 2024. Vol. 9. № 1. P. 9–18. DOI: 10.54732/jeeecs.v9i1.2.
13. Голуб Ю.И., Старовойтов В.В. Нормализация данных в машинном обучении // *Информатика*. 2021. Т. 18. № 3. С. 83–96. DOI: 10.37661/1816-0301-2021-18-3-83-96.
14. Чекан М. Автоматизация разработки мультиагентных систем на основе иерархических машин состояний // *Современные наукоемкие технологии*. 2025. № 3. С. 59–65. DOI: 10.17513/snt.40324.
15. Kontopoulou V.I., Panagopoulos A.D., Kakkos I., Matsopoulos G.K. A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks // *Future Internet*. 2023. Vol. 15. № 8. P. 255. DOI: 10.3390/fi15080255.