

Научная статья

УДК 630*58

<https://doi.org/10.25686/2306-2827.2024.1.30>

EDN: JHUADY

Оценка алгоритма «Случайный лес» машинного обучения для классификации фитомассы лесов

Д. М. Дергунов, О. Н. Воробьев, Э. А. Курбанов[✉], С. А. Лежнин, А. В. Губаев

Поволжский государственный технологический университет,

Российская Федерация, 424000, Йошкар-Ола, пл. Ленина, 3

kurbanovea@volgatech.net[✉]

Введение. Леса играют решающую роль в смягчении последствий изменения климата, защите биоразнообразия и обеспечении других экосистемных услуг, как чистый воздух и вода. Оценка и мониторинг наземной фитомассы лесов являются важными для понимания углеродного баланса, изменений их состояния и продуктивности с течением времени. **Цель** исследования – оценка алгоритма «Случайный лес» машинного обучения для классификации пространственного распределения надземной фитомассы по данным спутникового снимка Sentinel-2 на примере заказника «Сосновоборский» Пензенской области. **Объекты и методы.** Объектом исследования явились лесные насаждения заказника. На исследуемой территории, по данным лесоустройства и полевых работ, заложены 110 тестовых участков. Для моделирования пространственного распределения фитомассы использован ансамблевый метод машинного обучения «Случайный лес». Входными данными для модели были 12 спектральных каналов спутника Sentinel-2 и 7 вегетационных индексов. Для подбора значений оптимальных параметров модели «Случайный лес» использовалась программа GridSearchCV. Моделирование распределения фитомассы лесного насаждения по алгоритму «Случайный лес» проводилось с помощью интерактивной веб-среды с открытым исходным кодом Jupyter Notebook. **Результаты.** При использовании выбранных параметров алгоритма «Случайный лес» была получена оптимальная модель ($R^2=0,60$) распределения фитомассы по территории заказника. Плотность остатков прогнозируемых значений вокруг нуля и их близость к нормальному распределению свидетельствуют об её адекватности. Максимальную связь с фитомассой лесов показали 11 (SWIR), b5, b3, b12, b2 и b7 спектральные каналы снимка Sentinel-2. В итоге получена карта пространственного распределения фитомассы лесов по территории заказника. **Выводы.** Результаты показывают, что сочетание спектральных каналов Sentinel-2 и вегетационных индексов повышают точность оценки фитомассы лесных насаждений с использованием алгоритма машинного обучения «Случайный лес». Разработанный алгоритм может быть рекомендован для определения пространственного распределения надземной фитомассы в лесных экосистемах Пензенской области.

Ключевые слова: спутниковые снимки; Sentinel-2; вегетационные индексы; лесные насаждения; дистанционное зондирование; Пензенская область

Финансирование: работа выполнена в рамках гранта Российского научного фонда № 22-16-00094 «Дистанционный мониторинг лесных экосистем в условиях меняющегося климата», <https://rscf.ru/project/22-16-00094/>.

© Дергунов Д. М., Воробьев О. Н., Курбанов Э. А., Лежнин С. А., Губаев А. В., 2024

Для цитирования: Дергунов Д. М., Воробьев О. Н., Курбанов Э. А., Лежнин С. А., Губаев А. В. Оценка алгоритма «Случайный лес» машинного обучения для классификации фитомассы лесов // Вестник Поволжского государственного технологического университета. Сер.: Лес. Экология. Природопользование. 2024. № 1 (61). С. 30–43. <https://doi.org/10.25686/2306-2827.2024.1.30>; EDN: JHUADY

Введение. Лесные экосистемы играют жизненно важную роль в балансе наземного цикла углерода и могут способствовать смягчению последствий глобального изменения климата на жизнь человека [1]. В последние десятилетия этому направлению посвящено большое число исследований российских и зарубежных учёных [2–4]. Обобщённые результаты по региональным и глобальным оценкам депонирования углерода лесными насаждениями входят в отчёты межправительственной комиссии экспертов по изменению климата [5]. Эффективность и точность подобных исследований в значительной степени связана с использованием современных информационных технологий и методов обработки больших данных.

Основным общепризнанным фактором, определяющим скорость и объём депонирования углерода лесными экосистемами, является их биологическая продуктивность, которая обычно выражается пространственно-временным распределением надземной фитомассы. Для определения фитомассы насаждений с течением времени использовались разные методические решения. Традиционные полевые исследования предоставляют основные таксационные данные (диаметр, высоту, полноту, возраст) деревьев, что позволяет точно определять фитомассу насаждения с использованием аллометрических уравнений, но для этого требуются значительные временные и финансовые затраты [6]. В связи с этим для решения этого вопроса всё чаще используются данные дистанционного зондирования (ДЗЗ). Наиболее точными для этих целей являются лидарные технологии, позволяющие оценивать вертикальную структуру насаждения, а также радарная съёмка с синтезированной апертурой [7–9]. Оптические данные дистанционного зондирования с высоким пространственным разрешением также позволяют получать важную спектральную информацию и проводить мониторинг растительного покрова на больших территориях [10–12].

В последние годы для оценки фитомассы лесного покрова всё большее применение находят спутники Sentinel Европейского космического агентства в рамках программы Copernicus [13]. Sentinel-2 – миссия, направленная на изучение земной поверхности с помощью широкополосных мультиспектральных спутников высокого разрешения Sentinel-2A и Sentinel-2B. Мультиспектральный сенсор MSI Sentinel-2 (*англ.* MultiSpectral Instrument) сканирует поверхность земли в 13 спектральных диапазонах, от видимого до коротковолнового инфракрасного электромагнитного спектра. Группы диапазонов имеют различное пространственное разрешение: четыре спектральных канала имеют 10-метровое пространственное разрешение, шесть каналов – 20-метровое разрешение, и два канала – 60-метровое разрешение [14]. Этот сенсор нашёл широкое применение в мониторинге лесного покрова, а также для оценки запасов леса и ущерба от пожаров [15–17].

Для оценки надземной биомассы (фитомассы) AGB (*англ.* above ground biomass) с использованием ДЗЗ широкое применение находят алгоритмы машинного обучения, позволяющие идентифицировать ключевые спектральные и текстурные переменные [18]. Использование методов машинного обучения также способствует более эффективной обработке спутниковых данных [19, 20], в том числе при оценке фитомассы лесов [21]. Обычно при современных исследованиях применяют такие методы машинного обучения, как дерево решений (*англ.* Decision tree), ближайшего соседа (*англ.* Nearest Neighbor), искусственная нейронная сеть ANN (*англ.* Artificial Neural Network) и метод опорных векторов SVM (*англ.* Support Vector Machine), «Случайный лес» RF (*англ.* Random Forest). Эффективность каждого из этих методов машинного обучения зависит от сложности взаимосвязи между фитомассой и данными дистанционного зондирования [22, 23].

Цель данного исследования – оценить алгоритм машинного обучения «Случайный лес» для классификации пространственного распределения надземной фитомассы лесного покрова по спектральным каналам Sentinel-2 и серии спектральных индексов на примере заказника «Сосновоборский» Пензенской области. Для этого были решены следующие **задачи**:

- 1) заложены тестовые участки на территории исследования;
- 2) получены тематические карты вегетационных индексов по спутниковым снимкам на территорию заказника;
- 3) подобраны оптимальные параметры для моделирования фитомассы по алгоритму «Случайный лес» машинного обучения;
- 4) проведена оценка полученной модели и создана карта распределения плотности фитомассы на территории заказника.

Материалы и методы исследования. Исследования проводились на территории государственного природного зоологического заказника регионального значения «Сосновоборский» Пензенской области (рис. 1). Общая площадь исследуемой территории составляет 8 067 га.

Заказник является особо охраняемой природной территорией регионального значения, предназначенной для сохранения и восстановления редких и исчезающих видов животных, в том числе ценных видов в хозяйственном, научном и культурном отношении, и среды их обитания¹. Задача заказника – особая охрана редких и исчезающих видов животных, а также выполнение функции по сохранению и восстановлению ценных в хозяйственном, научном и культурном отношении охотничьих ресурсов и их воспроизводство (<http://www.oopt.aari.ru/oopt/>).



Рис. 1. Расположение тестовых участков на исследуемой территории
Fig. 1. Location of test sites in the study area

¹ О Внесении изменений в постановление правительства Пензенской области от 13.12.2017 N 604-пп (с последующими изменениями). Постановление правительства Пензенской области от 13.09.2021 № 587-пп [Электронный ресурс] URL: <https://clck.ru/33ujz7>

Во время полевых исследований в 2022 году на территории заповедника было заложено 20 тестовых участков (ТУ) в древостоях различных по таксационным характеристикам. Кроме того, по данным последнего лесоустройства были подобраны 90 ТУ с описанием их основных дендрометрических показателей и привязкой на местности. В целом в работе для классификации и оценки точности были подобраны 40 ТУ в хвойных, 38 в лиственных, 16 в смешанных и 16 на не покрытых лесом участках. Для каждого исследуемого насаждения были определены следующие таксационные показатели: средний возраст, высота и диаметр, породный состав и полнота. Эти данные были использованы для определения запаса насаждений с помощью таблиц для таксации лесов [24]. Полученные данные по запасам насаждений пересчитывались в фитомассу с помощью существующих конверсионных коэффициентов для региона исследования [25, 26].

В работе был использован спутниковый снимок Sentinel-2B MSI L2A T38 UNE от 26 июня 2022 года, полученный с электронного ресурса открытого доступа Copernicus (Sentinels Scientific Data Hub), <https://scihub.copernicus.eu/dhus/#/home>. Для учёта влияния атмосферы и устранения геометрического искажения исходное спутниковое изображение было скорректировано в модуле Sen2Cor, встроенном в программное обеспечение SNAP.

В работе был применён алгоритм машинного обучения «Случайный лес» [27], который основан на комбинировании выходных данных нескольких ансамблей деревьев решений, что позволяет получать оптимальное качество классификации спутниковых изображений. Это гибкий и простой в использовании алгоритм, который можно применять как для задач классификации, так и для задач регрессионного анализа [28, 29]. «Случайный лес» также используется учёными при расчётах фитомассы лесных насаждений путём объединения нескольких деревьев реше-

ний и оптимизации их прогнозов [30]. Алгоритм «Случайный лес» позволяет выявлять сложную нелинейную связь между растительным покровом и изображениями ДЗЗ с неопределённым набором данных, а также способен оптимально комбинировать данные из разных источников для повышения точности прогноза [31, 32]. В данном исследовании для разработки модели «Случайного леса» по оценке фитомассы лесного насаждения использовалась интерактивная веб-среда с открытым исходным кодом Jupyter Notebook, которая позволяет пользователям создавать и обмениваться информацией, содержащей коды, уравнения, визуализацию и текст. Jupyter Notebook поддерживает более 40 языков программирования, включая Python, R, Julia и Scala [33]. Блок-схема последовательности работ в исследовании приведена на рис. 2.



Рис. 2. Блок-схема оценки надземной фитомассы с помощью машинного обучения

Fig. 2. Flowchart of estimating above-ground phytomass using machine learning

В качестве входных параметров в модель оценки фитомассы лесных насаждений были использованы данные, полученные как из спектральных каналов MSI Sentinel-2, так и вегетационных индексов на территорию исследования. Для этого были использованы: нормализованный разностный вегетационный индекс NDVI (*англ.* Normalized Differenced Vegetation Index), расширенный вегетационный индекс EVI (*англ.* Enhanced Vegetation Index), вегетационный индекс с поправкой на почву SAVI (*англ.* Soil adjusted vegetation index), нормализованный разностный индекс влажности NDMI (*англ.* Normalized Differenced Moisture Index), относительный индекс вегетации RVI (*англ.* Ratio Vegetation Index), нормализованный индекс вегетации по краевому диапазону красной зоны спектра NDVI_{re} (*англ.* Red-edge Normalized Difference Vegetation Index), индекс хлорофилла по краевому диапазону красной зоны спектра Cl_{re} (*англ.* Red-edge Chlorophyll Index) (табл.). На первом этапе проведено разделение тестовых участков на обучающую (77 ТУ) и контрольную выборки (33 ТУ).

Далее проводился подбор основных параметров модели «Случайный лес» для оптимизации её производительности:

Вегетационные индексы, используемые в алгоритме «Случайный лес»

Vegetation indices used in the Random Forest algorithm

Индексы	Формула спектральных каналов	Источник
NDVI	$\frac{Band8 - Band4}{Band8 + Band4}$	[35]
EVI	$2,5 \times \frac{Band8 - Band4}{Band8 + 6 \times Band4 - 7,5 \times Band4}$	[36]
SAVI	$(1 + 0,5) \times \frac{Band8 - Band4}{Band8 + Band4 + 0,5}$	[37]
NDMI	$\frac{Band8 - Band11}{Band8 + Band11}$	[38]
RVI	$\frac{Band8}{Band4}$	[39]
NDVI _{re}	$\frac{Band8 - Band5}{Band8 + Band5}$	[40]
Cl _{re}	$\frac{Band8}{Band5} - 1$	[41]

• *n_estimators* – количество деревьев решений в алгоритме Random Forest. Увеличение количества деревьев может повысить точность модели, что в свою очередь приводит к излишней по времени обработке данных на её оптимизацию;

• *max_features* – количество независимых переменных (случайных подмножеств), выбранных случайным образом для использования при каждом решении в RF. Это может быть фиксированное число или доля от общего количества переменных (например, двоичный логарифм от числа переменных – $\log_2$, квадратный корень от числа переменных – $\sqrt{}$). Меньшее значение этого показателя может снизить время на переобучение, а большее повысить точность модели;

• *max_depth* – определяет максимальную глубину (величину), до которой может проходить оптимизация изучаемого дерева решения. Это один из самых важных параметров, когда речь идёт о повышении точности модели. Увеличение этого показателя приводит к повышению точности модели до определённого предела, после которого она начинает постепенно снижаться в связи с её переобучением. Поэтому при таких исследованиях важно изначально корректно установить его значение, чтобы избежать процесса переобучения [34].

Выбор количества параметров – одна из важных задач при построении модели машинного обучения, потому что изменение параметров модели может принципиально повлиять на её качество и время на её обработку. Для подбора параметров алгоритма «Случайного леса» в веб-среде с открытым исходным кодом Jupyter Notebook используется метод поиска по программе GridSearchCV, который автоматически подбирает параметры машинного обучения и создаёт оптимальный набор для каждой из возможной комбинации. Поиск по программе GridSearchCV включает проверку диапазона значений параметров $n_estimators$, $max_features$ и max_depth . После этого выбирается комбинация этих параметров, обеспечивающая максимальную производительность модели.

Заключительным этапом работ является запуск модели и оценка её эффективности. Для этого в работе использована десятикратная перекрёстная проверка, при которой набор данных разделён на десять частей, каждая из которых содержала одинаковое количество тестовых участков. В процессе моделирования на каждом этапе одна часть данных выбиралась в качестве тестового набора данных, а остальные девять частей использовались в качестве обучающих моделей для прогнозирования целевого значения [42, 43]. Такая процедура проводилась десять раз для каждого набора наблюдений. Статистические характеристики модели «Случайный лес» оценивали с помощью коэффициента детерминации (R^2), среднеквадратичной ошибки RMSE (*англ.* Root mean square error) и средней абсолютной ошибки MAE (*англ.* Mean absolute error).

Результаты исследования и их обсуждение. В работе на основе модуля GridSearchCV были получены оптимальные значения параметров модели «Случайный лес» по оценке фитомассы лесных

насаждений заповедника «Сосновоборский» по данным вегетационных индексов: $n_estimators = 500$, $max_features = \sqrt{}$, $max_depth = 10$. Как видно из рис. 3, а, R^2 достигает максимальных значений при увеличении числа «деревьев решений» до 500 штук. Затем наблюдается его небольшое снижение в связи с процессом переобучения модели. Распределение значений R^2 оказалось одинаково для $\log_2$ и $\sqrt{}$, поэтому выбор значения параметров $max_features$ может быть любым. На рис. 3, б видно резкое повышение R^2 при увеличении параметра max_depth числа «деревьев решений» с 0 до 3, затем наблюдается относительная стабилизация зависимости при их увеличении до 10. Максимальный показатель max_depth «дерева решений» позволяет разделить каждый исследуемый узел (*англ.* Node) на оптимальное количество наблюдений, что приводит к получению более точных данных по фитомассе насаждений.

Интеграция оптимальных значений параметров в модель «Случайного леса» с использованием наборов данных ТУ для обучения и тестирования позволила получить следующие статистики: $R^2=0,60$, $MAE=17,6$ т/га, $RMSE=20,9$ т/га, что достаточно для формирования модели расчёта фитомассы на территории исследования. На рис. 4 показано распределение данных фитомассы, сгенерированных по алгоритму «Случайный лес» и по тестовым участкам, на территории заказника «Сосновоборский». Отдельные прогнозные значения фитомассы (в виде точек) располагаются близко к оптимальной модели (сплошная линия). В целом, на рис. 4 наблюдается низкий уровень дисперсии в модели «Случайный лес» в сравнении с оптимальной линией, в которой распределение точно соответствует модели, что свидетельствует о приемлемом качестве принятой модели для оценки фитомассы.

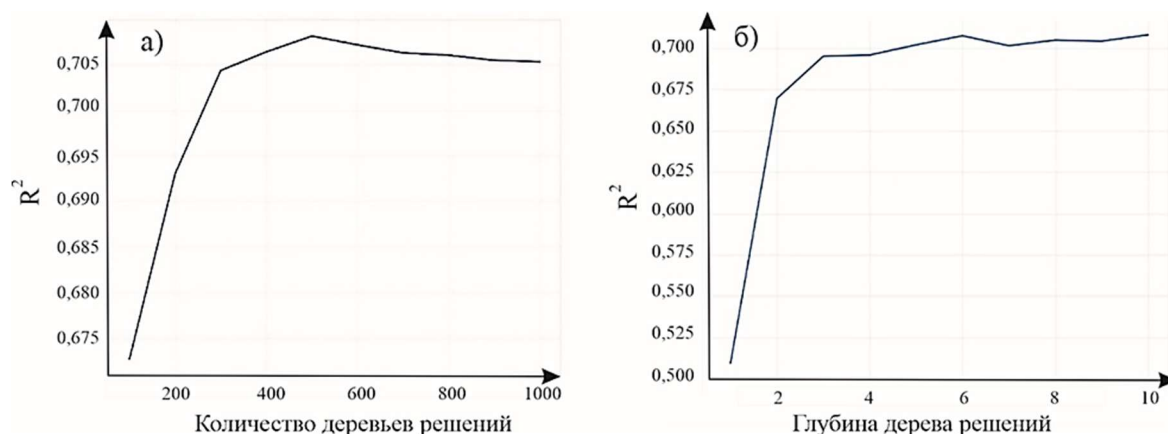


Рис. 3. Распределение значений R^2 по параметрам модели «дерево решений»: а) $n_estimators$; б) max_depth
 Fig. 3. Distribution of R^2 values by parameter of the Decision Tree model: а) $n_estimators$; б) max_depth

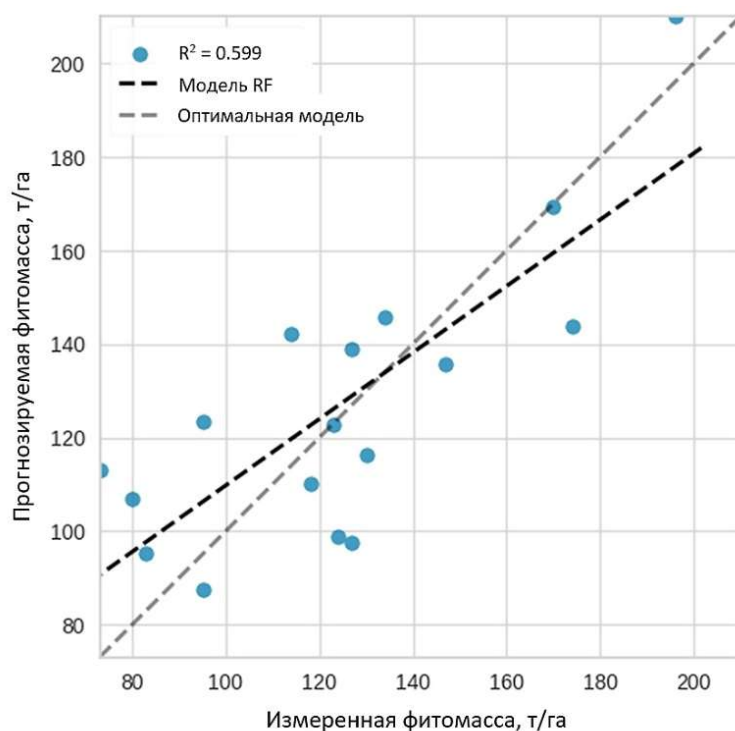


Рис. 4. Модели распределения и полевые данные фитомассы лесного покрова заказника «Сосновоборский»
 Fig. 4. Distribution of predicted and measured forest cover phytomass in the Sosnovoborsky nature reserve

На рис. 5 представлено распределение между фактическими и прогнозируемыми значениями фитомассы и остатками, случайно рассеянными вокруг нулевых значений. Гистограмма остатков демонстрирует нормальное распределение, а плотность прогнозируемых показателей также подтверждает адекватность модели (рис. 5). В работе установлено, что моделирование по алгоритму «Случайный лес» переоценивает небольшие значения фи-

томассы и недооценивает большие, что частично объясняет наличие систематической ошибки в прогнозе и указывает на проблему насыщения. Следует отметить, что проблемы завышения и занижения оценок до сих пор не удалось полностью устранить с помощью алгоритмов машинного обучения. Эта проблема также существовала в предыдущих исследованиях с непараметрическим алгоритмом оценки фитомассы [44].

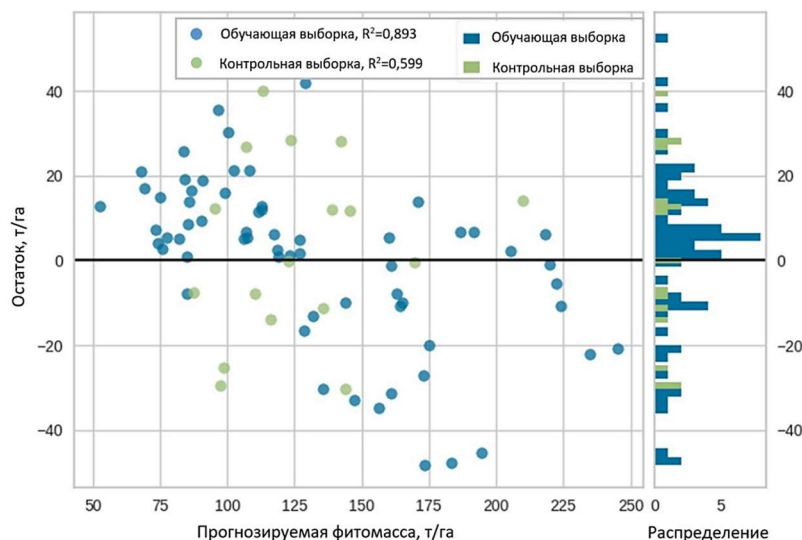


Рис. 5. График остатков, показывающий разницу между фактическими и моделируемыми значениями обучающей выборки (синий цвет) и контрольной выборки (зелёный цвет) фитомассы лесного покрова (т/га)

Fig. 5. Residue plot showing the difference between the actual and predicted values of the training sample (blue) and values of the control sample (green) of the forest cover phytomass (t/ha)

Оптимальные результаты модели оценки фитомассы могут быть объяснены использованием спектральных каналов Sentinel-2, которые лучшим образом фиксируют растительный покров, особенно в краевой красной зоне спектра Red Edge (каналы 5–7) (рис. 6). На рис. 6 показано распределение спектральных каналов и используемых в работе вегетационных индексов по их относительной важности при оценке фитомассы. Среди всех оцениваемых параметров, использованных в модели расчёта фитомассы по алгоритму «Случайный лес», максимум влияния оказывают значения спектральных показателей канала 11 (SWIR – коротковолновый

инфракрасный) и данные по b5, b3, b12, b2 и b7 каналам. При этом подходе машинного обучения вегетационные индексы показывают меньшее влияние на точность при оценке фитомассы. Это является серьёзным недостатком их использования, так как приводит к недооценке продуктивности районов с густыми лесами. Исследуемые индексы могут иметь низкую важность из-за своей высокой насыщенности в летний период, когда достигается пик вегетационного развития растений [45]. Для картографирования наземной фитомассы использовались показатели с относительной важностью более 50 %.

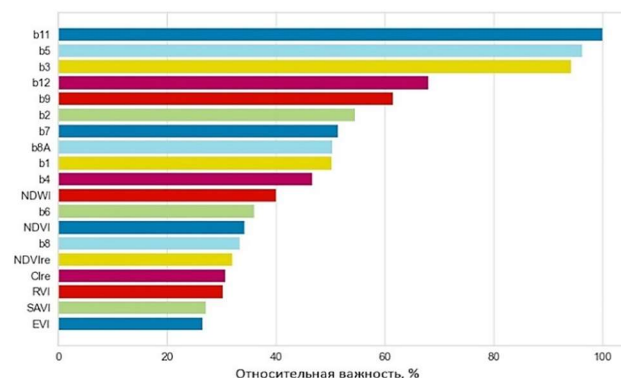


Рис. 6. Параметры модели, ранжированные в соответствии с объяснённой дисперсией, которую каждый из них вносит в модель

Fig. 6. Features ranked according to the explained variance that each feature contributes to the model

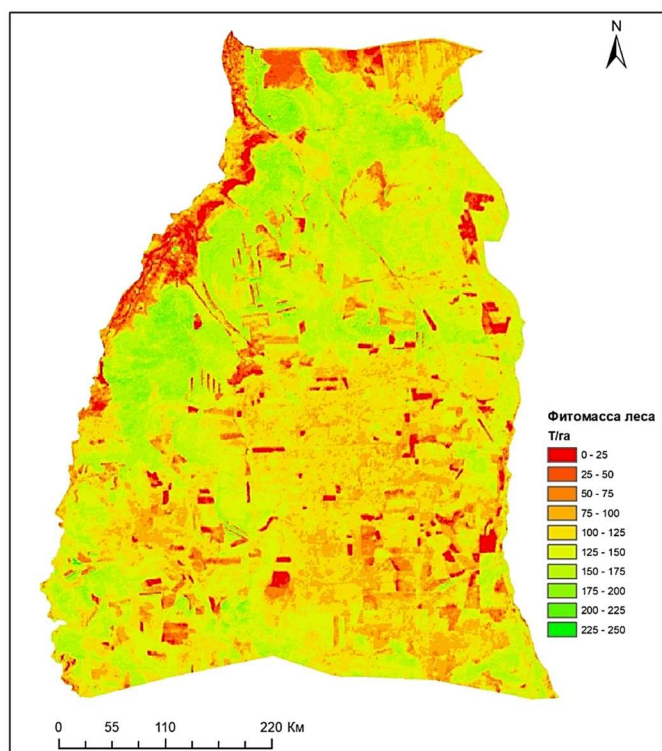


Рис. 7. Распределение фитомассы на исследуемой территории, полученной с использованием модели «Случайный лес»

Fig. 7. Phytomass distribution across the study area obtained using the Random Forest model

На рис. 7 представлена итоговая карта распределения фитомассы лесов заказника «Сосновоборский» Пензенской области, полученная на основе модели «Случайный лес» по данным спутниковых изображений и тестовых участков. Минимальные значения плотности фитомассы (красным цветом) наблюдаются на открытых участках заказника (вырубки, поля), сосредоточенных в основном в северной и северо-западной частях заказника. Также их небольшие участки можно найти в центральной части территории исследования. Максимальная плотность фитомассы лесных насаждений (светло-зелёный цвет), представленная в основном спелыми сосновыми насаждениями, сосредоточена в северной и центральной частях заказника.

Выводы. В работе проведена оценка набора спектральных каналов и вегетационных индексов, полученных по данным изображения Sentinel-2 MSI, с целью последующего моделирования распределения фитомассы в разновозрастных лесных ланд-

шафтах заказника «Сосновоборский» Пензенской области с помощью алгоритма машинного обучения «Случайный лес». Применение автоматического машинного обучения позволило ускорить процесс моделирования распределения фитомассы и оценку точности классификации спутниковых данных. Приемлемость использования алгоритма «Случайный лес» в таких оценках обусловлена возможностью использования нескольких значимых параметров спутниковых данных, необходимых для выбора наиболее оптимальной модели. В работе были использованы только два из них: спектральные каналы и вегетационные индексы.

Перспективным направлением оценки фитомассы с помощью машинного обучения является использование других источников данных ДЗЗ, таких как радиолокатор с синтезированной апертурой, гиперспектральные изображения с высоким пространственным разрешением, разновременные спутниковые данные для учёта сезонности, а также вспомогательные данные

(например, климат, почва и рельеф). Полученная модель может быть использована для более точного определения запасов депонированного углерода в лесных экосистемах Пензенской области по спутниковым данным. В свою очередь, это позволит

повысить точность оценки запасов углерода в лесных насаждениях на региональном уровне, что необходимо для выполнения Российской Федерацией обязательств по Парижскому соглашению об изменении климата.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. By 2050 the Mitigation Effects of EU Forests Could Nearly Double through Climate Smart Forestry / G.-J. Nabuurs, P. Delacote, D. Ellison et al. // *Forests*. 2017. Vol. 8. Iss. 12. Art. 484. DOI: 10.3390/f8120484
2. Reichstein M., Carvalhais N. Aspects of Forest Biomass in the Earth System: Its Role and Major Unknowns // *Surveys in Geophysics*. 2019. Vol. 40. Iss. 4. Pp. 693–707. DOI: 10.1007/s10712-019-09551-x
3. Анализ трендов временных рядов вегетационных индексов по данным MODIS для оценки влияния засух на лесные насаждения Среднего Поволжья с 2000 по 2020 год / О. Н. Воробьев, Э. А. Курбанов, Дж. Ша и др. // *Современные проблемы дистанционного зондирования Земли из космоса*. 2022. Т. 19. № 4. С. 181–194. [Vorobyov O. N., Kurbanov E. A., Sha J. et al. Trend analysis of MODIS time series vegetation indices to assess the impact of droughts on forest stands in the Middle Volga from 2000 to 2020. *Current Problems in Remote Sensing of the Earth from Space*. 2022. Vol. 19. Iss. 4. Pp. 181–194.] DOI: 10.21046/2070-7401-2022-19-4-181-194; EDN: LIMGBP
4. Прогноз состояния лесных экосистем Среднего Поволжья с использованием самообучающихся моделей / С. А. Лежнин, А. В. Губаев, О. Н. Воробьев и др. // *Лесные экосистемы в условиях изменения климата: биологическая продуктивность и дистанционный мониторинг*. 2022. № 8. С. 95–107. [Lezhnin S. A., Gubaev A. V., Vorobiev O. N. et al. Forecasting the state of forest ecosystems in the Middle Volga region using self-learning models. *Forest Ecosystems under Climate Change: Biological Productivity and Remote Monitoring*. 2022. No 8. Pp. 95–107.] DOI: 10.25686/10.25686.2022.48.81.010; EDN: PPMKHO
5. IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, 184 p., DOI: 10.59327/IPCC/AR6-9789291691647
6. Consequences of Landsat Image Strata Classification Errors on Bias and Variance of Inventory Estimates: A Forest Inventory Case Study / M. K. Crosby, T. G. Matney, E. B. Schultz et al. // *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2017. Vol. 10. Iss. 1. Pp. 243–251. DOI: 10.1109/JSTARS.2016.2597762
7. Estimation and mapping of above-ground biomass of mangrove forests and their replacement land uses in the Philippines using Sentinel imagery / J. A. Castillo, A. A. Apan, T. N. Maraseni et al. // *ISPRS Journal of Photogrammetry and Remote Sensing*. 2017. Vol. 134. Pp. 70–85. DOI: 10.1016/j.isprsjprs.2017.10.016
8. Combining LiDAR and hyperspectral data for aboveground biomass modeling in the Brazilian Amazon using different regression algorithms / C. T. Almeida, L. S. Galvao, L. E. O. C. Aragao et al. // *Remote Sensing of Environment*. 2019. Vol. 232. DOI: 10.1016/j.rse.2019.111323
9. Регрессионные модели для оценки фитомассы древостоев на основе бортового лидара / В. А. Усольцев, И. С. Цепордей, В. П. Часовских и др. // *Вестник Поволжского государственного технологического университета. Серия: Лес. Экология. Природопользование*. 2023. № 3 (59). С. 24–41. [Usoltsev V. A., Tsepordey I. S., Chasovskikh V. P. et al. Regression Models for Estimating the Phytomass of Forest Stands Using Airborne LiDAR. *Vestnik of Volga State University of Technology. Ser.: Forest. Ecology. Nature Management*. 2023. No 3 (59). Pp. 24–41.] DOI: 10.25686/2306-2827.2023.3.24; EDN: ATXTSY
10. Распознавание лесных насаждений и доминирующих древесных пород Пензенской области по данным спутника Sentinel-2 / Э. А. Курбанов, О. Н. Воробьев, С. А. Меньшиков и др. // *Современные проблемы дистанционного зондирования Земли из космоса*. 2018. Т. 15. № 5. С. 154–166. [Kurbanov E. A., Vorobiev O. N., Menshikov S. A. et al. Identification of forest stands and dominant tree species in Penza Region using Sentinel-2 imagery. *Current Problems in Remote Sensing of the Earth from Space*. 2018. Vol. 15. Iss. 5. Pp. 154–166.] DOI: 10.21046/2070-7401-2018-15-5-154-166; EDN: YOOOXJ
11. Пространственная динамика фитомассы березняков на бывших сельскохозяйственных землях Марийского Заволжья / Э. А. Курбанов, О. Н. Воробьев, Л. С. Устюгова и др. // *Известия вузов. Лесной журнал*. 2010. № 3. С. 8–14. [Kurbanov E. A., Vorobiev O. N., Ustyugova L. S. et al. Spatial Dynamics of Birch Forests Phytomass on Abandoned Agricultural Lands of Mari Transvolga Region. *Russian Forestry Journal*. 2010. No 3. Pp. 8–14.] EDN: MUROND

12. The Influence of Data Density and Integration on Forest Canopy Cover Mapping Using Sentinel-1 and Sentinel-2 Time Series in Mediterranean Oak Forests / V. Nasiri, S. M. M. Sadeghi, F. Moradi et al. // ISPRS International Journal of Geo-Information. 2022. Vol. 11. Iss. 8. Art. 423. DOI: 10.3390/ijgi11080423
13. Comparison of Sentinel-2 and Landsat 8 in the estimation of boreal forest canopy cover and leaf area index / L. Korhonen, Hadi, P. Packalen et al. // Remote Sensing of Environment. 2017. Vol. 195. Pp. 259–274. DOI: 10.1016/j.rse.2017.03.021
14. Landsat-8 and Sentinel-2 burned area mapping – A combined sensor multi-temporal change detection approach / D. P. Roy, H. Huang, L. Boschetti et al. // Remote Sensing of Environment. 2019. Vol. 231. DOI: 10.1016/j.rse.2019.111254
15. Nandy S., Srinet R., Padalia H. Mapping forest height and aboveground biomass by integrating ICESat-2, Sentinel-1 and Sentinel-2 data using Random forest algorithm in northwest Himalayan foothills of India // Geophysical Research Letters. 2021. Vol. 48. Iss. 14. Art. e93799. DOI: 10.1029/2021GL093799
16. Оценка разносезонных изображений Sentinel-2 для классификации лесного покрова водохранимых зон рек Марийского Заволжья / Л. В. Тарасова, Э. А. Курбанов, О. Н. Воробьев и др. // Вестник Поволжского государственного технологического университета. Сер: Лес. Экология. Природопользование. 2023. № 2 (58). С. 77–92. [Tarasova L. V., Kurbanov E. A., Vorobiev O. N. et al. Assessment of Multi-Season Sentinel-2 Images for Classification of Forest Cover in Riparian Zones of Mari Zavolzhye. Vestnik of Volga State University of Technology. Ser.: Forest. Ecology. Nature Management. 2023. No 2 (58). Pp. 77–92.] DOI: 10.25686/2306-2827.2023.2.77; EDN: OUDRPC
17. Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam / A. T. N. Dang, S. Nandy, R. Srinet et al. // Ecological Informatics. 2019. Vol. 50. Pp. 24–32. DOI: 10.1016/j.ecoinf.2018.12.010
18. The use of machine learning methods to estimate aboveground biomass of grasslands: A review / T. G. Morais, R. F. Teixeira, M. Figueiredo et al. // Ecological Indicators. 2021. Vol. 130. Art. 108081. DOI: 10.1016/j.ecolind.2021.108081
19. A survey on the use of GIS and remote sensing for sustainable forestry and ecology in Russia and China / E. Kurbanov, O. Vorobiev, J. Sha et al. // Современные проблемы дистанционного зондирования Земли из космоса. 2020. Т. 17. № 5. С. 9–20. DOI: 10.21046/2070-7401-2020-17-5-9-20; EDN: KVLZRZ
20. Evaluation of alternative approaches for landscape-scale biomass estimation in a mixed-species northern forest / C. M. Hoover, M. J. Ducey, R. A. Colter et al. // Forest Ecology and Management. 2018. Vol. 409. Pp. 552–563. DOI: 10.1016/j.foreco.2017.11.040
21. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms / Y. Li, M. Li, C. Li et al. // Scientific Reports. 2020. Vol. 10. Art. 9952. DOI: 10.1038/s41598-020-67024-3
22. Machine learning-based estimates of above-ground biomass of subalpine forests using Landsat 8 OLI and Sentinel-2B images in the Jiuzhaigou National Nature Reserve, Eastern Tibet Plateau / K. Luo, Y. Wei, J. Du et al. // Journal of Forestry Research. 2022. Vol. 33. Iss. 4. Pp. 1329–1340. DOI: 10.1007/s11676-021-01421-w
23. Machine learning and geostatistical approaches for estimating aboveground biomass in Chinese subtropical forests / H. Su, W. Shen, J. Wang et al. // Forest Ecosystems. 2020. Vol. 7. Art. 64. DOI: 10.1186/s40663-020-00276-7
24. Общесоюзные нормативы для таксации лесов / В. В. Загребев, В. И. Сухих, А. З. Швиденко и др. М: Колос, 1992. 495 с. [Zagreev V. V., Sukhikh V. I., Shvidenko A. Z. et al. All-Union standards for forest inventory. Moscow: Kolos, 1992. 495 p.]
25. Замолодчиков Д. Г., Уткин А. И., Честных О. В. Коэффициенты конверсии запасов насаждений в фитомассу для основных лесобразующих пород России // Лесная таксация и лесоустройство. 2003. № 1 (32). С. 119–127. [Zamolodchikov D. G., Utkin A. I., Chestnykh O. V. Conversion Rates of Planting Stocks into the Phytomass of the Main Forest-Forming Species of Russia. Forest Inventory and Forest Planning. 2003. No 1 (32). Pp. 119–127.] EDN: YYIDOX
26. Курбанов Э. А. Бюджет углерода сосновых экосистем Волго-Вятского района: монография. Йошкар-Ола: Марийский государственный технический университет, 2002. 300 с. [Kurbanov E. A. Carbon budget of pine ecosystems in the Volga-Vyatka region. Monograph. Yoshkar-Ola: MarGTU Publ., 2002. 300 p.] EDN: RTPWCT
27. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45. Iss. 1. Pp. 5–32. DOI: 10.1023/A:1010933404324
28. Liess M., Schmidt J., Glaser B. Improving the Spatial Prediction of Soil Organic Carbon Stocks in a Complex Tropical Mountain Landscape by Methodological Specifications in Machine Learning Approaches // PLOS ONE. 2016. Vol. 11. Iss. 4. Art. e0153673. DOI: 10.1371/journal.pone.0153673
29. Belgiu M., Dragut L. Random Forest in Remote Sensing: A Review of Applications and Future Directions // ISPRS Journal of Photogrammetry and Remote Sensing. 2016. Vol. 114. Pp. 24–31. DOI: 10.1016/j.isprsjprs.2016.01.011
30. Above-ground biomass estimation from LiDAR data using random forest algorithms / L. Torre-Tojal, A. Bastarrika, A. Boyano et al. // Journal of Computational Science. 2022. Vol. 58. Art. 101517. DOI: 10.1016/j.jocs.2021.101517
31. Evaluation of ALOS/PALSAR L-band data for the estimation of Eucalyptus plantations above-

ground biomass in Brazil / N. Baghdadi, G. Le Maire, J.-S. Bailly et al. // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2015. Vol. 8. Iss. 8. Pp. 3802–3811. DOI: 10.1109/JSTARS.2014.2353661

32. Non-Parametric Retrieval of Aboveground Biomass in Siberian Boreal Forests with ALOS PALSAR Interferometric Coherence and Backscatter Intensity / M. A. Stelmaszczyk-Górska, P. Rodriguez-Veiga, N. Ackermann et al. // Journal of Imaging. 2016. Vol. 2. Iss. 1. Art. 1. DOI: 10.3390/jimaging2010001

33. Five Guiding Principles to Make Jupyter Notebooks Fit for Earth Observation Data Education / J. Wagemann, F. Fierli, S. Mantovani et al. // Remote Sensing. 2022. Vol. 14. Iss. 14. Art. 3359. DOI: 10.3390/rs14143359

34. Probst P., Boulesteix A.-L. To tune or not to tune the number of trees in random forest // The Journal of Machine Learning Research. 2018. Vol. 18. Art. 181. DOI: 10.48550/arXiv.1705.05654

35. Monitoring vegetation systems in the great plains with ERTS / Rousel J., Haas R., Schell J. et al. // Proceedings of the Third Earth Resources Technology Satellite-1 Symposium. Volume I: Technical Presentations. Section A. NASA SP-351. NASA: Washington, DC, USA, 1974. Pp. 309–317.

36. Huete A., Justice C., Van Leeuwen W. MODIS Vegetation index (MOD 13) // EOS MODIS Algorithm-Theoretical Basis Document. NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA, 1999. Version 3. 115 p. URL: https://modis.gsfc.nasa.gov/data/atbd/atbd_mod13.pdf (дата обращения: 15.12.2023).

37. Huete A. R. A soil-adjusted vegetation index (SAVI) // Remote Sensing of Environment. 1988. Vol. 25. Iss. 3. Pp. 295–309. DOI: 10.1016/0034-4257(88)90106-X

38. Gao B. C. NDWI – a normalized difference water index for remote sensing of vegetation liquid water from space // Remote Sensing of Environment. 1996. Vol. 58. Iss. 3. Pp. 257–266. DOI: 10.1016/S0034-4257(96)00067-3

39. Baret F., Guyot G. Potentials and limits of vegetation indices for LAI and APAR assessment // Remote Sensing of Environment. 1991. Vol. 35. Iss. 2–3. Pp. 161–173. DOI: 10.1016/0034-4257(91)90009-U

40. Assessment of red-edge vegetation indices for crop leaf area index estimation / T. Dong, J. Liu, J. Shang et al. // Remote Sensing of Environment. 2019. Vol. 222. Pp. 133–143. DOI: 10.1016/j.rse.2018.12.032

41. Gitelson A. A., Gritz Y., Merzlyak M. N. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves // Journal of Plant Physiology. 2003. Vol. 160. Iss. 3. Pp. 271–282. DOI: 10.1078/0176-1617-00887

42. Fushiki T. Estimation of prediction error by using K-fold cross-validation // Statistics and Computing. 2011. Vol. 21. Iss. 2. Pp. 137–146. DOI: 10.1007/s11222-009-9153-8

43. Liaw A., Wiener M. Classification and regression by Random Forest // R News. 2002. Vol. 2. Iss. 3. Pp. 18–22.

44. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data / I. Ali, F. Greifeneder, J. Stamenkovic et al. // Remote Sensing. 2015. Vol. 7. Iss. 12. Pp. 16398–16421. DOI: 10.3390/rs71215841

45. Tesfaye A. A., Awoke G. B. Evaluation of the saturation property of vegetation indices derived from sentinel-2 in mixed crop-forest ecosystem // Spatial Information Research. 2021. Vol. 29. Iss. 1. Pp. 109–121. DOI: 10.1007/s41324-020-00339-5

Статья поступила в редакцию 20.12.2023; одобрена после рецензирования 10.01.2024; принята к публикации 04.03.2024

Информация об авторах

ДЕРГУНОВ Денис Михайлович – аспирант кафедры лесоводства и лесоустройства, Поволжский государственный технологический университет. Область научных интересов – дистанционное зондирование лесов и ГИС. Автор 15 научных публикаций. SPIN-код: 1356-4021

ВОРОБЬЁВ Олег Николаевич – кандидат сельскохозяйственных наук, доцент кафедры лесоводства и лесоустройства, Поволжский государственный технологический университет. Область научных интересов – дистанционное зондирование лесов и ГИС, депонирование углерода лесными экосистемами и их мониторинг. Автор 86 научных и учебно-методических публикаций. ORCID: <https://orcid.org/0000-0002-4897-677X>; SPIN-код: 2870-0763

КУРБАНОВ Эльдар Аликрамович – доктор сельскохозяйственных наук, профессор кафедры лесоводства и лесоустройства, руководитель международного центра устойчивого управления и дистанционного мониторинга лесов, Поволжский государственный технологический университет. Область научных интересов – устойчивое управление лесами, ДЗ лесов и ГИС, биологическая продуктивность и депонирование углерода лесными экосистемами. Автор 180 научных и учебно-методических публикаций. ORCID: <https://orcid.org/0000-0001-5330-9990>; SPIN-код: 9202-6359

ЛЕЖНИН Сергей Анатольевич – кандидат сельскохозяйственных наук, доцент кафедры лесоводства и лесоустройства, Поволжский государственный технологический университет. Область научных интересов – дистанционное зондирование земли, биологическая продуктивность лесных экосистем. Автор 67 научных публикаций. ORCID: <https://orcid.org/0000-0001-6319-4752>; SPIN-код: 3433-2617

ГУБАЕВ Александр Владимирович – кандидат сельскохозяйственных наук, старший научный сотрудник сектора научных программ и грантов, Поволжский государственный технологический университет. Область научных интересов – дистанционное зондирование земли, биологическая продуктивность лесных экосистем. Автор 30 научных публикаций. ORCID: <https://orcid.org/0000-0001-9410-9178>; SPIN-код: 8458-2979

Вклад авторов:

Дергунов Д. М. – концепция, методика, валидация, полевые данные, подготовка статьи, визуализация.

Воробьев О. Н. – методика, валидация, формальный анализ, проверка данных, редактирование.

Курбанов Э. А. – концепция, методика, проверка данных, подготовка статьи, руководство, финансирование.

Лежнин С. А. – программное обеспечение, формальный анализ, редактирование.

Губаев А. В. – программное обеспечение, визуализация.

Авторы заявляют об отсутствии конфликта интересов.

Все авторы прочитали и одобрили окончательный вариант рукописи.

Scientific article

UDC 630*58

<https://doi.org/10.25686/2306-2827.2024.1.30>

EDN: JHUADY

Estimation of the Random Forest Machine Learning Algorithm for Classification of Forest Phytomass

D. M. Dergunov, O. N. Vorobev, E. A. Kurbanov[✉], S. A. Lezhnin, A. V. Gubaev

Volga State University of Technology,
3, Lenin Sq., Yoshkar-Ola, 424000, Russian Federation
kurbanovea@volgatech.net[✉]

ABSTRACT

Introduction. Forests play a critical role in mitigating climate change, protecting biodiversity and providing ecosystem services such as clean air and water. Assessing and monitoring above-ground forest phytomass is important for understanding the carbon balance and changes in forest condition and productivity over time. The **purpose** of the study is to evaluate the Random Forest machine learning algorithm for classifying the spatial distribution of above-ground phytomass based on Sentinel-2 satellite image data using the case of the Sosnovoborsky nature reserve in the Penza region. **Objects and methods.** The object of the study is the forest stands of the Sosnovoborsky nature reserve. According to forest management and field data, 110 test plots were established in the study area. To model the spatial distribution of phytomass over the study area, the ensemble machine learning method Random Forest was used. The input data for the model were 12 spectral bands and 7 vegetation indices. To select the values of the optimal parameters of the Random Forest model, the GridSearchCV program was employed. Modeling of the distribution of the phytomass of a forest stand using the Random Forest algorithm was carried out using the interactive open-source web environment Jupyter Notebook. **Results.** Using the selected parameters of the Random Forest algorithm, an optimal model ($R^2=0.60$) of phytomass distribution throughout the reserve was built. The density of the residuals of the predicted values around zero and their closeness to the normal distribution indicate the adequacy of the model. The maximum connection with forest phytomass was demonstrated by 11 (SWIR), b5, b3, b12, b2 and b7 spectral bands of the Sentinel-2 image. As a result, a map of the spatial distribution of the forest phytomass across the study area was obtained. **Conclusions.** The results show that the combination of Sentinel-2 spectral bands and vegetation indices improves the accuracy of estimating the phytomass of forest stands using the Random Forest machine learning algorithm. The developed algorithm can be recommended for determining the spatial distribution of the above-ground phytomass in the forest ecosystems of the Penza region.

Keywords: satellite imagery; Sentinel-2; vegetation indexes; forest ecosystems; remote sensing; Penza region

Funding: the work was carried out within the framework of a grant from the Russian Science Foundation № 22-16-00094 “Remote monitoring of forest ecosystems in a changing climate”, <https://rscf.ru/project/22-16-00094/>.

The article was submitted 20.12.2023; approved after reviewing 10.01.2024;
accepted for publication 04.03.2024

For citation: Dergunov D. M., Vorobev O. N., Kurbanov E. A., Lezhnin S. A., Gubaev A. V. Estimation of the Random Forest Machine Learning Algorithm for Classification of Forest Phytomass. *Vestnik of Volga State University of Technology. Ser.: Forest. Ecology. Nature Management.* 2024. № 1 (61). Pp. 30–43. (In Russ.). <https://doi.org/10.25686/2306-2827.2024.1.30>; EDN: JHUADY

Information about the authors

Denis M. Dergunov – Postgraduate student, Chair of Silviculture and Forest Inventory, Volga State University of Technology. Research interests – forest remote sensing and GIS. Author of 15 scientific publications. SPIN: 1356-4021

Oleg N. Vorobev – Candidate of Agricultural Sciences, Associate Professor at the Chair of Silviculture and Forest Inventory, Volga State University of Technology. Research interests – forest remote sensing and GIS, carbon sequestration by forest ecosystems, forest eco-system monitoring. Author of 86 scientific publications and textbooks. ORCID: <https://orcid.org/0000-0002-4897-677X>; SPIN: 2870-0763

Eldar A. Kurbanov – Doctor of Agricultural Sciences, Professor at the Chair of Silviculture and Forest Inventory; Head of the International Center for Sustainable Forest Management and Remote Sensing, Volga State University of Technology. Research interests – sustainable forest management, forest remote sensing and GIS, biological productivity of forest ecosystems, carbon sequestration by forest ecosystems. Author of 180 scientific publications and textbooks. ORCID: <https://orcid.org/0000-0001-5330-9990>; SPIN: 9202-6359

Sergey A. Lezhnin – Candidate of Agricultural Sciences, Associate Professor at the Chair of Silviculture and Forest Inventory, Volga State University of Technology. Research interests – remote sensing of the Earth, biological productivity of forest ecosystems. Author of 67 scientific publications. ORCID: <https://orcid.org/0000-0001-6319-4752>; SPIN: 3433-2617

Alexandr V. Gubaev – Candidate of Agricultural Sciences, Senior Researcher, Division of Scientific Programs and Grants, Volga State University of Technology. Research interests – remote sensing of the Earth, biological productivity of forest ecosystems. Author of 30 scientific publications. ORCID: <https://orcid.org/0000-0001-9410-9178>; SPIN: 8458-2979

Contribution of the authors:

Dergunov D. M. – conceptualization, methodology, validation, field data, manuscript preparation, visualization.

Vorobev O. N. – methodology, validation, formal analysis, data verification, reviewing and editing.

Kurbanov E. A. – conceptualization, methodology, data verification, manuscript preparation, supervision, funding.

Lezhnin S. A. – software, formal analysis, reviewing and editing.

Gubaev A. V. – software, visualization.

The authors declare that they have no conflict of interest.

All authors read and approved the final manuscript.