

Использование методов машинного обучения для радарно-оптического синтеза серий безоблачных спутниковых изображений высокого пространственного и временного разрешения

Д. Е. Плотников¹, Ц. Чжоу²

¹ *Институт космических исследований РАН, Москва, 117997, Россия
E-mail: dmitplot@d902.iki.rssi.ru*

² *Московский государственный университет имени М. В. Ломоносова
E-mail: ziyaozhou48@gmail.com*

Исследован потенциал метода машинного обучения Random Forest для решения задачи синтеза безоблачных изображений нормализованного относительного индекса растительности NDVI (*англ.* Normalized Difference Vegetation Index) на основе совместного применения радарных данных С-диапазона Sentinel-1 и спутниковой системы оптической съёмки MSI (*англ.* Multispectral Instrument) Sentinel-2. В рамках исследования был разработан метод построения и оптимизации модели для кросс-сенсорного синтеза безоблачных значений NDVI на основе ряда радарных, топографических и тематических предикторов, построенных для синхронных пар спутниковых снимков Sentinel-1 и Sentinel-2, полученных за вегетационный период 2021 г. на территорию Калининградской обл. В процессе оптимизации модели использовалось рекурсивное исключение малоинформативных признаков и кросс-валидация, а для поиска наилучшей комбинации параметров модели — сеточный поиск. В результате работы модели на основе только радарных измерений были сформированы ряды безоблачных изображений NDVI с высоким пространственным и временным разрешением на всю территорию Калининградской обл. На эффективность разработанного подхода указывают высокие значения характеристик точности модели, полученные по тестовому набору: средняя абсолютная ошибка прогнозирования MAE (*англ.* mean absolute error) NDVI равна 0,08, среднеквадратическое отклонение RMSE (*англ.* root mean squared error) составляет 0,12, коэффициент корреляции R равен 0,78 и индекс Уиллмота WIA (*англ.* Willmot Index of Agreement) — 0,87. Полученные результаты показывают, что Random Forest имеет значительный потенциал для применения в кросс-сенсорном синтезе безоблачных изображений для восстановления временных рядов NDVI.

Ключевые слова: кросс-сенсорное синтезирование, Sentinel-1, Sentinel-2, восстановление временных рядов, машинное обучение, кросс-валидация, recursive feature elimination, сеточный поиск

Одобрена к печати: 22.07.2024

DOI: 10.21046/2070-7401-2024-21-4-33-46

Введение

Временные серии дистанционных измерений характеристик растительного покрова незаменимы при спутниковом картографировании типов растительности и оценки её состояния (Лупян и др. 2018; Плотников и др., 2020; Савин и др. 2011; Середа и др. 2020; Шабанов и др., 2018; Waldner et al., 2019; Wu et al., 2023). Ввиду высокой динамичности и мозаичности агроландшафтов, при дистанционном мониторинге сельскохозяйственной растительности особый интерес представляют серии спутниковых наблюдений высокого временного и пространственного разрешения.

После работы (Gao et al., 2006), посвящённой новаторскому методу синтеза высокодетальных измерений оптического диапазона длин волн на основе данных Landsat (TM/ETM+/OLI, *англ.* Thematic Mapper/Enhanced Thematic Mapper Plus/Operational Land Imager) и Terra/Aqua (MODIS, *англ.* Moderate resolution Imaging Spectroradiometer) путём взвешенной комбинации спектральной, временной и координатной компонент измерений (метод STARFM, *англ.* Spatial and Temporal Adaptive Reflectance Fusion Model), дальнейшие исследования были

нацелены на универсализацию и повышение эффективности работы семейства этих методов (Müller et al., 2015; Wu et al., 2015, 2016). Параллельно развивались методы восстановления серий высокого пространственного и временного разрешения на основе их моделирования (Roy, Yan, 2018), пространственно-временного анализа (Плотников и др., 2018а; Yan, Roy, 2018) и геостатистического кригинга (Wang et al., 2016). Было показано, что в результате применения указанных методов восстановления могут быть построены региональные карты растительного покрова и используемых пахотных земель с точностью не ниже типичных значений для такого рода задач (Гаврилюк и др., 2018; Плотников и др., 2018б). Мультисенсорный разноплатформенный синтез данных оптического диапазона длин волн во многом утратил актуальность после начала распространения свободно доступных данных Sentinel-2 (MSI) уровня L2A программы ESA (*англ.* European Space Agency) Copernicus в конце 2018 – начале 2019 года (<https://scihub.copernicus.eu/>). С этого момента восстановление сезонных и многолетних безоблачных серий ежедневных измерений высокого пространственного разрешения стало возможным на базе ранее разработанных универсальных методов интерполяции, использующих данные одной спутниковой платформы, включая Sentinel-2 (MSI), «Метеор-М» (КМСС — комплекс многозональной спутниковой съёмки) и другие (Plotnikov et al., 2022).

В настоящее время современными созвездиями спутниковых систем (например, Dove/RapidEye/SkySat коммерческой компании Planet Labs) обеспечивается возможность получения ежедневных высокодетальных (лучше 1 м) спутниковых наблюдений. Тем не менее, в случаях перманентного отсутствия данных видимого диапазона длин волн по причине слишком длительного экранирования земной поверхности сплошной облачностью мультисенсорный синтез оптических данных, а тем более и классические методы восстановления временных серий на основе данных одной платформы или созвездия, практически полностью теряют свою эффективность. При этом основанные на других физических принципах радарные спутниковые данные обеспечивают возможность получения информации о характеристиках растительного покрова ввиду проницаемости облачного покрова для радиолокационного излучения и независимости от условий солнечного освещения подстилающей поверхности. Росту числа локальных исследований на тему совместного использования радарных и мультиспектральных оптических измерений высокого пространственного разрешения способствовало дальнейшее развитие программы Copernicus, предоставившей в открытый доступ данные со спутников серий Sentinel-1 и Sentinel-2 с максимальной из свободно доступных альтернатив пространственной детализацией 10–20 м на пиксель.

Следует отметить, что и радарные данные имеют ряд затрудняющих их использование особенностей, таких как наличие спекл-шума, а также влияние геометрии подстилающей поверхности на характеристики отражённого сигнала, и, таким образом, требуют тщательной предварительной обработки. Ввиду значительных объёмов высокодетальных данных и ресурсов на их обработку, проекты по решению задач спутникового мониторинга глобального и субконтинентального охвата поддерживаются специализированными центрами хранения и распространения спутниковой информации, такими как NASA (*англ.* National Aeronautics and Space Administration), ESA, DLR (*нем.* Deutsches Zentrum für Luft- und Raumfahrt) и других на основе методов обработки больших данных и машинного обучения. Кроме этого, ряд задач мониторинга широкого пространственного охвата могут быть решены с использованием платформ облачных вычислений, таких как Google Earth Engine (Gorelick et al., 2017), однако такие возможности имеются и у относительно крупных коммерческих компаний и научных центров коллективного пользования, включая Центр коллективного пользования «ИКИ-Мониторинг» (Лупян и др., 2019).

Возможности комбинирования мультиспектральной оптической съёмки с радарными данными (Van Tricht et al., 2018; Veloso et al., 2017; Zhou et al., 2017) особенно востребованы при дистанционном мониторинге регионов мира с перманентным облачным покровом, в том числе высокоширотных и тропических зон, а также регионов с морским климатом. Так, радарные данные спутниковой системы ASAR (*англ.* Advanced Synthetic Aperture Radar) Envisat (С-диапазон) применялись для картографирования сахарного тростника в тропическом Китае и дистанционного определения индекса листовой поверхности LAI (*англ.* Leaf Area

Index) (Lin et al., 2009). В работе (Alvarez-Mozos et al., 2021) была продемонстрирована тесная связь между значениями NDVI и различными радиолокационными индексами. Ряд новых радарных индексов, включая DPSVIm (*англ.* Dual Polarization SAR Vegetation Index modified), были разработаны и успешно использовались в качестве предикторов в различных методах машинного обучения для реконструкции значений NDVI растительного покрова в работе (Dos Santos et al., 2022), где лучшим алгоритмом для синтеза оказался метод случайного леса.

Дистанционный мониторинг растительного покрова ряда российских регионов, включая северо-западную и дальневосточную часть РФ, также зачастую затруднён из-за местных климатических особенностей, что приводит к снижению частоты безоблачных высокодетальных наблюдений земной поверхности в оптическом диапазоне длин волн. Низкая частота безоблачных наблюдений препятствует получению ключевой фенологической информации из-за невозможности устойчивого формирования рядов необходимого временного разрешения. В настоящей работе продемонстрирован потенциал радарных данных С-диапазона Sentinel-1 и спутниковой системы Sentinel-2 (MSI) для синтеза временных рядов безоблачных значений вегетационного индекса NDVI с пространственным разрешением 20 м на территорию Калининградской обл.

Тестовый регион и спутниковые данные

Калининградская обл. расположена в бассейне р. Преголя и выходит к побережью Балтийского моря на западе и северо-западе (*рис. 1*). В регионе преобладают низкие равнины, болота и пологие холмы, на 20 % территории расположены смешанные леса. Климат влажный морской с годовым количеством осадков 600–700 мм, среднегодовой уровень облачности более 65 %, а среднегодовое число ясных дней в региональном центре не превышает 35. Земли сельскохозяйственного назначения характеризуются высокой мозаичностью агроландшафта и составляют более половины площади области. Среднее число безоблачных наблюдений любого участка земной поверхности за полгода (в период с мая по октябрь) при использовании спутниковой системы Sentinel-2 не превышает 10.

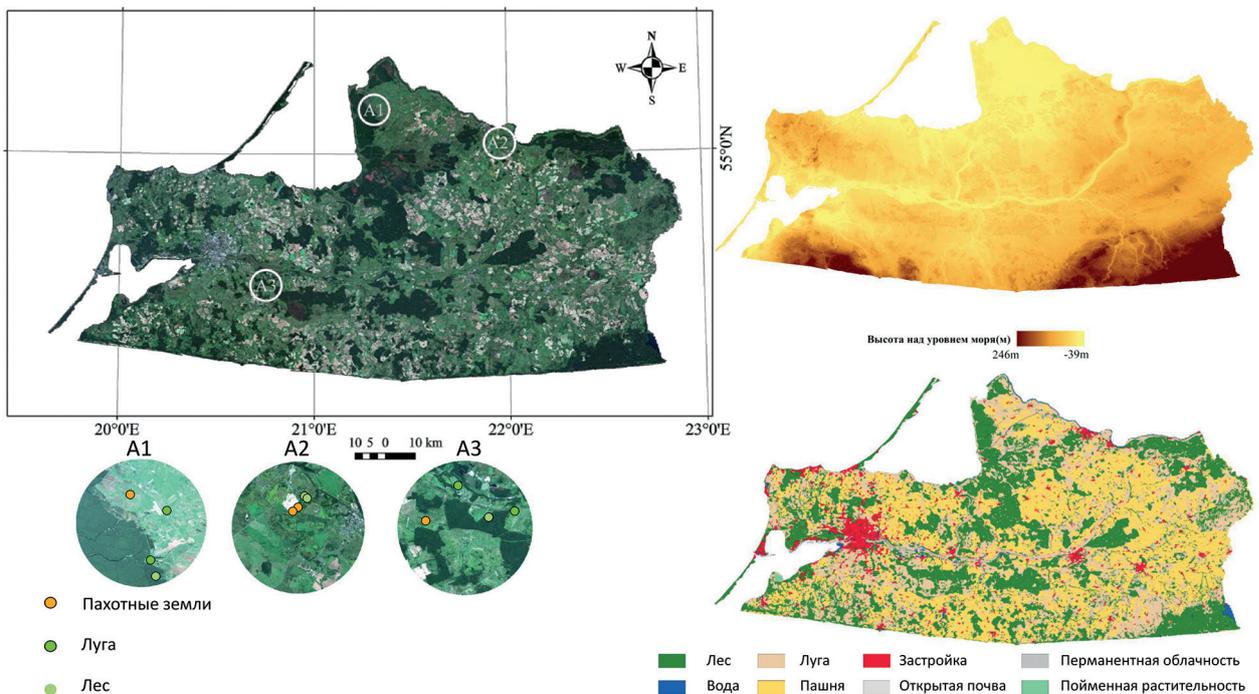


Рис. 1. Спутниковое изображение Калининградской обл., тестовых участков А1–А3 с положением точек внутри зон для визуального анализа временных серий, карта рельефа и карта растительного покрова Калининградской обл.

Для оперативного спутникового мониторинга растительного покрова региона, включая дистанционную оценку быстроменяющейся сельскохозяйственной растительности на основе вегетационных индексов, часто требуется использование серий высокого временного и пространственного разрешения.

В нашем исследовании использовались 22 отдельных изображения Sentinel-1 уровня GRD (*англ.* Ground Range Detected) и 75 изображений Sentinel-2 уровня L2A — всего 97 спутниковых снимков всей территории Калининградской обл., охватывающих временной диапазон с мая по октябрь 2021 г. Кроме этого, в качестве признаков были использованы: цифровая карта высот DEM (*англ.* Digital Elevation Model) миссии SRTM (*англ.* Shuttle Radar Topographic Mission) с пространственным разрешением 30 м и карта растительного покрова LULC (*англ.* Land Use Land Cover) от ESRI (*англ.* Environmental Systems Research Institute) с пространственным разрешением 10 м за 2021 г., построенная на основе спутниковых данных Sentinel-2 (Karra et al., 2021).

Для детальной оценки результатов восстановления временных серий были выбраны три неоднородных (включающих сразу несколько тематических классов карты растительного покрова) тестовых участка A1, A2 и A3 на севере и западе области (см. *рис. 1*) размером около 10 км, для которых были построены временные ряды всех исходных и синтезированных изображений NDVI с акцентом на трёх классах растительного покрова — пахотные земли, луга и лесная растительность.

Для визуального анализа временных рядов NDVI внутри каждого участка случайным образом были выбраны пиксели указанных типов растительного покрова, для которых были построены исходные безоблачные (по маскам SCL, *англ.* Scene Classification Land) и синтезированные временные ряды NDVI.

Методы обработки спутниковых данных и построение модели

Для снижения влияния спекл-шума и рельефа местности на радарные измерения Sentinel-1 уровня L1 была проведена калибровка, спекл-фильтрация (фильтром Ли с окном 7×7) и коррекция (*англ.* terrain correction) спутниковых данных с получением коэффициентов обратного рассеяния. С использованием данных поляризаций Sentinel-1 были вычислены следующие индексы: радиолокационный индекс растительности RVI (*англ.* Radar Vegetation Index), $RVI = 4VH / (VV + VH)$; индекс вертикальной двойной деполаризации VDDPI (*англ.* Vertical Dual De-polarization Index), $VDDPI = (VV + VH) / VV$; нормализованный индекс разностной поляризации NDPI (*англ.* Normalized Differential Phenology Index), $NDPI = (VV - VH) / (VV + VH)$; улучшенный биполярный индекс растительности $DPSVI_m = (VV^2 + VV \cdot VH) / \sqrt{2}$, где VV и VH — коэффициенты обратного рассеяния для вертикально-вертикальной и вертикально-горизонтальной поляризации. Кроме этого, в качестве предикторов использовались метрики отношения (VV/VH) и суммы $(VV + VH)$ коэффициентов обратного рассеяния на соответствующих поляризациях. Указанные индексы часто используются для решения задач дистанционной оценки сельскохозяйственной растительности (Dos Santos et al., 2021; Mishra et al., 2022), являющейся в настоящем исследовании региональной доминантой среди остальных классов.

Предварительная обработка данных прибора MSI (Sentinel-2) заключалась в маскировании облачности и теней с использованием масок мешающих факторов SCL для продукта L2A и расчёте вегетационного индекса NDVI.

Поскольку исходное пространственное разрешение данных спутниковых систем Sentinel-1 и Sentinel-2 неодинаково, а геометрия пикселей Sentinel-1 отличалась от геометрии Sentinel-2, был выполнен ресемплинг исходных изображений в проекцию UTM (*англ.* Universal Transverse Mercator) на основе бикубической интерполяции с пространственным разрешением 20 м.

В качестве исследуемого метода машинного обучения для создания модели был использован метод случайного леса (*англ.* random forest) (Breiman, 2001), реализованный на языке Python в библиотеке sklearn. При построении признакового пространства для модели прогно-

зирования NDVI на основе случайного леса был сформирован набор из 13 исходных признаков, а именно: радарные поляризации и комбинации VV, VH, VV/VH, VV+VH, радарные индексы RVI, VDDPI, NDPI, DPSVIm, а также цифровая модель рельефа SRTM и карта растительного покрова ESRI (LULC), географические координаты (*англ.* lon, lat) и дата (*англ.* time).

Для сокращения разницы во времени между моментами съёмки системами Sentinel-1 и Sentinel-2 в целях минимизации влияния фенологии и погодных условий на точность модели были выбраны 15 «синхронных» пар изображений Sentinel-2/Sentinel-1, даты получения которых отличались не более чем на четыре дня. Пространственный шаг регулярной сети выборки был задан 500×500 м, что позволило достаточно полно описать изменчивость фрагментированной территории исследования при разумном объёме обучающих данных (см. *рис. 2б*). Всего было собрано 307 221 точек выборки в 15 парах синхронных изображений Sentinel-2/Sentinel-1.

Для оценки эффективности модели использовались следующие метрики: среднеквадратическое отклонение RMSE (*англ.* root mean squared error), коэффициент корреляции R , средняя абсолютная ошибка MAE (*англ.* mean absolute error) и индекс Уиллмота WIA (*англ.* Willmot Index of Agreement).

Индекс WIA рассчитывается следующим образом:

$$WIA = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (|P_i - \bar{O}| + |O_i - \bar{O}|)^2}, \quad \bar{O} = \sum_{i=1}^n \frac{O_i}{n}, \quad 0 \leq WIA \leq 1,$$

где обозначение O относится к наблюдаемым значениям, P — к прогнозируемым значениям, \bar{O} — среднее наблюдаемое значение. Индекс WIA принимает значение от 0 до 1, при этом равное единице значение WIA означает, что расчётное значение полностью совпадает с фактическим, а нулю — полное несовпадение. Обычно считается, что модель имеет практическую предсказательную ценность, если WIA превышает 0,6.

Выборка случайным образом была поделена на две части — 70 % использовалось в качестве обучающего набора и 30 % — в качестве тестового набора. Чтобы уменьшить размерность пространства и для нахождения наилучшей комбинации признаков, использовалось рекурсивное исключение признаков RFE (*англ.* Recursive Feature Elimination) и 10-кратная кросс-валидация, а для поиска наилучшей комбинации параметров модели (параметризации) — сеточный поиск и 10-кратная кросс-валидация. Затем было проведено обучение на обучающем наборе по найденному наилучшему сочетанию полученных признаков и параметров модели с получением предварительной модели, которая оценивалась с помощью тестового набора. На последнем этапе оптимальная комбинация признаков и полный объём выборки были использованы для обучения окончательной параметризованной модели, которая была использована для синтеза безоблачных изображений NDVI для всей области исследования на каждую целевую дату.

Результаты

Рекурсивное исключение признаков методом RFE выполнялось итеративно, начиная с исходного набора из 13 исходных признаков, на каждой итерации осуществлялось построение модели, которая 10-кратно кросс-валидировалась с расчётом метрик MAE, RMSE, R^2 и WIA, после чего исключался наименее значимый признак и процедура повторялась. Распределение значимости исходного набора признаков и динамика показателей точности в процессе рекурсивного исключения признаков приведены на *рис. 2* (см. с. 38).

В результате работы RFE было установлено оптимальное сочетание признаков, достигнутое на четвёртой итерации (см. экстремумы метрик на *рис. 2, справа*), ими оказались: VH, VH+VV, VH/VV, NDPI, DEM, time, LULC, lon, lat (значимость девяти оставшихся признаков приведена на *рис. 3, слева*, см. с. 38).

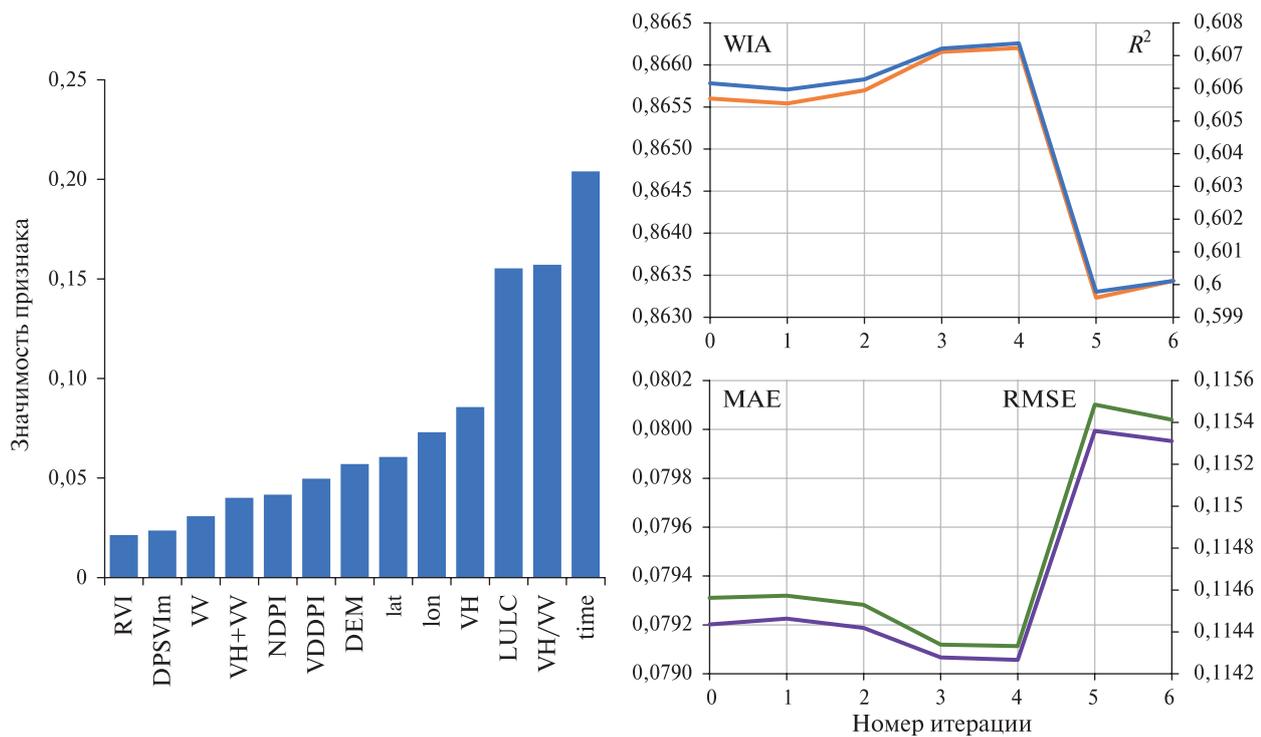


Рис. 2. Распределение значимости исходного набора из 13 признаков (слева) и динамика показателей точности модели между итерациями RFE (справа)

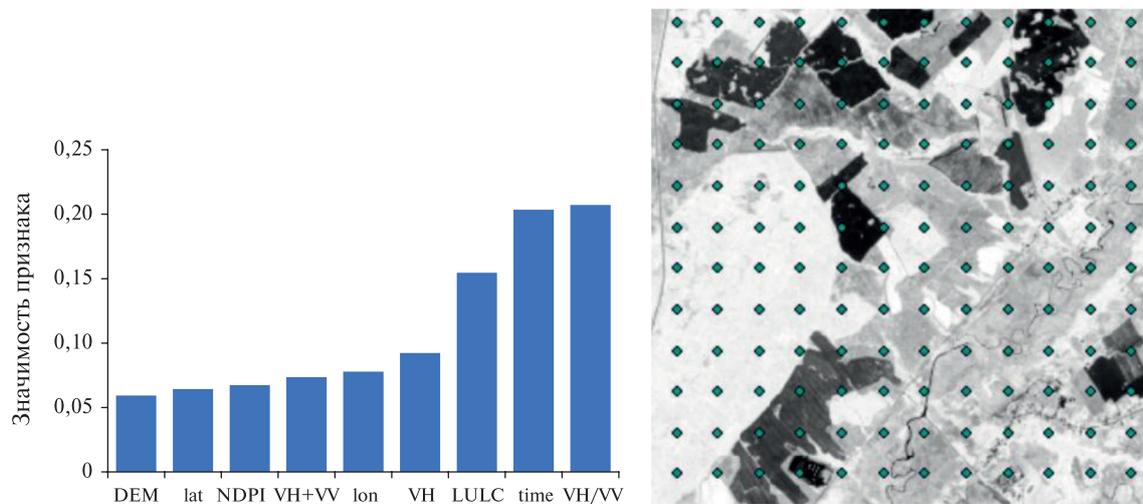


Рис. 3. Фрагмент сети регулярного сэмплинга с шагом 500 м на фоне изображения NDVI объектов земной поверхности по данным MSI (справа) и распределение значимости оставшихся признаков в наилучшей модели (слева)

Параметризация модели на сетке дала следующие значения параметров случайного леса в терминах библиотеки sklearn: $n_estimators$ (число деревьев в случайном лесу) = 1000, $min_samples_leaf$ (минимальное число сэмплов в финальном узле) = 4, $min_samples_split$ (минимальное число сэмплов в промежуточном узле) = 10.

Диаграмма рассеяния в виде тепловой карты для исходных значений NDVI по тестовому набору и модельной оценкой NDVI приведена на рис. 4 (см. с. 39) вместе с основными характеристиками полученной модели, включая среднеквадратическое отклонение ($RMSE = 0,12$), коэффициент корреляции ($R = 0,78$), среднюю абсолютную ошибку ($MAE = 0,08$) и индекс Уиллмота ($WIA = 0,87$). Можно увидеть, что подавляющее число изме-

рений за полугодовой период исследования смещено в область высоких значений и среднее по сезону и территории значение NDVI превышает 0,7.

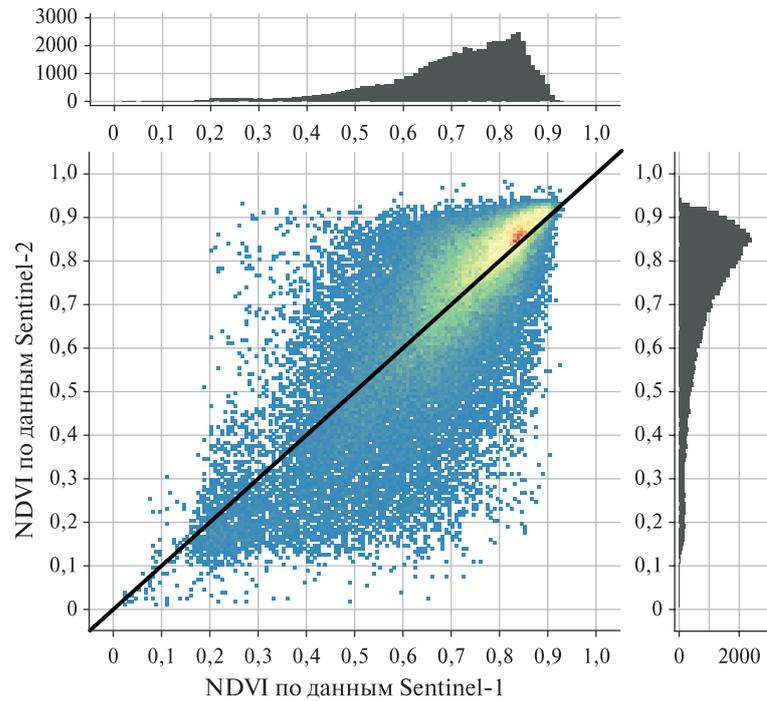


Рис. 4. Диаграмма рассеяния для исходных (по данным Sentinel-2) и прогнозируемых (по данным Sentinel-1) значений NDVI, полученная по тестовому набору обученной модели, одномерные гистограммы NDVI по каждому источнику по отдельности приведены сверху и справа

С помощью окончательной модели были синтезированы изображения NDVI для некоторых изображений Калининградской обл. по данным Sentinel-1. Пример результатов синтеза изображений NDVI всего региона исследования для трёх синхронных дат приведён на рис. 5. Можно отметить, что синтезированные изображения корректно передают пространственные градиенты: устойчиво высокие значения NDVI для участков лесной растительности в центре, на северо-востоке и на юго-востоке области и низкие значения NDVI для населённых пунктов и других урбанизированных территорий.

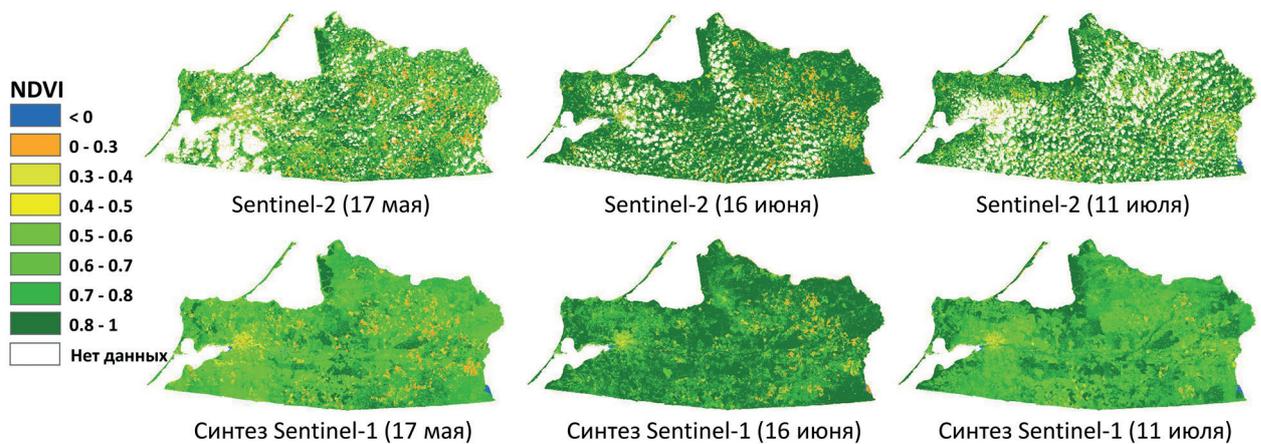


Рис. 5. Пример результатов работы модели на всю территорию Калининградской обл. за три различных месяца: реконструированные безоблачные изображения NDVI по данным Sentinel-1 (внизу), изображения NDVI по данным Sentinel-2 после маскирования мешающих факторов (вверху)

Ниже (рис. 6) приведена серия всех исходных и восстановленных изображений NDVI на территорию трёх тестовых зон (A1, A2, A3) для 15 пар Sentinel-2/Sentinel-1 с различием дат съёмки не более четырёх дней. Красным цветом отмечены области маскирования мешающих факторов, пары сгруппированы в три блока, соответствующих тестовым зонам A1–A3, где в верхней части блока приведены изображения NDVI по данным Sentinel-2, а в нижней — синтезированные измерения по данным модели и измерений Sentinel-1.

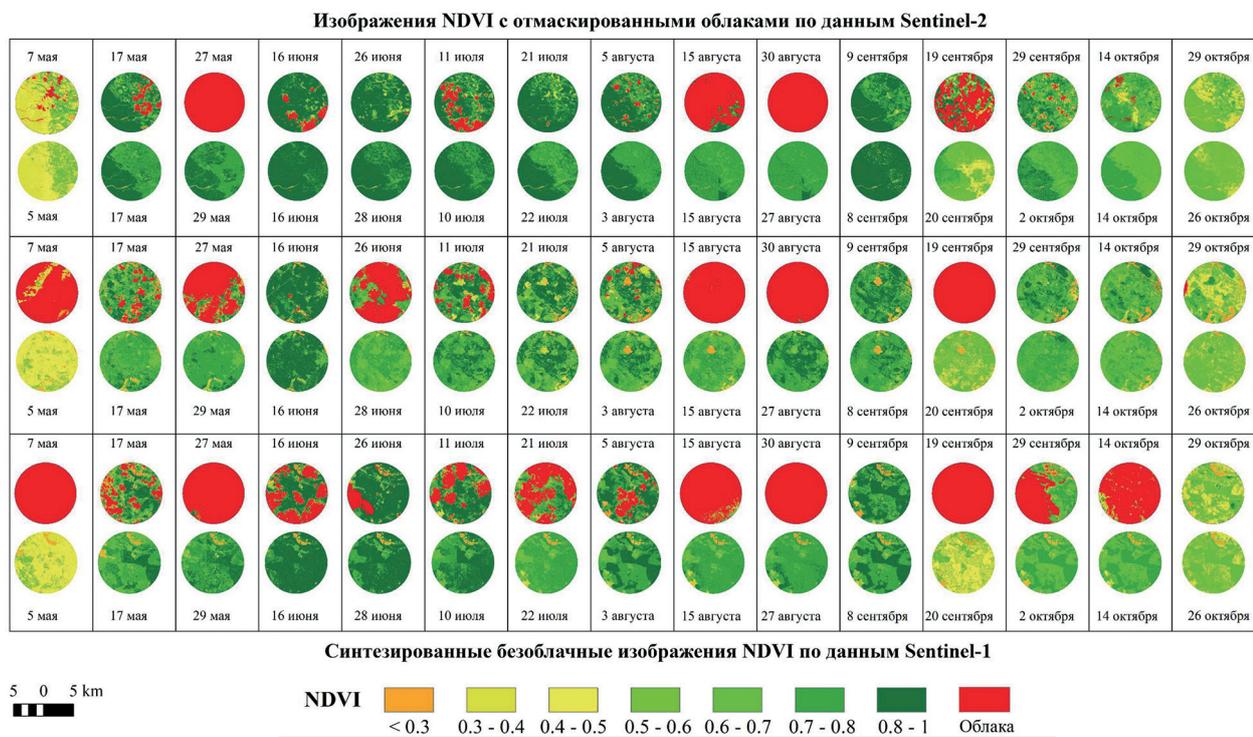


Рис. 6. Серии 15 пар исходных (по данным Sentinel-2, верхние ряды блоков) и синтезированных (по данным Sentinel-1, нижние ряды блоков) изображений NDVI для тестовых зон A1 (вверху), A2 (посередине), A3 (внизу) с различием дат съёмки не более четырёх дней

Внутри тестовых зон случайным образом были выбраны точки, соответствующие трём различным типам растительного покрова, для которых было выполнено построение и визуальный анализ временного ряда синтезированных по данным Sentinel-1 значений NDVI в сравнении с реальными («оптическими») безоблачными (по маске мешающих факторов из слоя SCL) значениями NDVI по данным Sentinel-2 (рис. 7, см. с. 41). Кроме этого, для каждой даты синхронной пары в нижней части рис. 7 приведено число сэмплов обучающей выборки, которое фактически было использовано при построении модели случайного леса.

Рисунки 6 и 7 указывают на высокую долю облачности в тестовых зонах, что отражает ситуацию для всего региона исследования. В целом временные ряды синтезированных значений NDVI хорошо соотносятся с рядами оптических NDVI большую часть анализируемого периода времени, отражая полученные на этапе оценки общей модели значения характеристик точности, в частности достаточно низкие значения ошибок RMSE и MAE (0,12 и 0,08 соответственно).

Тем не менее, на графиках рис. 7 можно заметить особенности во временной серии синтезированных изображений, когда восстановленные значения NDVI отклоняются от ожидаемых значений, например, значения за 29 мая и 20 сентября. В эти даты синтезированное значение отклоняется от тренда, возвращаясь к нему к следующему наблюдению, чего, как ожидается, не должно происходить синхронно для указанных классов, в частности, в осенний период времени. Выбросы в указанные даты связаны с низкой репрезентативностью обучаю-

щей выборки (серые столбики на *рис. 7*), из-за чего регрессионная модель не смогла усвоить достаточно информации о признаках в этой точке. В дальнейшем для минимизации ошибок модели в таких ситуациях может потребоваться динамическое изменение пространственной частоты выборки, а также подключение многолетней информации.

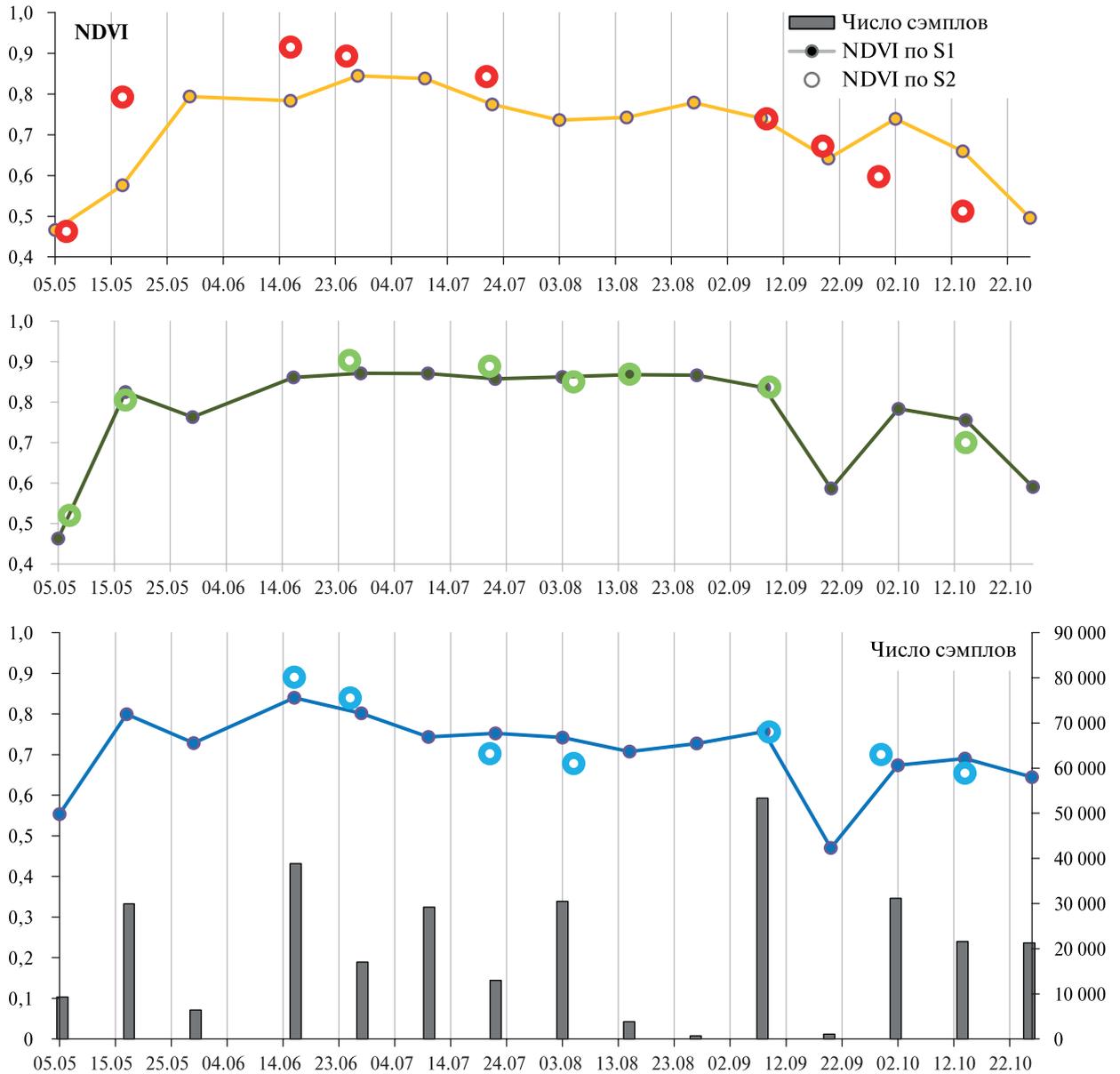


Рис. 7. Временные серии исходных (Sentinel-2, большие выколотые точки) и восстановленных (по данным Sentinel-1, линии) значений NDVI на уровне отдельных пикселей для участков пахотных земель (*вверху*), лесной растительности (*посередине*) и луговой растительности (*внизу*); восстановленные значения за соседние даты соединены прямыми линиями, по вспомогательной оси на нижнем графике серыми столбиками показано число сэмплов, использованных для обучения модели за соответствующие даты

В целом созданная модель отражает наблюдаемые пространственные и временные тренды значений NDVI, а её достоверность и точность коррелирует с доступностью репрезентативной выборки. Полученные результаты говорят о том, что алгоритм случайного леса обладает значительным потенциалом в задаче восстановления временных рядов NDVI высокого временного разрешения.

Выводы и обсуждение

В результате настоящего исследования был разработан алгоритм построения и создана регрессионная модель для синтеза безоблачных изображений NDVI на основе методов машинного обучения и радарных данных спутниковой системы Sentinel-1 на территорию Калининградской обл. Модель позволяет восстанавливать одномоментные изображения NDVI Калининградской обл. с последующим формированием безоблачных серий высокого пространственного и временного разрешения даже в условиях полного отсутствия безоблачных измерений в оптическом диапазоне длин волн, добавляя, таким образом, новую информацию о состоянии и сезонной динамике растительного покрова, которая не может быть получена в этих условиях путём классической интерполяции временных рядов спутниковых данных Sentinel-2. В частности, на приведённых выше примерах (см. *рис. 7*) можно отметить, что при чрезвычайно малом числе исходных безоблачных наблюдений, полученных в течение полугода для класса пахотных земель, лесной растительности и лугов (средняя частота наблюдений — 1,5; 1,33 и 1,16 наблюдений в месяц соответственно), для каждого пикселя было создано 15 новых измерений, что увеличило общее число заведомо безоблачных наблюдений до четырёх в месяц.

Было отмечено, что тренд синтезированных значений для участка пахотных земель на *рис. 7* в осенний период демонстрирует заметные отклонения от тренда реальных значений NDVI, полученных на основе измерений в оптическом диапазоне. Такая особенность наблюдается для некоторых типов сельскохозяйственной растительности и может указывать на неточное описание моделью сценариев увядания некоторых типов растительности, в частности, когда растение лишается хлорофилла, не теряя при этом своей вертикальной архитектуры. Характерным примером могут быть участки созревшего подсолнечника, которые зачастую остаются на полях до поздней осени и могут не убираться вплоть до выпадения снега. Такие участки продолжают соответствовать сильному рассеянию на изображениях Sentinel-1, что обычно характерно для зелёной биомассы. Дальнейшее развитие описанных подходов, включая обучение модели, должно происходить с учётом указанных особенностей фенологии соответствующих типов растительного покрова.

Следует отметить, что в процессе автоматической оптимизации из модели были исключены предикторы на основе радарных индексов VDDPI и DPSVIm, которые часто используются для оценки сельскохозяйственной растительности и достаточно хорошо проявляют себя при её отделении от участков открытой почвы. Однако в данном случае доля выборки, соответствующей открытой почве, ассоциированной с распашкой, оказалась очень мала (среднесезонные значения NDVI за выбранный период превышают 0,7 (см. *рис. 4, 7*) для любого класса), несмотря на значительную долю сельскохозяйственных земель в площади Калининградской обл. Это явление связано с особенностями местного севооборота, преобладанием в этом регионе на участках пахотных земель технических и кормовых культур (многолетних трав), в то время как более половины оставшихся площадей (под зерновыми) соответствовали озимым культурам. Указанные культуры не сеются в текущем сезоне вегетации, а уборка многолетних трав также зачастую не сопровождается полной экспозицией открытой почвы.

Таким образом, для повышения эффективности моделей, которые бы лучше описывали состояние открытой почвы, а также другие особенности фенологии сельскохозяйственных культур, представляется необходимым дальнейшее развитие описанных подходов, включающее использование разносезонных данных, при этом оптимизация модели должна происходить регионально для учёта реального соотношения между типами растительного покрова и их фенологии.

Исследование выполнено в рамках научной темы «Мониторинг», госрегистрация № 122042500031-8.

Литература

1. Гаврилюк Е. А., Плотникова А. С., Плотников Д. Е. Картографирование наземных экосистем Печоро-Илычского заповедника и его окрестностей на основе восстановленных мультивременных спутниковых данных Landsat // Современные проблемы дистанционного зондирования Земли из космоса. 2018. Т. 15. № 5. С. 141–153. DOI: 10.21046/2070-7401-2018-15-5-141-153.
2. Луян Е. А., Барталев С. А., Крашенинникова Ю. С. и др. Анализ развития озимых культур в южных регионах европейской части России весной 2018 года на основе данных дистанционного мониторинга // Современные проблемы дистанционного зондирования Земли из космоса. 2018. Т. 15. № 2. С. 272–276. DOI: 10.21046/2070-7401-2018-15-2-272-276.
3. Луян Е. А., Прошин А. А., Буццев М. А. и др. Опыт эксплуатации и развития центра коллективного пользования системами архивации, обработки и анализа спутниковых данных (ЦКП «ИКИ-Мониторинг») // Современные проблемы дистанционного зондирования Земли из космоса. 2019. Т. 16. № 3. С. 151–170. DOI: 10.21046/2070-7401-2019-16-3-151-170.
4. Плотников Д. Е., Хвостиков С. А., Барталев С. А. (2018а) Метод автоматического распознавания сельскохозяйственных культур на основе спутниковых данных и имитационной модели развития растений // Современные проблемы дистанционного зондирования Земли из космоса. 2018. Т. 15. № 4. С. 131–141. DOI: 10.21046/2070-7401-2018-15-4-131-141.
5. Плотников Д. Е., Колбудаев П. А., Барталев С. А., Луян Е. А. (2018б) Автоматическое распознавание используемых пахотных земель на основе сезонных временных серий восстановленных изображений Landsat // Современные проблемы дистанционного зондирования Земли из космоса. 2018. Т. 15. № 2. С. 112–127. DOI: 10.21046/2070-7401-2018-15-2-112-127.
6. Плотников Д. Е., Ёлкина Е. С., Дунаева Е. А. и др. Развитие метода автоматического распознавания озимых культур на основе спутниковых данных для оценки их состояния на территории Республики Крым // Таврический вестн. аграр. науки. 2020. № 1(21). С. 64–83. DOI: 10.33952/2542-0720-2020-1-21-64-83.
7. Савин И. Ю., Барталев С. А., Луян Е. А., Толпин В. А., Медведева М. А., Плотников Д. Е. Спутниковый мониторинг воздействия засухи на растительность (на примере засухи 2010 года в России) // Современные проблемы дистанционного зондирования Земли из космоса. 2011. Т. 8. № 1. С. 150–162.
8. Середа И. И., Денисов П. В., Трошко К. А. и др. Уникальные условия развития озимых культур, наблюдаемые по данным спутникового мониторинга на европейской территории России в октябре 2020 г. // Современные проблемы дистанционного зондирования Земли из космоса. 2020. Т. 17. № 5. С. 304–310. DOI: 10.21046/2070-7401-2020-17-5-304-310.
9. Шабанов Н. В., Барталев С. А., Ерошенко Ф. В., Плотников Д. Е. Развитие возможностей дистанционной оценки индекса листовой поверхности по данным MODIS // Современные проблемы дистанционного зондирования Земли из космоса. 2018. Т. 15. № 4. С. 166–178. DOI: 10.21046/2070-7401-2018-15-4-166-178.
10. Alvarez-Mozos J., Villanueva J., Arias M., Gonzalez-Audicana M. Correlation between NDVI and Sentinel-1 derived features for maize // 2021 IEEE Intern. Geoscience and Remote Sensing Symp. IGARSS. 2021. P. 6773–6776. DOI: <https://doi.org/10.1109/IGARSS47720.2021.9554099>.
11. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45. No. 1. DOI: 10.1023/A:1010933404324.
12. Dos Santos E. P., Da Silva D. D., Do Amaral C. H. Vegetation cover monitoring in tropical regions using SAR-C dual-polarization index: seasonal and spatial influences // Intern. J. Remote Sensing. 2021. V. 42. Iss. 19. P. 7581–7609. DOI: 10.1080/01431161.2021.1959955.
13. Dos Santos E. P., Da Silva D. D., Do Amaral C. H. et al. A Machine Learning approach to reconstruct cloudy affected vegetation indices imagery via data fusion from Sentinel-1 and Landsat-8 // Computers and Electronics in Agriculture. 2022. V. 194. Article 106753. <https://doi.org/10.1016/j.compag.2022.106753>.
14. Gao F., Masek J., Schwaller M., Hall F. On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance // IEEE Trans. Geoscience and Remote Sensing. 2006. V. 44. No. 8. P. 2207–2218. DOI: 10.1109/TGRS.2006.872081.
15. Gorelick N., Hancher M., Dixon M. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone // Remote Sensing of Environment. 2017. V. 202. P. 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>.
16. Karra K., Kontgis C., Statman-Weil Z. et al. Global land use/land cover with Sentinel-2 and deep learning // 2021 IEEE Intern. Geoscience and Remote Sensing Symp. IGARSS. 2021. P. 4704–4707. DOI: 10.1109/IGARSS47720.2021.9553499.
17. Lin H., Chen J., Pei Z., Zhang S., Hu X. Monitoring sugarcane growth using Envisat ASAR // IEEE Trans. Geoscience and Remote Sensing. 2009. V. 47. No. 8. P. 2572–2579.
18. Mishra D., Pathak G., Singh B. P. et al. Crop classification by using dual-pol SAR vegetation indices derived from Sentinel-1 SAR-C data // Environmental Monitoring Assessment. 2022. V. 195. No. 1. Article 115. DOI: 10.1007/s10661-022-10591-x.

19. Müller H., Rufin P., Griffiths P. et al. Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian savanna landscape // *Remote Sensing of Environment*. 2015. V. 156. P. 490–499. <https://doi.org/10.1016/j.rse.2014.10.014>.
20. Plotnikov D. E., Loupian E. A., Kolbudaev P. A. et al. Daily surface reflectance reconstruction using LOWESS on the example of various satellite systems // 8th Intern. Conf. Information Technology and Nanotechnology (ITNT). 2022. DOI: 10.1109/ITNT55410.2022.9848630.
21. Roy D. P., Yan L. Robust Landsat-based crop time series modeling // *Remote Sensing of Environment*. 2018. V. 238. Article 110810. <https://doi.org/10.1016/j.rse.2018.06.038>.
22. Van Tricht K., Gobin A., Gilliams S., Piccard I. Synergistic use of radar Sentinel-1 and optical Sentinel-2 imagery for crop mapping: a case study for Belgium // *Remote Sensing*. 2018. V. 10. No. 10. Article 1642. DOI: 10.3390/rs10101642.
23. Veloso A., Mermoz S., Bouvet A. et al. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications // *Remote Sensing of Environment*. 2017. V. 199. P. 415–426. <https://doi.org/10.1016/j.rse.2017.07.015>.
24. Waldner F., Schucknecht A., Lesiv M. et al. Conflation of expert and crowd reference data to validate global binary thematic maps // *Remote Sensing of Environment*. 2019. V. 221. P. 235–246. DOI: 10.1016/j.rse.2018.10.039.
25. Wang Q., Shi W., Li Z., Atkinson P. M. Fusion of Sentinel-2 images // *Remote Sensing of Environment*. 2016. V. 187. P. 241–252. <https://doi.org/10.1016/j.rse.2016.10.030>.
26. Wu M., Zhang X., Huang W. et al. Reconstruction of daily 30 m data from HJ CCD, GF-1 WFV, Landsat, and MODIS data for crop monitoring // *Remote Sensing*. 2015. V. 7. No. 12. P. 16293–16314. DOI: 10.3390/rs71215826.
27. Wu M., Wu C., Huang W., Niu Z., Wang C., Li W., Hao P. An improved high spatial and temporal data fusion approach for combining Landsat and MODIS data to generate daily synthetic Landsat imagery // *Information Fusion*. 2016. V. 31. Iss. 31. P. 14–25.
28. Wu B., Zhang M., Zeng H. et al. Challenges and opportunities in remote sensing-based crop monitoring: A review // *National Science Review*. 2023. V. 10. Iss. 4. Article nwac290. DOI: doi.org/10.1093/nsr/nwac290.
29. Yan L., Roy D. P. Large-area gap filling of Landsat reflectance time series by Spectral-Angle-Mapper Based Spatio-Temporal Similarity (SAMSTS) // *Remote Sensing*. 2018. V. 10. No. 4. Article 609. DOI: 10.3390/rs10040609.
30. Zhou T., Pan J., Zhang P. et al. Mapping winter wheat with multi-temporal SAR and optical images in an Urban Agricultural Region // *Sensors*. 2017. V. 17. No. 7. Article 1210. DOI: 10.3390/s17061210.

Leveraging machine learning for radar-optical cloud-free time series synthesis of high spatial and temporal resolution

D. E. Plotnikov¹, Z. Zhou²

¹ *Space Research Institute RAS, Moscow 117997, Russia*
E-mail: dmitplot@d902.iki.rssi.ru

² *Lomonosov Moscow State University, Moscow 119991, Russia*
E-mail: ziyaozhou48@gmail.com

In the paper, we study the potential of the Random Forest machine learning algorithm with respect to the problem of synthesizing cloud-free NDVI images based on the joint use of Sentinel-1 C-band radar and Sentinel-2 optical satellite imaging system MSI (Multispectral Instrument) paired datasets. We developed a method for building and optimizing a machine learning model aimed at cross-sensor synthesis of cloud-free NDVI values based on radar, topographic and thematic predictors constructed for synchronous pairs of Sentinel-1 and Sentinel-2 satellite images acquired during the growing season of 2021. In the process of model optimization, recursive features elimination (RFE) and cross-validation were used, while grid search was used to find the best model parameters combination. As a result of model predictions, time series of cloud-free NDVI images of high spatial and temporal resolution were produced for Kaliningrad region area on the basis of only Sentinel-1 radar data. The effective-

ness of the developed approach is indicated by the high performance metrics of the model obtained from the test set: root mean squared error (RMSE = 0.12), correlation coefficient ($R = 0.78$), mean absolute error (MAE = 0.08) and Willmott Index of Agreement (WIA = 0.87). We demonstrate that Random Forest has significant potential for use in cross-sensor synthesis of cloud-free NDVI time series reconstruction.

Keywords: cross-sensor synthesis, Sentinel-1, Sentinel-2, time series reconstruction, machine learning, cross-validation, recursive feature elimination, grid search

Accepted: 22.07.2024

DOI: 10.21046/2070-7401-2024-21-4-33-46

References

1. Gavrilyuk E. A., Plotnikova A. S., Plotnikov D. E., Land cover mapping of the Pechora-Ilych Nature Reserve and its vicinity based on reconstructed multitemporal Landsat satellite data, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2018, Vol. 15, No. 5, pp. 141–153 (in Russian), DOI: 10.21046/2070-7401-2018-15-5-141-153.
2. Loupian E. A., Bartalev S. A., Krashennikova Yu. S. et al., Analysis of winter crops development in the southern regions of the European part of Russia in spring of 2018 with use of remote monitoring, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2018, Vol. 15, No. 2, pp. 272–276 (in Russian), DOI: 10.21046/2070-7401-2018-15-2-272-276.
3. Loupian E. A., Proshin A. A., Bourtsev M. A. et al., Experience of development and operation of the IKI-Monitoring center for collective use of systems for archiving, processing and analyzing satellite data, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2019, Vol. 16, No. 3, pp. 151–170 (in Russian), DOI: 10.21046/2070-7401-2019-16-3-151-170.
4. Plotnikov D. E., Khvostikov S. A., Bartalev S. A. (2018a), Method for automated crop types mapping using remote sensing data and a plant growth simulation model, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2018, Vol. 15, No. 4, pp. 131–141 (in Russian), DOI: 10.21046/2070-7401-2018-15-4-131-141.
5. Plotnikov D. E., Kolbudaev P. A., Bartalev S. A., Loupian E. A. (2018b), Automated annual cropland mapping from reconstructed time series of Landsat data, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2018, Vol. 15, No. 2, pp. 112–127 (in Russian), DOI: 10.21046/2070-7401-2018-15-2-112-127.
6. Plotnikov D. E., Elkina E. S., Dunaieva Ie. A. et al., Development of the method for automatic winter crops mapping by means of remote sensing aimed at crops state assessment over the Republic of Crimea, *Taurida Herald of the Agrarian Sciences*, 2020, No. 1(21), pp. 64–83 (in Russian), DOI: 10.33952/2542-0720-2020-1-21-64-83.
7. Savin I. Yu., Bartalev S. A., Loupian E. A., Tolpin V. A., Medvedeva M. A., Plotnikov D. E., Satellite monitoring of vegetation affected by drought (using drought 2010 in Russia as an example), *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2011, Vol. 8, No. 1, pp. 150–162 (in Russian).
8. Sereda I. I., Denisov P. V., Troshko K. A. et al., The unique situation of winter crops development observed from remote sensing data in the European territory of Russia in October 2020, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2020, Vol. 17, No. 5, pp. 304–310 (in Russian), DOI: 10.21046/2070-7401-2020-17-5-304-310.
9. Shabanov N. V., Bartalev S. A., Eroshenko F. V., Plotnikov D. E., Development of capabilities for remote sensing estimate of Leaf Area Index from MODIS data, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2018, Vol. 15, No. 4, pp. 166–178 (in Russian), DOI: 10.21046/2070-7401-2018-15-4-166-178.
10. Alvarez-Mozos J., Villanueva J., Arias M., Gonzalez-Audicana M., Correlation between NDVI and Sentinel-1 derived features for maize, *2021 IEEE Intern. Geoscience and Remote Sensing Symp. IGARSS*, 2021, pp. 6773–6776, DOI: <https://doi.org/10.1109/IGARSS47720.2021.9554099>.
11. Breiman L., Random Forests, *Machine Learning*, 2001, Vol. 45, No. 1, DOI: 10.1023/A:1010933404324.
12. Dos Santos E. P., Da Silva D. D., Do Amaral C. H., Vegetation cover monitoring in tropical regions using SAR-C dual-polarization index: seasonal and spatial influences, *Intern. J. Remote Sensing*, 2021, Vol. 42, Issue 19, pp. 7581–7609, DOI: 10.1080/01431161.2021.1959955.
13. Dos Santos E. P., Da Silva D. D., Do Amaral C. H. et al., A Machine Learning approach to reconstruct cloudy affected vegetation indices imagery via data fusion from Sentinel-1 and Landsat-8, *Computers and Electronics in Agriculture*, 2022, Vol. 194, Article 106753, DOI: <https://doi.org/10.1016/j.compag.2022.106753>.

14. Gao F., Masek J., Schwaller M., Hall F., On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance, *IEEE Trans. Geoscience and Remote Sensing*, 2006, Vol. 44, No. 8, P. 2207–2218, DOI: 10.1109/TGRS.2006.872081.
15. Gorelick N., Hancher M., Dixon M. et al., Google Earth Engine: Planetary-scale geospatial analysis for everyone, *Remote Sensing of Environment*, 2017, Vol. 202, pp. 18–27, <https://doi.org/10.1016/j.rse.2017.06.031>.
16. Karra K., Kontgis C., Statman-Weil Z. et al., Global land use/land cover with Sentinel-2 and deep learning, *2021 IEEE Intern. Geoscience and Remote Sensing Symp. IGARSS*, 2021, pp. 4704–4707, DOI: 10.1109/IGARSS47720.2021.9553499.
17. Lin H., Chen J., Pei Z., Zhang S., Hu X., Monitoring sugarcane growth using Envisat ASAR, *IEEE Trans. Geoscience and Remote Sensing*, 2009, Vol. 47, No. 8, pp. 2572–2579.
18. Mishra D., Pathak G., Singh B. P. et al., Crop classification by using dual-pol SAR vegetation indices derived from Sentinel-1 SAR-C data, *Environmental Monitoring Assessment*, 2022, Vol. 195, No. 1, Article 115, DOI: 10.1007/s10661-022-10591-x.
19. Müller H., Rufin P., Griffiths P. et al., Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian savanna landscape, *Remote Sensing of Environment*, 2015, Vol. 156, pp. 490–499, <https://doi.org/10.1016/j.rse.2014.10.014>.
20. Plotnikov D. E., Loupian E. A., Kolbudaev P. A. et al., Daily surface reflectance reconstruction using LOWESS on the example of various satellite systems, *8th Intern. Conf. Information Technology and Nanotechnology (ITNT)*, 2022, DOI: 10.1109/ITNT55410.2022.9848630.
21. Roy D. P., Yan L., Robust Landsat-based crop time series modeling, *Remote Sensing of Environment*, 2018, Vol. 238, Article 110810, <https://doi.org/10.1016/j.rse.2018.06.038>.
22. Van Tricht K., Gobin A., Gilliams S., Piccard I., Synergistic use of radar Sentinel-1 and optical Sentinel-2 imagery for crop mapping: a case study for Belgium, *Remote Sensing*, 2018, V. 10, No. 10, Article 1642, DOI: 10.3390/rs10101642.
23. Veloso A., Mermoz S., Bouvet A. et al., Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications, *Remote Sensing of Environment*, 2017, Vol. 199, pp. 415–426, <https://doi.org/10.1016/j.rse.2017.07.015>.
24. Waldner F., Schucknecht A., Lesiv M. et al., Conflation of expert and crowd reference data to validate global binary thematic maps, *Remote Sensing of Environment*, 2019, Vol. 221, pp. 235–246, DOI: 10.1016/j.rse.2018.10.039.
25. Wang Q., Shi W., Li Z., Atkinson P. M., Fusion of Sentinel-2 images, *Remote Sensing of Environment*, 2016, Vol. 187, pp. 241–252, <https://doi.org/10.1016/j.rse.2016.10.030>.
26. Wu M., Zhang X., Huang W. et al., Reconstruction of daily 30 m data from HJ CCD, GF-1 WFV, Landsat, and MODIS data for crop monitoring, *Remote Sensing*, 2015, Vol. 7, No. 12, pp. 16293–16314, DOI: 10.3390/rs71215826.
27. Wu M., Wu C., Huang W., Niu Z., Wang C., Li W., Hao P., An improved high spatial and temporal data fusion approach for combining Landsat and MODIS data to generate daily synthetic Landsat imagery, *Information Fusion*, 2016, Vol. 31, Issue 31, pp. 14–25.
28. Wu B., Zhang M., Zeng H. et al., Challenges and opportunities in remote sensing-based crop monitoring: A review, *National Science Review*, 2023, Vol. 10, Issue 4, Article nwac290, DOI: doi.org/10.1093/nsr/nwac290.
29. Yan L., Roy D. P., Large-area gap fFilling of Landsat reflectance time series by Spectral-Angle-Mapper Based Spatio-Temporal Similarity (SAMSTS), *Remote Sensing*, 2018, Vol. 10, No. 4, Article 609, DOI: 10.3390/rs10040609.
30. Zhou T., Pan J., Zhang P. et al., Mapping winter wheat with multi-temporal SAR and optical images in an Urban Agricultural Region, *Sensors*, 2017, Vol. 17, No. 6, Article 1210, DOI: 10.3390/s17061210.