УДК: 528.852+004.855.5 **DOI: 10.35595/2414-9179-2024-1-30-295-305**

А. А. Воробьева[1]

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ МЕТОДОМ СЛУЧАЙНОГО ЛЕСА И СИСТЕМ УПРАВЛЕНИЯ БОЛЬШИМИ ПРОСТРАНСТВЕННЫМИ ДАННЫМИ ДЛЯ ВОССТАНОВЛЕНИЯ РЯДОВ ДАННЫХ ВЕГЕТАЦИОННЫХ ИНДЕКСОВ

АННОТАЦИЯ

В данной статье рассмотрены содержание и результаты работы, посвященной разработке модели машинного обучения, позволяющей осуществить восстановление неполных данных с применением технологий облачных вычислений. Задача рассмотрена на примере исследования, посвященного моделированию данных для восполнения отсутствующих значений вегетационных индексов, основываясь на открытых каталогах данных платформ облачных вычислений. Предложенная методика основана на использовании многолетней периодичной выборки значений вегетационных индексов и обучения модели на больших объемах данных для повышения качества восстановления рядов. Указанный в работе подход позволяет добиться более высокой точности, нежели использование при восстановлении данных классических способов интерполяции, что делает моделируемые значения пригодными для использования при решении различных практических задач. Предложенная в работе методика реализована на примере восстановления значений нормализованного разностного вегетационного индекса, используемого для мониторинга и оценки состояния растительного покрова. В качестве исходных данных использовались массивы значений, полученные из каталогов облачной среды Google Earth Engine, предназначенной для обработки и анализа данных дистанционного зондирования Земли, по территории центральной части Новгородской области. Также, для ускорения процесса обучения модели и увеличения эффективности и производительности, использовались возможности платформы Google Colaboratory, что позволило не применять в исследовании локальные вычислительные мощности и специализированное программное обеспечение. Этот подход может быть адаптирован для восстановления других индексов или разрешения неполноты данных в различных предметных областях, что подчеркивает его универсальность и потенциальное практическое применение.

КЛЮЧЕВЫЕ СЛОВА: Google Earth Engine, регрессия, NDVI, Python

[1] Санкт-Петербургский Государственный Университет, Институт наук о Земле, Кафедра картографии и геоинформатики, Менделеевская линия, д. 2, Санкт-Петербург, Россия, 199034,
e-mail: **st096985@student.spbu.ru**

Anna A. Vorobyeva[1]

APPLICATION OF RANDOM FOREST MACHINE LEARNING AND BIG GEOSPATIAL DATA MANAGEMENT SYSTEMS APPLIED TO RECONSTRUCT THE VEGETATION INDEX DATA SERIES

ABSTRACT

This article discusses the content and results of the work devoted to the development of a machine learning model that allows for data incompleteness recovery using cloud computing. The problem is considered using the example of a study devoted to data modeling to fill in missing values of vegetation indices based on open data catalogs of cloud computing platforms. The proposed methodology is based on the use of a multi-year periodic sampling of vegetation index values and model training on large amounts of data to improve the quality of series reconstruction. The approach indicated in the work allows for higher accuracy than using classical interpolation methods for data recovery, which makes the modeled values suitable for use in solving various practical problems. The proposed method is implemented using the example of restoring the values of the Normalized Difference Vegetation Index used for monitoring and evaluating the state of vegetation cover. Arrays of values obtained from the catalogs of the Google Earth Engine cloud environment intended for processing and analyzing data from remote sensing of the Earth (on the territory of the central part of the Novgorod Region) were used as initial data. To accelerate the learning process of the model and increase efficiency and productivity, the capabilities of the Google Colaboratory platform were used, which made it possible not to use local computing capacity and do not use specialized software in the study. This approach can be adapted to reconstruct other indexes or resolve data incompleteness in various subject areas, which emphasizes its versatility and potential practical application.

KEYWORDS: Google Earth Engine, regression, NDVI, Python

ВВЕДЕНИЕ

В области наук об окружающей среде вычисление вегетационных индексов [*Черепанов*, 2017] является важным методом исследования, мониторинга и оценки динамики и состояния растительного покрова в различных условиях. Но получение полных и корректных данных часто оказывается невозможным из-за преград, связанных с атмосферным воздействием, особенностями датчиков сканирующих систем и сложным взаимодействием факторов окружающей среды. Для осуществления анализа состояния растительного покрова необходимо устранить возникающую неполноту данных. Использование методов машинного обучения [*Müller* et al., 2016; *Sarafanov* et al., 2020] позволяет моделировать значения вегетационных индексов, опираясь на многолетнюю статистику, а также на значения релевантных показателей, что способствует повышению точности и надежности восстановления пропусков NDVI[2] (Normalized Difference Vegetation Index — нормализованный разностный вегетационный индекс). Машинное обучение включает в себя широкий спектр вычислительных алгоритмов и методов, разработанных таким образом, чтобы модели могли автоматически обучаться и совершенствоваться, получая новую информацию. Практически

[1] Saint-Petersburg State University, Institute of Earth Sciences, Department of Cartography and Geoinformatics, 2, Mendeleevskaya line, Saint Petersburg, 199034, Russia,
e-mail: **st096985@student.spbu.ru**

[2] База данных индексных показателей: Index Database. Электронный ресурс: https://www.indexdatabase.de/ (дата обращения 17.05.2024)

машинное обучение основано на анализе закономерностей для составления прогнозов или принятия решений.

Однако на сегодняшний день нет универсального решения, позволяющего использовать методы машинного обучения для восстановления данных вегетационных индексов. В настоящей статье рассмотрен подход к решению задачи восстановления данных на примере устранения неполноты данных NDVI, возникающей из-за проблем при получении исходных мультиспектральных снимков.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

В настоящей работе в качестве исходных данных использовались значения вегетационного индекса для территории центральной части Новгородской области, рассчитанные с помощью коллекций данных облачной платформы Google Earth Engine[1].

Неполнота данных, возникающая при расчете NDVI, может быть обусловлена несколькими факторами, каждый из которых в разной степени зависит от методов, используемых при сборе и обработке данных (рис. 1).

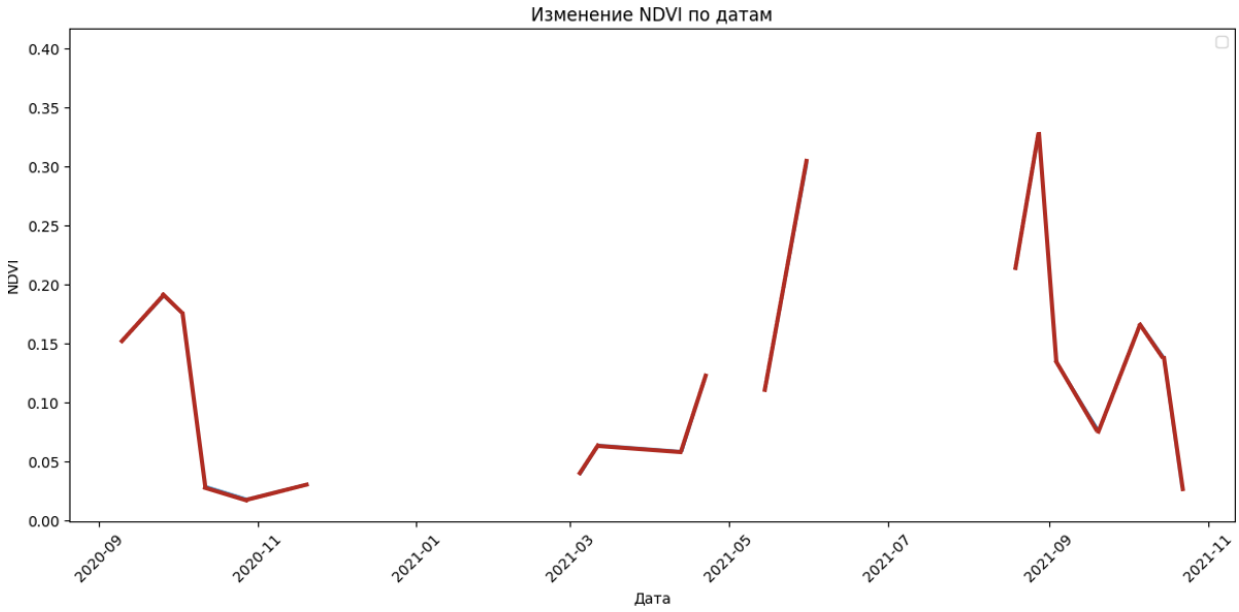

Рис. 1. График NDVI с отсутствующими значениями
Fig. 1. NDVI graph with missed values

Одной из основных причин возникновения неполноты данных является наличие облачного покрова на спутниковых снимках. Облака перекрывают обзор земной поверхности, что приводит к отсутствию полученных исходных данных. Существенным фактором являются также ограничения, связанные с работой датчиков, ошибками калибровки и спектральным разрешением, которые могут привести к получению ошибочных данных или полному их отсутствию. Кроме того, атмосферные помехи, такие как аэрозоли и дымка, искажают получаемые снимки и делают невозможным использование их для расчетов. Фенология растительности [*Шнелле*, 1961] также влияет на динамику NDVI, вызывая временное отсутствие корректных данных из-за циклов роста растительности и изменений растительного покрова с течением времени. Наконец, методы обработки данных и

[1] Облачный ресурс для обработки данных дистанционного зондирования: Google Earth Engine. Электронный ресурс: https://earthengine.google.com/ (дата обращения 21.04.2024)

алгоритмы, используемые для расчета индекса, также способствуют неполноте данных в том случае, если они работают некорректно. В совокупности эти факторы подчеркивают трудности, связанные с получением подходящих данных для экологического мониторинга и анализа.

Методы восстановления пропусков [*Julien* et al., 2019] в данных обычно представляют собой интерполяцию и экстраполяцию для заполнения недостающих значений на основе имеющихся данных. Подобные способы направлены на анализ значений индекса в тех областях, где данные являются неполными. В данном исследовании подход к анализу исходных данных отличался от классического — оценивались данные многолетней выборки [*Weigend*, 2018; *Saad* et al., 2020], а не конкретно участки образования пропусков. Восстановление выполнялось с помощью обученной на выборке модели случайного леса [*Тараканов*, 2023; *Hastie* et al., 2009; *Zhu*, 2020]. Для осуществления обучения использовались мощности облачных сервисов [*Мордовина*, 2012; *Бучев* и др., 2017] Google: Google Colaboratory (Google Colab)[1] и Google Earth Engine. Google Earth Engine — это облачный сервис, предоставляющий доступ к данным дистанционного зондирования и инструментам для их анализа. Он включает в себя библиотеки данных и редактор кода, что позволяет пользователям создавать различные алгоритмы и программы для обработки данных. Общедоступный каталог данных Google Earth Engine — это большое, непрерывно обновляемое хранилище часто используемых наборов геопространственных данных. В настоящем исследовании взаимодействие с платформой Google Earth Engine осуществлялось через Google Colab. Google Colab — это облачная платформа, разработанная для совместного программирования и анализа данных с использованием Jupyter Notebook[2] на базе Python[3]. Она предоставляет бесплатный доступ к графическим (GPU — Graphics Processing Unit) и тензорным процессорам (TPU — Tensor Processing Unit), что делает ее удобным инструментом для решения задач машинного обучения [*Pessoa* et al., 2018].

Точность алгоритмов для вычисления пропущенных данных является критически важным аспектом обработки и анализа данных. В качестве оценочных метрик в данной работе применялись показатели производительности, используемые для количественной оценки полученных результатов:

1. Среднеквадратичная ошибка (MSE — Mean Squared Error)[4] — измеряет среднюю разницу между наблюдаемыми и прогнозируемыми значениями, при этом более низкие значения указывают на более высокую точность вычисления.
2. Средняя абсолютная ошибка (MAE — Mean Absolute Error)[5] — вычисляет среднюю абсолютную разницу между наблюдаемыми и прогнозируемыми значениями, обеспечивая меру точности алгоритма без учета направления ошибок.

---

1 Облачный ресурс для работы с Python: Google Colaboratory. Электронный ресурс: https://colab.research.google.com/ (дата обращения 20.04.2024)

2 Пользовательская среда взаимодействия с Python: Jupyter Notebook. Электронный ресурс: https://jupyter.org/ (дата обращения 20.04.2024)

3 Язык программирования Python: Python. Электронный ресурс: https://www.python.org/ (дата обращения 13.04.2024)

4 Метрика MSE в документации библиотеки Scikit-learn: Scikit-learn documentation. Электронный ресурс: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html (дата обращения 26.03.2024)

5 Метрика MAE в документации библиотеки Scikit-learn: Scikit-learn documentation. Электронный ресурс: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html (дата обращения 26.04.2024)

3. R2 (коэффициент детерминации)[1] — R2 количественно определяет долю отклонения в наблюдаемых значениях, которая объясняется расчетными значениями. Более высокое значение (более близкое к единице) R2 указывает на более высокую эффективность расчета.

Разработанная методика предполагает такие этапы, как:
1) получение данных, предобработка, визуализация;
2) обучение модели и предсказание;
3) оценка точности.

Географическая область, представляющая интерес, определяется с использованием полигональной геометрии с указанием региона для анализа. В данной статье эксперимент по восстановлению значений вегетационных индексов проводился на участке, расположенном в пределах территории Новгородской области. Апробация методики проводилась на территориях, находящихся в Ямало-Ненецком автономном округе и Краснодарском крае. Все участки расположены в различных природных зонах:

- тундровой (остров Белый, Ямало-Ненецкий автономный округ);
- лесной (Новгородская область);
- степной (Усть-Лабинский район Краснодарского края).

Каждый из этих участков имеет свои особенности и преобладающий тип землепользования. Площадь каждого участка составляла порядка 3,5 тыс. га.

Тундровый участок находится на острове Белом, который является частью Ямало-Ненецкого автономного округа. Этот участок характеризуется суровыми климатическими условиями и низкой продуктивностью растительности. Здесь преобладает использование земли под пастбища.

Лесной участок расположен в Новгородской области. Он отличается умеренным климатом и богатым разнообразием растительности. Основными типами землепользования являются сельское хозяйство и лесопользование.

Степной участок находится в Усть-Лабинском районе Краснодарского края. Климат здесь теплый и сухой, а растительность представлена преимущественно злаковыми травами. Земледелие является здесь основным типом землепользования.

Таким образом, выбор рассматриваемых территорий был обусловлен их различными природными условиями и типами землепользования, что позволяет провести более полный анализ эффективности методов восстановления растительности в разных экосистемах. Выбираются коллекции спутниковых снимков Landsat-8 [*Schmid*, 2017] и Sentinel-2 с примененной атмосферной коррекцией [*Pacifici* et al., 2014]. Эти коллекции охватывают требуемые временные рамки (в данном исследовании коллекции выбирались за диапазон с января 2015 г. по январь 2023 г.) и географический район. Всего было использовано около 300 снимков, которые имели облачность не более 15 %.

Значения NDVI рассчитываются для каждого изображения в выбранных коллекциях по формуле[2]:

---

[1] Метрика R2 в документации библиотеки Scikit-learn: Scikit-learn documentation. Электронный ресурс: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (дата обращения 26.04.2024)

[2] База данных индексных показателей: Index Database. Электронный ресурс: https://www.indexdatabase.de/ (дата обращения 02.02.2024)

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

Средние значения NDVI для исследуемой области извлекаются из каждого изображения. Результаты сохраняются в виде объектов в массиве данных. Полученные значения NDVI обрабатываются и визуализируются с использованием библиотеки Pandas[1] для обработки данных и Matplotlib[2] для построения графиков (рис. 2).

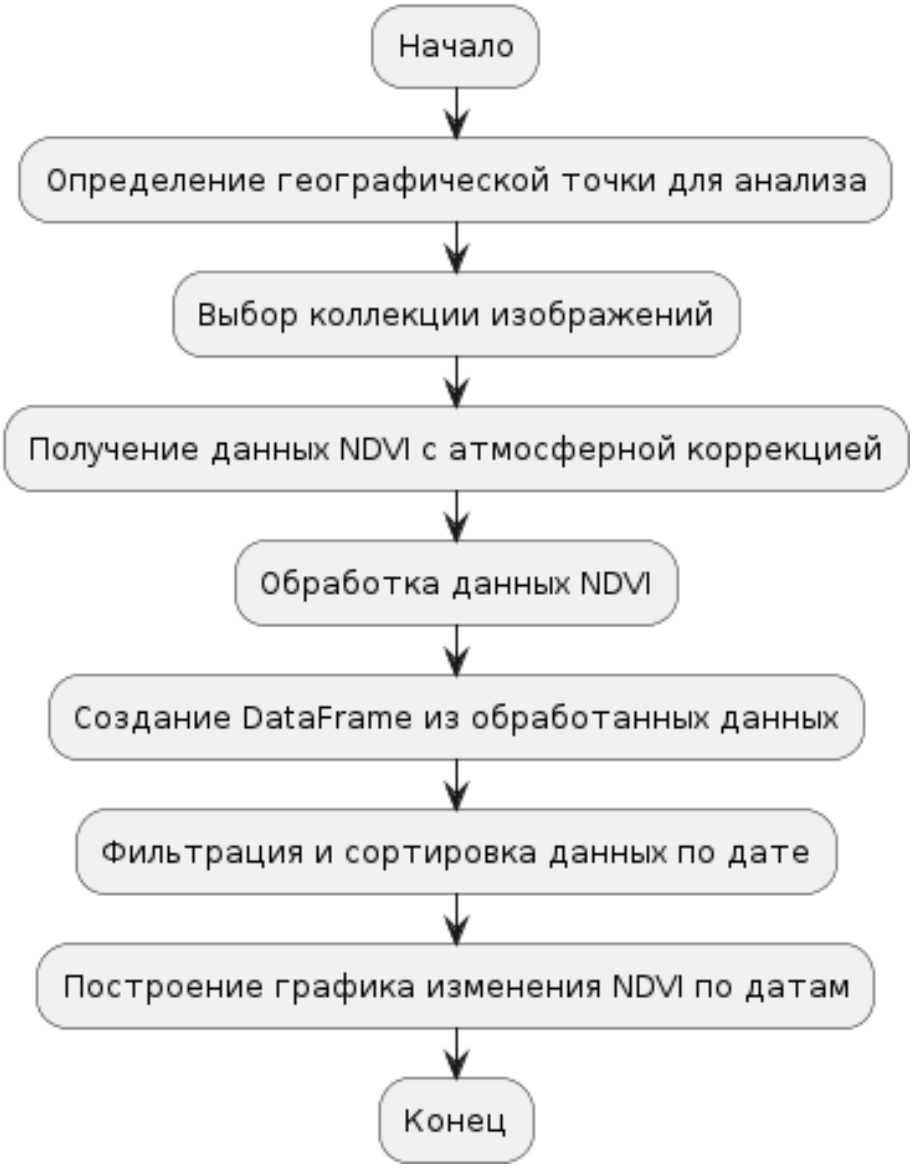

Рис. 2. Схема алгоритма предобработки данных
Fig. 2. Diagram of the data preprocessing algorithm

[1] Документация библиотеки Pandas: Pandas documentation. Электронный ресурс: https://pandas.pydata.org/docs (дата обращения 24.03.2024)

[2] Документация библиотеки Mathplotlib: Mathplotlib documentation. Электронный ресурс: https://matplotlib.org/stable/index.html (дата обращения 22.03.2024)

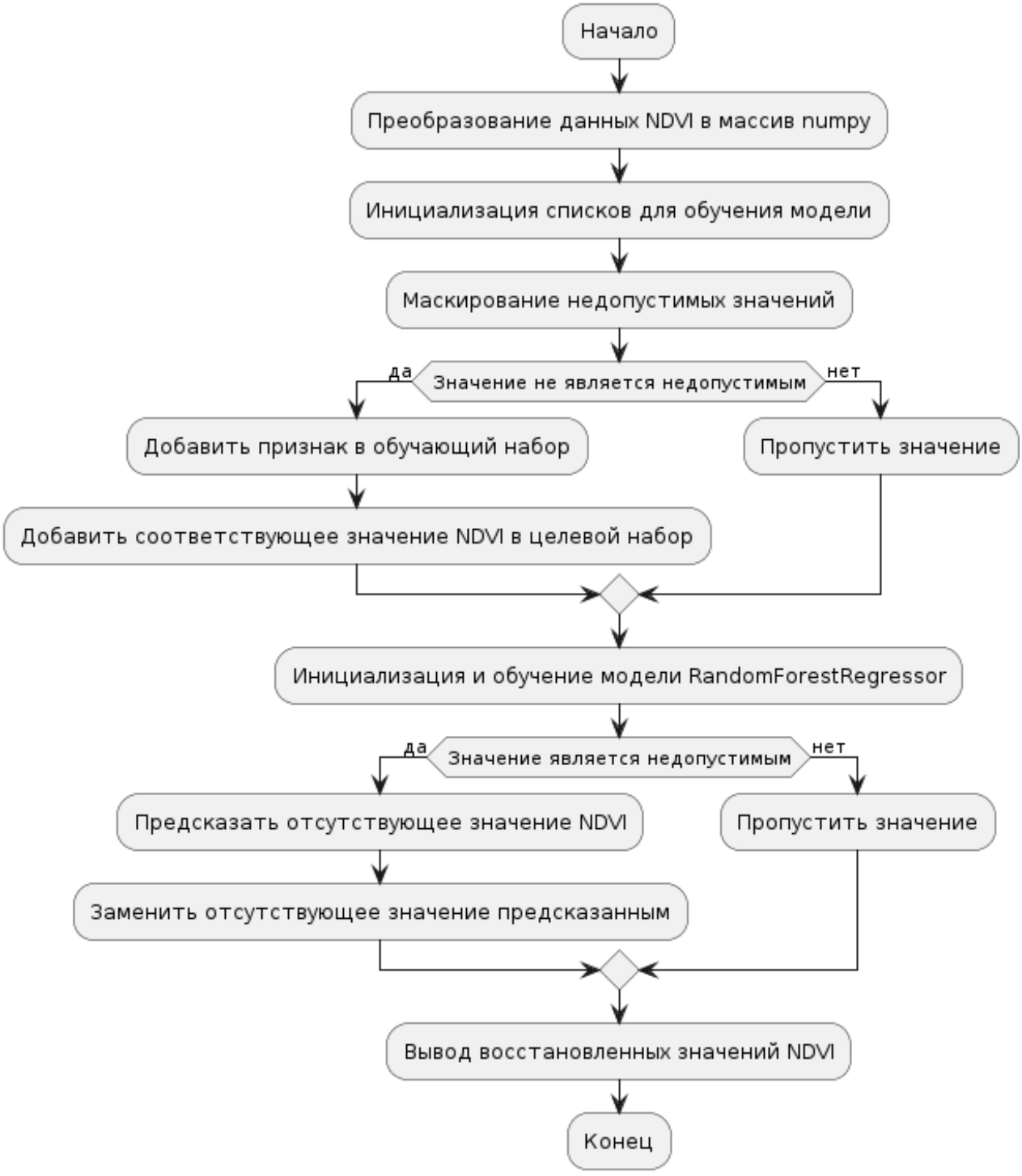

Рис. 3. Схема разработанного алгоритма для предсказания значений
Fig. 3. The scheme of the developed algorithm

Алгоритм (рис. 3) извлекает данные NDVI из массива данных (DataFrame) и преобразует их в массив NumPy[1]. Пропущенные значения в данных NDVI обрабатываются с помощью маскированного массива NumPy, чтобы идентифицировать и замаскировать недопустимые (отсутствующие) значения. Допустимые (не пропущенные) точки данных разбиваются на объекты и метки, которые будут использоваться для обучения регрессионной модели случайного леса. Создается экземпляр RandomForestRegressor[2], и модель обучается с использованием данных объекта и метки. Пропущенные значения в данных NDVI вычисляются с использованием обученной модели. Для каждого отсутствующего значения индекса используется соответствующее значение признака, чтобы предсказать пропущенное значение NDVI и визуализировать его с помощью графика с восстановленными значениями (рис. 4).

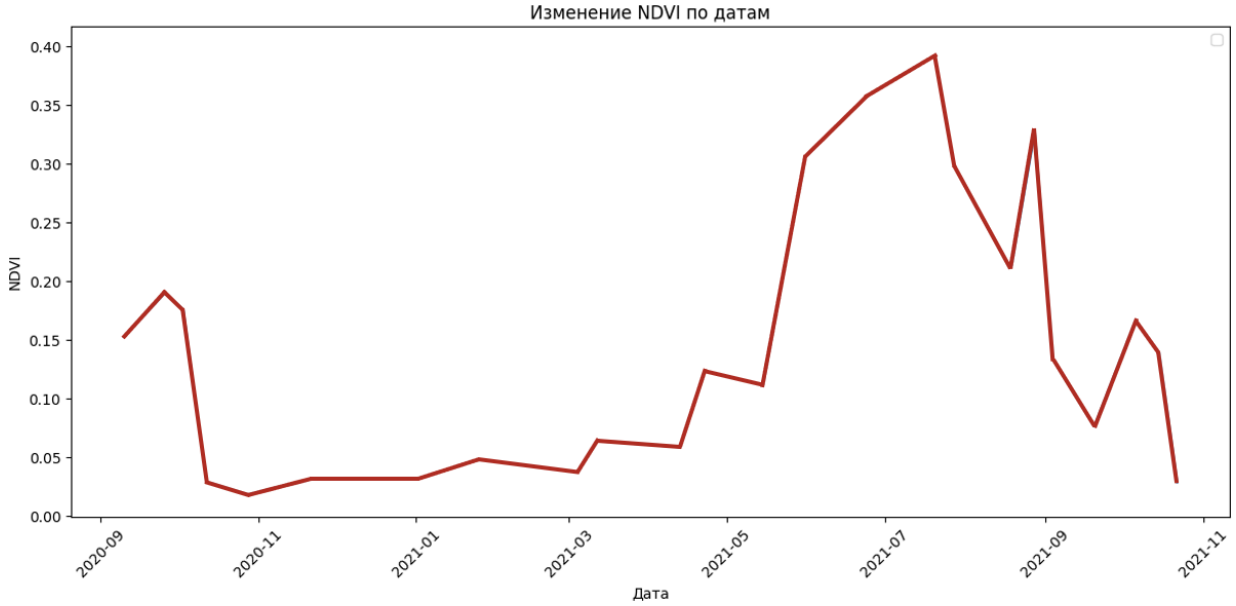

Рис. 4. График NDVI с восстановленными значениями
Fig. 4. NDVI graph with restored values

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Эффективность процесса расчета оценивается с использованием таких показателей, как R2 (коэффициент детерминации), MAE (средняя абсолютная ошибка) и MSE (среднеквадратичная ошибка).

Коэффициент детерминации измеряет долю отклонения в наблюдаемых значениях, которая объясняется прогнозируемыми значениями. Значение, близкое к 1, указывает на то, что модель объясняет значительную долю изменчивости в наблюдаемых данных, в то время как значение, близкое к 0, указывает на низкую эффективность. В результате данного исследования на примере территории Новгородской области значение R2 составляет 0,510. Это говорит о том, что расчетные значения объясняют чуть больше половины вариабельности наблюдаемых данных, а это указывает на умеренную прогностическую

[1] Документация библиотеки NumPy: NumPy documentation. Электронный ресурс: https://numpy.org/doc/ (дата обращения 06.04.2024)

[2] Функция RandomForestRegressor в документации библиотеки Scikit-learn: Scikit-learn documentation. Электронный ресурс: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble. RandomForestRegressor.html (дата обращения 18.04.2024)

эффективность. Средняя абсолютная ошибка (MAE) обеспечивает простую оценку точности алгоритма, при этом более низкие значения указывают на более высокую производительность. Значение MAE, равное 0,081, указывает на то, что в среднем расчетные значения примерно на 0,081 единицы отличаются от наблюдаемых значений. Хотя это значение может показаться небольшим, его следует интерпретировать в контексте данных и конкретных требований приложения. При анализе среднеквадратичной ошибки (MSE) более низкие значения будут указывать на высокую производительность. Значение MSE, равное 0,010, указывает на то, что в среднем квадратичная разница между наблюдаемыми и прогнозируемыми значениями составляет около 0,010. Стоит отметить, что эти расчеты проводились на изначально полных наборах данных, из которых для исследования в случайном порядке удалялись значения. Это позволило получить более объективную оценку точности моделей и минимизировать влияние пропущенных данных на результаты.

Апробация на территориях Ямало-Ненецкого автономного округа и Краснодарского края показала, что разработанный алгоритм имел даже чуть более высокую точность восстановления (чем на тестовом наборе) для более северной территории и чуть более низкую точность для южной территории. Но необходимо учесть, что значения NDVI могут быть ниже в зоне тундр из-за специфических условий окружающей среды. Это может объяснить лучшую точность модели для этой зоны. Тем не менее, важно помнить, что точность модели также зависит от других факторов, таких как качество данных, методы обработки и обучения модели. Сравнение результатов точности производилось на основе коэффициента детерминации, т. к. данная метрика наиболее полно отражает точность полученной модели (табл. 1).

Табл. 1. Оценка точности по результатам дополнительного тестирования
Table 1. Accuracy assessment based on the results of additional testing

Территории тестирования	**Метрики оценки точности**		
	MSE	**MAE**	**R2**
Область на территории Новгородской области	0,081	0,010	0,510
Область на территории Ямало-Ненецкого автономного округа	0,047	0,004	0,542
Область на территории Краснодарского края	0,077	0,011	0,501

ВЫВОДЫ

Предложенная методика может быть применена для оптимизации процессов и повышения качества моделирования данных при использовании временных рядов вегетационных индексов. Она позволяет уменьшить время, затрачиваемое на восстановление отсутствующих значений, а также повысить точность предсказания. В данном исследовании рассмотрено использование временных рядов, которые сами по себе играют решающую роль в точном вычислении пропущенных значений. В ходе исследования удалось явно оценить результативность и итоговую точность предложенного алгоритма восстановления данных, что позволяет оценить его пригодность для решения конкретных задач конечного пользователя.

СПИСОК ЛИТЕРАТУРЫ

Бучнев А. А., Пяткин В. П., Пяткин Ф. В. Модель облачной среды для обработки данных дистанционного зондирования Земли. ИТНОУ: Информационные технологии в науке, образовании и управлении, 2017. № 3. С. 57–61.

Мордовина Д. О. Облачные вычисления в сфере геоинформационных технологий и ДЗЗ. Геоматика, 2012. № 2. С. 9–11.

Тараканов Д. А. Восстановление пропущенных значений в данных гидрометеорологических наблюдений с использованием машинного обучения (на примере реки Белая, Республика Башкортостан). Вестник Евразийской науки, 2023. Т. 15. № 6.

Шнелле Ф. Фенология растений. Ленинград: Гидрометеоиздат, 1961. 259 с.

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer, 2009. 746 p.

Julien Y., Sobrino J. A. Optimizing and comparing gap-filling techniques using simulated NDVI time series from remotely sensed global data. International Journal of Applied Earth Observation and Geoinformation, 2019. V. 76. P. 93–111. DOI: 10.1016/j.jag.2018.11.008.

Pacifici F., Longbotham N., Emery W. J. The Importance of physical quantities for the analysis of multitemporal and multiangular optical very high spatial resolution images. IEEE Transactions on Geoscience and Remote Sensing, 2014. V. 52. No. 10. P. 6241–6256. DOI: 10.1109/TGRS.2013.2295819.

Pessoa T., Medeiros R., Nepomuceno T., Bian G., Albuquerque V. H. C., Filho P. P. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. IEEE Access, 2018. V. 6. P. 61677–61685. DOI: 10.1109/ACCESS.2018.2874767.

Saad M., Chaudhary M., Karray F., Gaudet V. Machine learning based approaches for imputation in time series data and their impact on forecasting. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020. P. 2621–2627.

Sarafanov M., Kazakov E., Nikitin N. O., Kalyuzhnaya A. V. A machine learning approach for remote sensing data gap-filling with open-source implementation: An example regarding land surface temperature, surface albedo and NDVI. Remote Sensing, 2020. V. 12. Iss. 23. P. 3865. DOI: 10.3390/rs12233865.

Schmid J. N. Using Google Earth Engine for Landsat NDVI time series analysis to indicate the present status of forest stands. 2017. DOI: 10.13140/RG.2.2.34134.14402/6.

Weigend A. S. Time series prediction: forecasting the future and understanding the past. Routledge, 2018. 663 p. DOI: 10.4324/9780429492648.

Zhu T. Analysis on the Applicability of the Random Forest. Journal of Physics: Conference Series, 2020. V. 1607. P. 012123. DOI: 10.1088/1742-6596/1607/1/012123.

REFERENCES

Buchnev A. A., Pyatkin V. P., Pyatkin F. V. Cloud environment model for processing Earth remote sensing data. ITNOU: Information technologies in science, education and management, 2017. No. 3. P. 57–61 (in Russian).

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer, 2009. 746 p.

Julien Y., Sobrino J. A. Optimizing and comparing gap-filling techniques using simulated NDVI time series from remotely sensed global data. International Journal of Applied Earth Observation and Geoinformation, 2019. V. 76. P. 93–111. DOI: 10.1016/j.jag.2018.11.008.

Mordovina D. O. Cloud computing in the field of geoinformation technologies and remote sensing. Geomatics, 2012. No. 2. P. 9–11 (in Russian).

Pacifici F., Longbotham N., Emery W. J. The Importance of physical quantities for the analysis of multitemporal and multiangular optical very high spatial resolution images. IEEE Transactions on

Geoscience and Remote Sensing, 2014. V. 52. No. 10. P. 6241–6256. DOI: 10.1109/TGRS.2013.2295819.

Pessoa T., Medeiros R., Nepomuceno T., Bian G., Albuquerque V. H. C., Filho P. P. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. IEEE Access, 2018. V. 6. P. 61677–61685. DOI: 10.1109/ACCESS.2018.2874767.

Saad M., Chaudhary M., Karray F., Gaudet V. Machine learning based approaches for imputation in time series data and their impact on forecasting. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020. P. 2621–2627.

Sarafanov M., Kazakov E., Nikitin N. O., Kalyuzhnaya A. V. A machine learning approach for remote sensing data gap-filling with open-source implementation: An example regarding land surface temperature, surface albedo and NDVI. Remote Sensing, 2020. V. 12. Iss. 23. P. 3865. DOI: 10.3390/rs12233865.

Schmid J. N. Using Google Earth Engine for Landsat NDVI time series analysis to indicate the present status of forest stands. 2017. DOI:10.13140/RG.2.2.34134.14402/6.

Schnelle F. Plant phenology. Leningrad: Gidrometeoizdat, 1961. 259 p. (in Russian).

Tarakanov D. A. Missing values recovering in hydrometeorological data using machine learning (a case study from the Belaya River, Republic of Bashkortostan). The Eurasian Scientific Journal, 2023. V. 15. No. 6 (in Russian).

Weigend A. S. Time series prediction: forecasting the future and understanding the past. Routledge, 2018. 663 p. DOI: 10.4324/9780429492648.

Zhu T. Analysis on the Applicability of the Random Forest. Journal of Physics: Conference Series, 2020. V. 1607. P. 012123. DOI: 10.1088/1742-6596/1607/1/012123.