

М. Е. Сиденко^{1,2}, С. А. Рылов¹✉

Гибридный алгоритм выделения водной поверхности по спутниковым снимкам Sentinel-2 на основе глубокого обучения и кластерного анализа

¹Федеральный исследовательский центр информационных и вычислительных технологий,
г. Новосибирск, Российская Федерация

²Новосибирский государственный университет, г. Новосибирск, Российская Федерация
e-mail: RylovS@mail.ru

Аннотация. В работе рассматривается задача автоматического выделения водной поверхности на мультиспектральных спутниковых снимках Sentinel-2, в условиях паводковой ситуации и повышенной облачности. Показан качественный характер различий спектральных характеристик объектов в освещённых и теневых зонах сцены, что обосновывает необходимость их раздельной обработки при выделении водной поверхности. Предложен гибридный алгоритм, сочетающий предварительное маскирование облачности, плотностную кластеризацию по спектральным признакам и последующую классификацию кластеров методом случайного леса. В качестве признаков для кластеризации использовались спектральные индексы NDWI, NDVI и MNDWI. Для выделения облачности применялся алгоритм OmniCloudMask. Экспериментальная оценка на наборе данных WorldFloods показала высокое качество выделения водной поверхности как в ясных, так и в облачных зонах сцен.

Ключевые слова: дистанционное зондирование Земли, паводки, сегментация водной поверхности, кластеризация, случайный лес, спектральные индексы, Sentinel-2

М. Е. Sidenko^{1,2}, S. A. Rylov¹✉

A hybrid algorithm for water surface segmentation from Sentinel-2 satellite images using deep learning and cluster analysis

¹Federal Research Center for Information and Computational Technologies,
Novosibirsk, Russian Federation

²Novosibirsk State University, Novosibirsk, Russian Federation
e-mail: RylovS@mail.ru

Abstract. The paper considers the problem of automatic water surface segmentation on Sentinel-2 multispectral satellite imagery under flood conditions and heavy cloud coverage. Qualitative differences in the spectral characteristics of objects in illuminated and shadowed areas of a scene are demonstrated, which justifies their separate processing during water surface extraction. A hybrid algorithm combining preliminary cloud masking, density-based clustering of spectral features, and subsequent cluster classification using the Random Forest method is proposed. The spectral indices NDWI, NDVI, and MNDWI were used as clustering features. The OmniCloudMask algorithm was applied for cloud detection. Experimental evaluation on the WorldFloods dataset demonstrated high-quality water surface segmentation both in clear and cloud-covered areas of the scenes.

Keywords: remote sensing, floods, water surface segmentation, clustering, random forest, spectral indices, Sentinel-2

Введение

Полноценное решение задачи оперативного мониторинга паводковой ситуации невозможно без использования данных дистанционного зондирования Земли. Среди современных спутниковых систем оптического наблюдения особый интерес для задач картографирования паводков представляют аппараты серии Sentinel-2. Их широкое применение обусловлено сочетанием относительно высокого пространственного разрешения, регулярности съёмки, глобального покрытия и открытой доступности данных, что делает возможным оперативный анализ паводковой обстановки на больших территориях.

Задача выделения водной поверхности в условиях паводков существенно сложнее выделения границ постоянных водных объектов. Паводковая вода смешивается с растительностью и почвой, формируя спектрально неоднородные области [1]. Дополнительные сложности создают облачность, тени от облаков и рельефа местности, а также низкий контраст сцен в условиях плохой освещенности [2]. В результате водные объекты по своим спектральным характеристикам могут пересекаться с тенями, влажными почвами и затемнёнными участками поверхности, что приводит к ошибкам сегментации. Поэтому разработка алгоритма автоматического выделения воды, способного работать в сложных условиях, является актуальной задачей.

Целью исследования является разработка алгоритма автоматического выделения водной поверхности на спутниковых снимках Sentinel-2, способного работать в условиях паводков и повышенной облачности.

Материалы и методы

В качестве источника данных использовались мультиспектральные снимки аппарата Sentinel-2, с уровнем предобработки L1C, из открытого набора данных WorldFloods [3]. Набор содержит 509 сцен паводков различной сложности, включая случаи с повышенной облачностью, тенями от облаков и низким контрастом. В набор также включены маски воды и облаков для каждого снимка. Также Сибирским Центром «НИЦ «Планета» были предоставлены аннотации снимков разливов рек, сделанные аппаратами Sentinel-2 в количестве 10 штук, которые были использованы для визуальной оценки качества работы алгоритма.

Для исследования влияния теней от облаков на спектральные характеристики объектов на одном из снимков предоставленных «НИЦ «Планета» были вычислены средние значения спектральных признаков NIR и NDWI на границе освещенной и теневой зон. Полученные результаты показали наличие существенных различий спектральных характеристик объектов в освещённых и теневых областях сцены. Это обусловило необходимость предварительного выделения зон различной освещенности и их независимой обработки в составе создаваемого алгоритма. С этой целью применялся ансамблевый алгоритм глубокого обучения OmniCloudMask [4], формирующий маски четырех классов: освещенной зоны, плотных облаков, полупрозрачной облачности и теней от облаков.

В основе разработанного алгоритма лежит подход, основанный на предварительной кластеризации изображения с последующей классификацией полученных

кластеров на водные и неводные, что позволяет повысить устойчивость результатов и снизить влияние выбросов [5, 6]. Настоящая работа развивает подход, представленный авторами в [7], адаптируя его к условиям повышенной облачности.

Для кластеризации был выбран сеточный иерархический алгоритм НСА [6]. Он обладает высоким быстродействием и способен выделять вложенные кластеры сложной формы, что обеспечивает высокое качество кластеризации.

Для классификации полученных кластеров был выбран ансамблевый алгоритм случайного леса [8]. Благодаря обучению каждого дерева на случайном подмножестве признаков и объектов выборки, а также формированию итогового результата по принципу мажоритарного голосования всех обученных деревьев, алгоритм демонстрирует устойчивость к переобучению и выбросам, что критически важно при высокой изменчивости отражательной способности воды, характерных для паводков. В качестве программной реализации использовался класс `RandomForestClassifier` из библиотеки `sklearn` [9].

Общая схема предлагаемого алгоритма включает следующие этапы.

1. Подготовка данных – извлечение спектральных характеристик R, G, B, NIR, SWIR1, SWIR2 и спектральных индексов NDWI [10], NDVI, MNDWI [11].

2. Формирование масок облачности с помощью алгоритма `OmniCloudMask`. В качестве входных признаков используются каналы R, G, NIR.

3. Линейное растяжение признаков NDWI, NDVI, MNDWI с обрезкой 1 % значений с краев гистограммы. Растяжение производится независимо для каждой из 4 масок облачности, полученных от алгоритма `OmniCloudMask`.

4. Кластеризация алгоритмом НСА (гиперпараметры $m: 48$, $t: 0,99$) каждой из 4 зон снимка по отдельности.

5. Вычисление средних значений спектральных признаков R, G, B, NIR, SWIR1, SWIR2 для каждого полученного кластера.

6. Классификация кластеров алгоритмом случайного леса по вычисленным средним значениям спектральных признаков. Формирование полномасштабной карты предсказаний, содержащей 3 класса – вода, не вода и не валидная область.

7. Постобработка медианным фильтром с размером окна 5×5 для сглаживания резких выступов и фильтрации шума на итоговой карте.

Для обучения алгоритма классификации разработана следующая методика:

1. Для каждого снимка из обучающего набора `WorldFloods` производится кластеризация и расчет статистик кластеров по ранее описанным правилам.

2. Для каждого кластера вычисляется метрика «чистоты» (Purity), определяемая как доля пикселей доминирующего класса.

3. После формирования итоговой таблицы, содержащей метрику «чистоты» и средние значения спектральных характеристик каждого кластера, выполняется фильтрация по критерию «чистоты» больше 80%.

4. Полученная таблица данных используется для обучения алгоритма классификации.

Результаты

По результатам анализа графиков распределения значений яркости пикселей

водных и не водных объектов (рис. 1) на примере признака NDWI было установлено, что при попадании объекта в зону затемнения его спектральные характеристики претерпевают качественные изменения, причем для разных типов объектов эти изменения имеют разный характер. Это обстоятельство обуславливает необходимость раздельного анализа освещенной и затемненной зон.

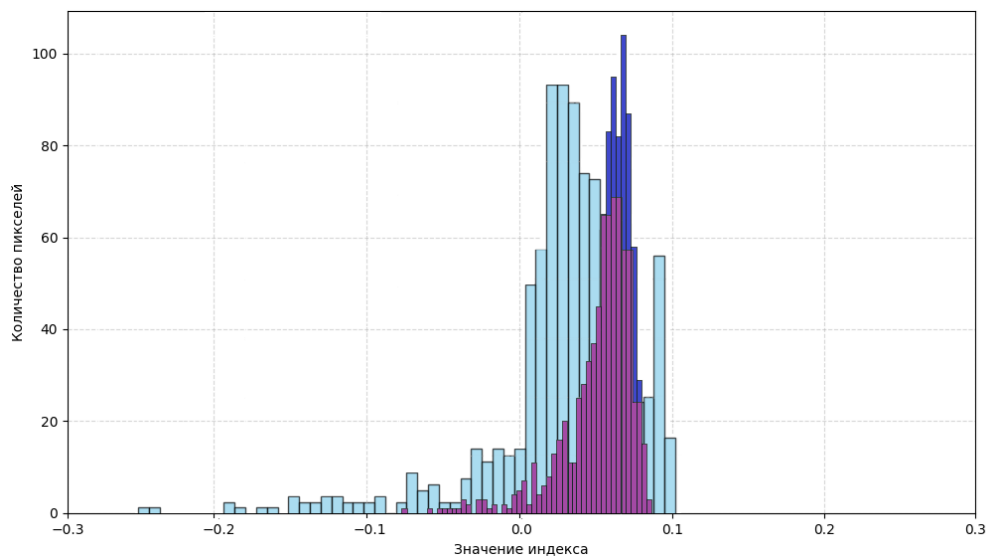


Рис. 1. Гистограммы распределения значений индекса NDWI для водного объекта. Голубым цветом – значение характеристики на освещенном участке, темно-синим – на затемненном, фиолетовым – пересечение распределений

Для оценки качества работы созданного алгоритма были выбраны метрики Precision, Recall, IoU для водного класса. Так как снимки тестовой выборки содержали сцены, которые могли относиться к одному и тому же паводковому явлению в разные моменты времени, для оценки качества использовались две стратегии усреднения – по всем снимкам сразу или с предварительным усреднением метрик для снимков из одного района съемки с последующим глобальным усреднением.

В табл. 1 представлены достигнутые предложенным алгоритмом метрики качества на тестовом наборе данных WorldFloods. В табл. 2 представлены метрики, достигнутые моделями глубокого обучения, предложенными авторами датасета [3]. Как следует из полученных результатов, разработанный алгоритм демонстрирует высокое качество работы в сложных условиях, сопоставимое с современными нейросетевыми моделями.

По результатам визуального анализа снимков, предоставленных «НИЦ Планета», был установлен перечень проблемных зон, в которых предложенный алгоритм допускает ошибки. Среди них можно выделить зоны с частично прозрачными облаками (рис. 2), с низкой яркостью водных объектов, где вода не выделяется в полном объеме, а также гористую и холмистую (рис. 3) местность, в которых алгоритм дает ложноположительные срабатывания на тени от рельефа.

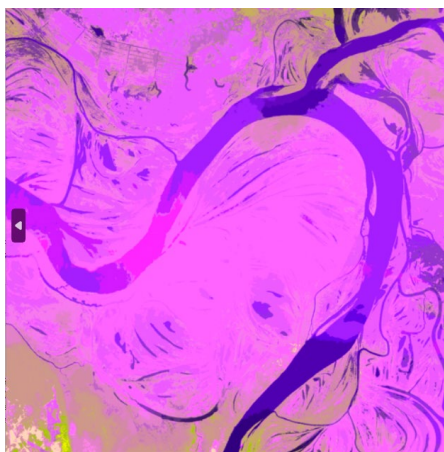


Рис. 2. Слева – композит из R-NIR-NDVI каналов, справа – результат предсказания алгоритма поверх RGB-композиата (вода выделена ярко-синим)

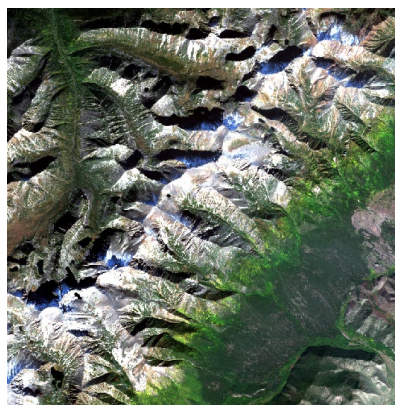


Рис. 3. Слева – RGB композит снимка в гористой местности, справа – результат выделения воды алгоритмом (белым цветом выделяется вода)

Заключение

В работе рассмотрена задача автоматического выделения водной поверхности на мультиспектральных спутниковых снимках Sentinel-2 в условиях паводков и повышенной облачности. Предложен гибридный алгоритм, сочетающий предварительное выделение зон облачности и теней, кластеризацию по спектральным индексам NDWI, NDVI, MNDWI и последующую классификацию кластеров с использованием ансамблевого алгоритма случайного леса.

В ходе исследования было показано, что спектральные характеристики объектов в освещённых и теневых областях сцены существенно различаются. Это обуславливает необходимость их независимой обработки при решении задачи сегментации водной поверхности. Для предварительного выделения зон различной освещённости использовался алгоритм OmniCloudMask, формирующий маски облаков и теней от облаков.

Экспериментальная оценка на наборе данных WorldFloods показала, что

предложенный алгоритм обеспечивает высокое качество выделения водной поверхности в сложных условиях съёмки, включая наличие облачности, теней и низкого контраста сцен. Полученные результаты оказались сопоставимы с результатами современных нейросетевых моделей, предложенных авторами набора данных. Алгоритм продемонстрировал высокую скорость обработки полноразмерных сцен Sentinel-2, что делает его пригодным для оперативного мониторинга паводковой ситуации.

По результатам визуального анализа были выявлены основные проблемные случаи, приводящие к ошибкам сегментации: частично прозрачная облачность, низкая яркость водных объектов и гористая местность, в которой возможны ложноположительные срабатывания на тенях от рельефа.

В дальнейшем планируется исследование альтернативных вариантов нормализации данных в освещенных, затемненных и облачных зонах, а также оптимизация алгоритма для уменьшения потребления ресурсов.

Таблица 1. Оценки качества работы предложенного алгоритма на наборе WorldFloods

Метрика	Усреднение по снимкам	Усреднение по событиям
По всему снимку		
IoU	91,46	91,29
Precision	92,63	92,46
Recall	98,24	98,32
По облачной зоне		
IoU	90,45	91,81
Precision	90,56	91,92
Recall	99,78	99,81
По зоне без облачности		
IoU	91,76	91,52
Precision	92,98	92,74
Recall	98,20	98,29

Таблица 2. Оценки качества работы моделей, предложенных авторами набора WorldFloods

Архитектура	Вид модели	Recall	Precision	IoU
Unet	Multiclass	92,23	91,51	85,29
Unet	Multioutput binary	94,19	86,94	82,45
Unet S2-to-L8	Multioutput binary	96,35	86,18	83,49
Unet RGBNIR	Multioutput binary	93,50	84,33	79,86

Благодарности

Авторы выражают благодарность Сибирскому центру ФГБУ «НИЦ «Планта» за предоставленные спутниковые снимки и комментарии по их обработке.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Portalés-Julià E., Mateo-García G., Gómez-Chova L. Understanding Flood Detection Models Across Sentinel-1 and Sentinel-2 Modalities and Benchmark Datasets. 2025. [Электронный ресурс]. URL: <https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=5118486> (дата обращения: 20.05.2025).
2. Amitrano D., Martino G.Di, Simone A.Di, Imperatore P. Flood detection with SAR: A review of techniques and datasets // Remote Sensing. – 2024. – Vol. 16. – No. 4. – P. 656.
3. Portalés-Julià E. et al. Global flood extent segmentation in optical satellite images // Scientific reports. – 2023. – Т. 13. – №. 1. – С. 20316.
4. Wright N. et al. Training sensor-agnostic deep learning models for remote sensing: Achieving state-of-the-art cloud and cloud shadow identification with OmniCloudMask // Remote Sensing of Environment. – 2025. – Т. 322. – С. 114694.
5. Рылов С.А., Новгородцева О.Г., Пестунов И.А., Дубровская О.А. Технология обработки данных с космических аппаратов "Канопус-В", "Ресурс-П" и "Метеор-М" для мониторинга и картографирования паводковой ситуации // Региональные проблемы дистанционного зондирования Земли : Материалы II Международной научной конференции, Красноярск, 22–25 сентября 2015 года. – Красноярск: СУФ, 2015. – С. 207-212.
6. Rylov S.A., Pestunov I.A. Fast hierarchical clustering of multispectral images and its implementation on NVIDIA GPU // Journal of Physics: Conference Series (JPCS). – 2018. – Vol. 1096. – P. 012039.
7. Сиденко М. Е., Рылов С. А. Влияние спектральных характеристик и методов предобработки спутниковых снимков при выделении водной поверхности // Интерэкспо Гео-Сибирь. – 2025. – Т. 4. – С. 51.
8. Breiman L. Random forests // Machine learning. – 2001. – Т. 45. – №. 1. – С. 5-32.
9. Pedregosa F. et al. Scikit-learn: Machine learning in Python // the Journal of machine Learning research. – 2011. – Т. 12. – С. 2825-2830.
10. McFeeters, S.K. Using the Normalized Difference Water Index (NDWI) within a geographic information system to detect swimming pools for mosquito abatement: a practical approach // Remote Sensing. – 2013. – Vol. 5. – No. 7. – P. 3544-3561.
11. Xu H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery // International journal of remote sensing. – 2006. – Т. 27. – №. 14. – С. 3025-3033.

© М. Е. Сиденко, С. А. Рылов, 2026