

Выбор информативных индикаторов и метода построения регрессии для оценки бонитета лесов по спутниковым данным дистанционного зондирования

С. А. Хвостиков, С. А. Барталев, В. А. Егоров, Е. А. Стыщенко

Институт космических исследований РАН, Москва, 117997, Россия

E-mail: khvostikov@d902.iki.rssi.ru

Развитие методов машинного обучения привело к появлению множества подходов к оценке характеристик растительного покрова на основе данных дистанционного зондирования. Проведённый анализ позволил выбрать оптимальный метод машинного обучения и набор информативных спутниковых индикаторов для решения задачи оценки в масштабах страны продуктивности лесов, в качестве характеристики которой в настоящем исследовании был принят показатель бонитета. Большой объём опорной выборки, состоящей из более чем 28 млн пикселей карты с известным классом бонитета лесов, и исходный набор из 669 полученных по спутниковым изображениям признаков привели к необходимости выбора наиболее значимых из них. Анализ различных методов оценки значимости признаков показал, что для поставленной задачи лучшие результаты демонстрируют методы на основе случайной перестановки признака на входе построенной регрессии. Применение указанных методов позволило сократить исходный набор из 669 признаков до 100 наиболее информативных из них с незначительной потерей точности. С использованием сокращённого набора признаков выполнен анализ нескольких методов машинного обучения, и сделан вывод об оптимальности подхода на основе градиентного бустинга в реализации LightGBM. Комбинация сокращённого набора признаков и метода градиентного бустинга использовалась для построения карты классов бонитета лесов страны. Оценка точности полученных результатов на основе отложенной части выборки показала высокий коэффициент детерминации, равный 0,87, среднеквадратическая ошибка составила 0,5 класса бонитета. Сопоставление полученной карты бонитета лесов с наземными данными показало более чем 80%-й уровень их совпадения.

Ключевые слова: леса, бонитет, дистанционное зондирование, оценка значимости признаков, градиентный бустинг

Одобрена к печати: 05.06.2025

DOI: 10.21046/2070-7401-2025-22-4-64-75

Введение

Оценка характеристик лесных экосистем — одна из важнейших задач дистанционного зондирования Земли (ДЗЗ) из космоса. Спутниковые данные позволяют распознавать типы земного покрова, строить карты породного состава лесов, определять их запас и биомассу, осуществлять мониторинг нарушений и оценивать множество других характеристик (Барталев и др., 2016). Для решения такого рода задач часто применяются разные методы машинного обучения, обычно опирающиеся на набор спутниковых индикаторов (признаков), связанных с исследуемой характеристикой (тип покрова, порода лесов, их биомасса и т.д.), и обучающий набор опорных данных с известными значениями искомой характеристики для выборочных участков.

Анализ литературы показывает большое разнообразие используемых в задачах дистанционного зондирования методов машинного обучения (Maxwell et al., 2018). Для оценки характеристик земного покрова могут применяться методы случайного леса, нейронных сетей, градиентного бустинга, опорных векторов, k ближайших соседей и др., в том числе основанные, например, на модельных подходах. Разнообразие методов, используемых в задачах дистанционного зондирования, обуславливает необходимость обоснованного выбора оптимального подхода для решения конкретной задачи. При этом существует ограниченное количество работ, сравнивающих возможности применения разных методов в задачах дистанционного

зондирования. Например, можно отметить работу (Khatami et al., 2016), в которой был проведён сравнительный мета-анализ точности разных методов. В указанной публикации сделан вывод о высокой эффективности нейронных сетей, а также показана возможность получения хороших результатов на основе методов случайного леса и опорных векторов. В другой работе (Sheykhmousa et al., 2020) сравниваются методы случайного леса и опорных векторов и делается заключение о небольшом преимуществе последнего. Но в целом малое количество подобных работ, особенно в такой динамично развивающейся сфере, как машинное обучение, не позволяет сформулировать однозначные априорные рекомендации по выбору оптимального метода для решения конкретной задачи.

С другой стороны, нельзя не отметить появление всё большего количества спутниковых индикаторов, которые позволяют оценивать состояние земного покрова. Это связано с возрастающим количеством спутниковых систем ДЗЗ, увеличением частоты наблюдений, развитием методов реконструкции (включая фильтрацию шумов и интерполяцию) временных рядов данных дистанционного зондирования. Большое количество признаков может затруднять применение методов машинного обучения за счёт возрастания требований к вычислительным ресурсам. Также некоторые исследователи отмечают, что исключение части признаков может увеличить точность классификации (Maxwell et al., 2018). В связи с этим выбор оптимального набора информативных признаков при оценке состояния земного покрова на основе спутниковых данных ДЗЗ становится актуальной задачей. Для решения этой задачи могут использоваться разные подходы к сокращению размерности пространства признаков, например метод главных компонент, различные методы оценки значимости спутниковых индикаторов.

В настоящей работе проведён анализ множества методов оценки значимости признаков и проанализирована точность нескольких методов машинного обучения для решения задачи оценки продуктивности лесов с помощью многолетних данных инструмента MODIS (*англ. Moderate Resolution Imaging Spectroradiometer*).

Бонитет (продуктивность) лесных насаждений может быть оценён по спутниковым данным с использованием двух подходов. Первый подход основан на зависимости бонитета от условий произрастания, что позволяет связать его с различными характеристиками местности. Например, в одной из работ (Bjelanovic et al., 2018) исследовалась связь между бонитетом и рельефом, используя его как косвенный индикатор режима увлажнения. В другой работе (Сочилова и др., 2018) информация о рельефе вместе со спектральными отражательными характеристиками земного покрова по данным Proba-V применялась для картографирования классов бонитета лесов Приморского края. Второй подход основан на использовании многолетних наблюдений для оценки изменений биомассы лесов, отражающих их продуктивность. В ряде исследований показана возможность оценки бонитета лесов исходя из дистанционных наблюдений их биомассы в течение нескольких лет (Huuva et al., 2023; Penner et al., 2023).

В настоящей работе используется второй подход, предполагающий оценку бонитета лесов на основе анализа большого набора многолетних признаков, что обусловило необходимость решения задачи выбора их оптимального подмножества. С помощью отобранных наиболее значимых признаков проанализированы несколько методов машинного обучения для выбора наилучшего. Использование набора выбранных информативных признаков и наилучшего из рассмотренной совокупности метода машинного обучения позволило провести дистанционную оценку бонитета лесов в масштабах страны.

Данные

Оценка бонитета лесов на уровне страны требует наличия значительного объёма репрезентативных обучающих данных, характеризующих продуктивность всего многообразия лесных экосистем на территории страны. Эти данные для всей территории страны получены по материалам актуализированной цифровой картографической основы, созданной в ходе

1-го цикла государственной инвентаризации лесов (ГИЛ), содержащей сведения о характеристиках лесной растительности, в том числе о классе бонитета, на уровне лесотаксационных выделов. Данные предоставлены ФГБУ «Рослесинфорг» в рамках важнейшего инновационного проекта государственного значения «Российская система климатического мониторинга климатически активных веществ (ВИП ГЗ)». Достаточно большая давность последнего обновления этих данных не является критическим препятствием для их использования в качестве обучающих ввиду достаточно медленного изменения бонитета лесов. Предварительный анализ обучающих данных показал наличие ряда аномалий, для коррекции которых была выполнена процедура фильтрации, включающая несколько этапов:

1. Сравнение с полученными ранее спутниковыми картами растительного покрова России (Барталев и др., 2016), доступными в системе «БЕГА-Science» (Лупян и др., 2021), и картой лесов по данным проекта Global Forest Change (Hansen et al., 2013) (<https://glad.earthengine.app/view/global-forest-change>). Информация не включалась в обучающую выборку, если она расходилась с оценками по данным ДЗЗ, например, если на этой территории по спутниковой карте отсутствовал лес или тип леса (хвойные, широколиственные, лиственные) не совпадал по спутниковой и наземной информации.
2. Исключение областей, на которых по данным спутниковых наблюдений в период 2001–2022 гг. были выявлены нарушения, такие как пожары или рубки. Участки нарушений лесов за период 2000–2022 гг. детектировались по данным проекта Global Forest Change (Hansen et al., 2013).
3. Фильтрация на основе моделей хода роста лесов. Для этого по повыведельным данным лесоустройства строились модели хода роста лесов (Швиденко и др., 2008). При этом лесные выделы, ход роста лесов в которых характеризовался отклонением от их модельной динамики более чем на два класса бонитета, исключались из дальнейшего рассмотрения.
4. Экспертное выявление и фильтрация очевидных ошибок. Например, в Центральной Сибири из выборки были исключены лесные насаждения наивысшей продуктивности (бонитет Ib), нехарактерные для данного региона.

Прошедшие фильтрацию повыведельные данные были осреднены на уровне пикселей MODIS (разрешение 230 м). На этом этапе исключались неоднородные пиксели, в границах которых по наземным данным наблюдалось более одного значения классов бонитета. В результате обработки была получена обучающая выборка, включающая более чем 28 млн пикселей с информацией о классе бонитета лесов.

Для валидации финальной карты бонитета лесов использовались данные 4868 таксационно-дешифровочных участков, полученные в ходе 1-го цикла ГИЛ.

Выбор признаков для оценки бонитета лесов основан на предположении, что он определяет многолетнюю динамику роста насаждений и может быть оценён по изменению запасов стволовой древесины и других характеристик лесов. В связи с этим для оценки бонитета использовался набор признаков за период наблюдения спутниковой системой MODIS с 2001 по 2022 г., включая:

1) очищенные от влияния облачного покрова ежедневные временные ряды изображений земной поверхности, реконструированные на основе методов фильтрации и интерполяции данных съёмки спутниковой системой MODIS (Миклашевич и др., 2019). Ввиду большого объёма данных использовались не все ежедневные наблюдения, а их подмножество, отобранное для решения задач классификации растительного покрова (Барталев и др., 2016). Использовались коэффициенты спектральной яркости в четырёх спектральных каналах MODIS (красном (0,62–0,67 мкм), ближнем инфракрасном (ИК) (0,84–0,88 мкм), голубом (0,46–0,48 мкм) и среднем ИК (1,63–1,65 мкм)), полученные в четырёх временных срезах: 165, 195, 225 и 255-й дни года;

2) многолетние ряды данных о характеристиках лесов России, полученные на основе обработки временных рядов спутниковых изображений, включая информацию о преобла-

дающей породе (Барталев и др., 2016), запаса стволовой древесины (Ворушилов и др., 2022), лесистости территории и относительной полноте насаждений (Ховратович, 2022);

3) информацию о рельефе, основанную на цифровой модели рельефа ASTER (*англ.* Advanced Spaceborne Thermal Emission and Reflection Radiometer) (Tachikawa et al., 2011).

Полный набор использованных спутниковых индикаторов включал 669 признаков, характеризующих динамику лесов за 22 года.

Методы

Оценка значимости признаков

Для оценки бонитета лесов России был выбран набор из 669 признаков, для которых получена обучающая выборка из более чем 28 млн элементов. Большой объём данных затрудняет анализ разных методов машинного обучения с настройкой их параметров. В связи с этим возникает вопрос о выборе оптимального поднабора признаков, обеспечивающего достаточную точность построения регрессии в условиях ограниченных вычислительных ресурсов.

Выбор оптимального набора признаков основывался на использовании двух методов построения регрессии, а именно методе случайного леса в реализации пакета *ranger* (Wright, Ziegler, 2017) и метода градиентного бустинга *LightGBM* (*англ.* Light Gradient Boosting Machine — *LGBM*) (Ke et al., 2017).

Для выбора оптимального набора признаков применялись девять методов оценки значимости признаков, которые можно разделить на три группы.

Первая группа методов основана на использовании информации, полученной при построении ансамблей деревьев решений. Деревья решений состоят из последовательности узлов, в каждом из которых происходит разделение исходной выборки на основе некоторого набора признаков. Большое число узлов и деревьев в ансамбле даёт возможность оценить, насколько хорошо каждый признак позволяет разделить разные значения целевой характеристики (бонитета) в обучающей выборке, что может использоваться для оценки значимости признака. Этот метод широко применяется для случайного леса, но он может работать не оптимально для категориальных переменных, отдавая приоритет переменным с большим количеством категорий (Strobl et al., 2007). Например, в настоящей статье это может означать, что карта породного состава (24 категории или породы) может иметь заниженную значимость по сравнению с коэффициентами спектральной яркости (1000+ категорий или уникальных значений).

Методы данной группы оценивались на примере трёх разных реализаций. Использовалась реализация случайного леса *ranger* (далее обозначаемая *ranger_impurity*), а также две реализации из библиотеки градиентного бустинга *LightGBM*. Одна из них основывается только на числе узлов, в котором использовался признак (*LGBM_split*), а вторая работает аналогично подходу случайного леса и оценивает качество разбиения выборки в каждом узле (*LGBM_gain*).

Вторая группа методов оценки значимости признаков основана на анализе уже построенной регрессии. Методы этой группы предполагают изменение одного признака с помощью случайной перестановки его значений с анализом изменения точности классификации/регрессии при использовании нового набора данных. Если случайная перестановка одного признака привела к большому падению точности, то можно сделать вывод о его высокой значимости. В литературе отмечено два недостатка этого подхода (Hooker et al., 2021). Во-первых, он может переоценивать значимость признаков с высокой взаимной корреляцией. Во-вторых, такая перестановка значений одного признака может приводить к созданию элементов, отсутствующих в исходной выборке и не имеющих аналогов в реальности, что также может ухудшить оценку значимости.

Для этой группы методов были проанализированы три разных реализации. Во-первых, использовалась стандартная функция пакета *scikit-learn* (Pedregosa et al., 2011) — *permuta-*

tion_importance, которая применялась для двух алгоритмов машинного обучения: случайного леса (ranger_permutation_1) и LightGBM (LGBM_permutation). Также использовался встроенный в библиотеку ranger метод перестановки признаков для случайного леса (ranger_permutation_2). В каждой реализации регрессия строилась по обучающей выборке, а значимость проверялась по отложенной части выборки.

Третья группа включает три дополнительных метода оценки значимости признаков. Первый метод данной группы основан на формировании оптимального набора признаков с помощью метода главных компонент (метод PCA (*англ.* principal component analysis)), использующего преобразование координат исходного признакового пространства. Преобразование выполняется таким образом, чтобы вариация значений признаков вдоль первой новой оси была максимизирована, после чего эта ось исключается из анализа и процесс повторяется итеративно. В результате такого преобразования формируется новый набор признаков, естественным образом упорядоченных по их значимости (вариации). В анализе использовалась реализация метода PCA из библиотеки scikit-learn (Pedregosa et al., 2011).

Кроме того, анализировались ещё два более простых метода выбора оптимального набора признаков. Один из них — случайный выбор признака для добавления в текущий набор (random). Метод ранжирует признаки по значимости итеративно, случайным образом выбирая «самый значимый» признак из оставшегося набора, исключая его из списка кандидатов и повторяя этот процесс до исчерпания набора признаков. Второй метод основан на ранжировании признаков по величине их корреляции с бонитетом лесов (r2_importance). Все методы третьей группы настраивались по обучающей части выборки.

Вышеперечисленные девять методов оценки значимости были применены к исходному набору из 669 признаков. Каждый из этих методов позволил построить список признаков, упорядоченных от наиболее до наименее значимых. Для сравнения методов оценки значимости по N лучшим признакам строилась регрессия, и по отложенной выборке, не участвовавшей в оценке значимости и построении регрессии, оценивалась среднеквадратическая ошибка полученной регрессии, которая должна отражать качество (фактическую значимость) конкретного набора признаков.

Было отмечено, что при выборе признака целесообразно учитывать его корреляцию с уже включёнными в набор признаками. Если признак с максимальной значимостью сильно коррелирует с уже сформированным набором, то добавление другого признака может дать более высокую точность итоговой регрессии. Чтобы учесть этот фактор, оценивались все корреляции между признаком-кандидатом и каждым признаком в уже сформированном наборе, выбиралась максимальная из них, после чего для получения скорректированной значимости признака исходная её оценка умножалась на величину, равную $1 - \max R^2$. Использование скорректированной значимости признаков дало небольшое увеличение точности построенной регрессии (рис. 1).

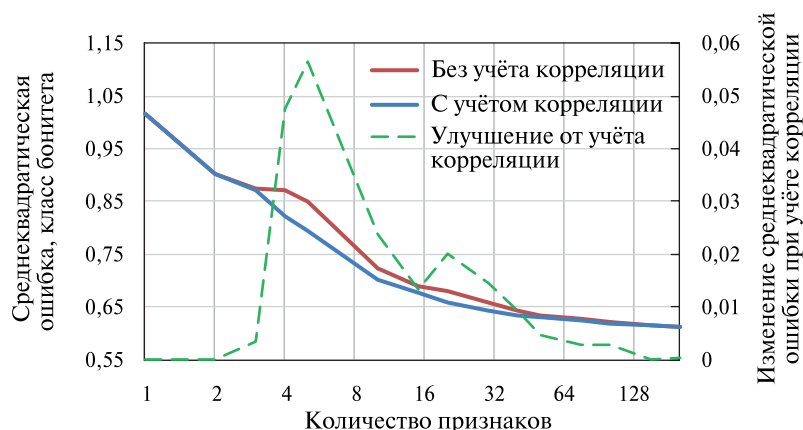


Рис. 1. Сравнение величины ошибки в зависимости от числа выбранных признаков (метод LGBM_gain) с использованием коррекции коэффициента корреляции и без неё

Методы машинного обучения

Несмотря на многообразие методов машинного обучения, на данный момент отсутствуют однозначные рекомендации по выбору оптимального метода для решения конкретной задачи. В связи с этим в данной работе был проведён анализ нескольких широко используемых методов машинного обучения и выполнено сравнение их точности для оценки бонитета лесов России. Для сравнения методов была сформирована обучающая выборка, состоящая из 1 млн элементов, и аналогичная ей по размеру контрольная выборка, половина которой использовалась для настройки методов, а вторая половина — для их сравнения между собой.

Предварительно была проведена оценка целесообразности применения библиотек автоматического машинного обучения, таких как AutoGluon (Erickson et al., 2020), H_2O automl (LeDell, Poirier, 2020) и Flaml (Wang et al., 2021). Однако все протестированные методы автоматического машинного обучения показали результаты не лучше, чем метод случайного леса. Вероятно, это может свидетельствовать о том, что стандартные настройки этих библиотек не оптимальны для решения конкретной задачи оценки продуктивности лесов с использованием данных ДЗЗ. Но при их применении было замечено, что наилучшие результаты дают методы на основе градиентного бустинга, что обосновывает целесообразность их дальнейшего более детального изучения.

Большинство опробованных методов — линейные регрессии, k ближайших соседей и опорных векторов — были взяты в реализации библиотеки scikit-learn (Pedregosa et al., 2011). Для метода случайного леса использовалась реализация из библиотеки ranger (Wright, Ziegler, 2017). Эта реализация показала более высокую скорость обучения в сравнении с реализацией из scikit-learn и позволила реализовать метод оценки значимости ranger_permutation_2, описанный в предыдущем разделе. Также использовались методы градиентного бустинга на основе деревьев решений в реализации библиотек XGBoost (Chen, Guestrin, 2016) и LightGBM (Ke et al., 2017). Была проведена настройка параметров всех методов машинного обучения. Для линейных регрессий применялся глобальный поиск на основе метода differential_evolution библиотеки scikit-learn, для более медленных методов градиентного бустинга использовался скалярный поиск оптимума по каждому параметру.

Для регрессионного анализа, основанного на ансамбле деревьев решений, характерна проблема исключения высоких и низких значений целевой характеристики (в данной работе — бонитета), что может приводить к искажению результирующей гистограммы, её «сужению». Для решения этой проблемы использовалась коррекция результата работы таких ансамблей на основе метода Empirical Distribution Matching (Belitz, Stackelberg, 2021).

Результаты

Результаты сравнения разных методов оценки значимости признаков представлены на рис. 2 (см. с. 70). Показанная на рисунке динамика среднеквадратической ошибки регрессии по разному числу признаков для различных методов оценки значимости позволяет делать выводы о качестве их работы. Наименее точными оказались случайный метод random и LGBM_split; наиболее точными — все подходы на основе случайной перестановки признака (LGBM_permutation, ranger_permutation_1, ranger_permutation_2), а также метод с использованием оценки точности признака в каждом узле (LGBM_gain). Второй метод, основанный на внутренней оценке значимости по деревьям решений (ranger_impurity), показал среднюю точность.

Также стоит отметить, что метод на основе корреляции ($r^2_{importance}$) продемонстрировал наилучшую точность оценки значимости на одном-двух признаках, но он оказался наименее точным при использовании 30 и более признаков. Это может быть связано с наличием значимых признаков, не коррелирующих с бонитетом лесов, например карт древесных пород и рельефа местности. Метод PCA также показывает хороший результат на небольшом числе признаков (до 60), но при дальнейшем их увеличении его погрешность начинает возрастать,

и он оказывается единственным методом, не выходящим на уровень среднеквадратической ошибки, равной 0,61.

На основе проведённого анализа было решено использовать метод с наименьшим значением среднеквадратической ошибки — LGBM_permutation. Был выбран набор из 100 наиболее значимых признаков по оценке этого метода, который даёт среднеквадратическую ошибку, равную 0,615, что незначительно хуже, чем ошибка для полного набора из 669 признаков (0,611). Стоит отметить, что в этой работе исключение незначимых признаков не позволяет добиться среднеквадратической ошибки выше уровня, достигаемого при использовании полного набора признаков. Это может быть связано с тем, что все признаки в той или иной степени характеризуют целевую характеристику и в наборе нет шумовых признаков.

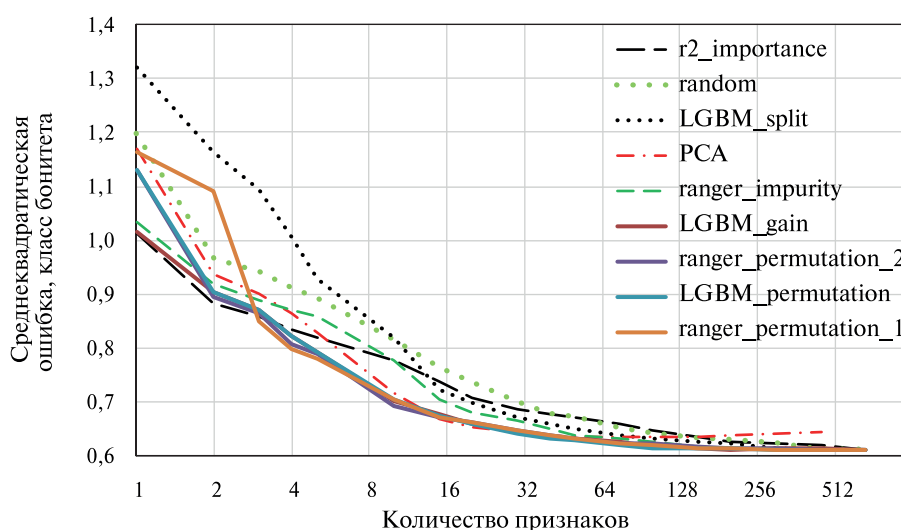


Рис. 2. Среднеквадратическая ошибка оценки класса бонитета лесов в зависимости от количества использованных наиболее значимых признаков, выбранных с помощью разных методов оценки значимости

Сравнение точности разных методов построения регрессии

Метод		R^2	RMSE	Время обучения, мин
Линейная регрессия	Базовая	0,65	0,81	0,05
	Ridge			0,05
	Lasso			2
	Elastic Net			2
<i>k</i> ближайших соседей		0,70	0,76	—
Опорных векторов		0,73	0,72	—
Случайного леса		0,776	0,65	75
XGBoost		0,783	0,64	5
LightGBM		0,816	0,59	8

Примечание: Курсивом выделена точность методов, проанализированных по ограниченной части выборки.

Набор из 100 наиболее значимых признаков применялся для сравнения разных методов машинного обучения, результаты которого приведены в таблице. Линейные регрессии работают устойчиво хуже других методов, а использование их вариантов с регуляризацией, таких как Elastic Net, не даёт изменения точности. В то же время при выбранном наборе призна-

ков и объёме выборки методы k ближайших соседей и опорных векторов работают слишком медленно, что не позволило выполнить их настройку и включить в сравнительный анализ. Их проверка на выборке небольшого объёма из 10 000 элементов показала, что метод опорных векторов демонстрирует несколько лучшие результаты по сравнению со случайным лесом (среднеквадратическая ошибка RMSE (англ. Root Mean Square Error) 0,72 против 0,73), а k ближайших соседей — хуже (0,76). Методы с применением деревьев решений дают хорошие результаты с явным преимуществом методов на основе градиентного бустинга. Среди последних наилучший результат показал LightGBM, который и было решено использовать для оценки бонитета лесов России. Можно отметить, что время обучения LightGBM превысило время обучения XGBoost, хотя первый метод включает ряд оптимизаций, направленных на ускорение построения модели. Предположительно это связано с выбранными настройками метода, приведшими к построению более глубоких деревьев с большим числом узлов для метода LightGBM.

Выполненный анализ позволил определить оптимальный подход к оценке бонитета лесов на основе спутниковых данных ДЗЗ. Ранжирование признаков по их информативности с применением случайной перестановки входных признаков дало возможность выбрать их подмножество из 100 наиболее значимых признаков с минимальной потерей точности восстановления регрессии для оценки показателя бонитета лесов. Регрессия между признаками и значениями бонитета лесов строилась на основе метода LightGBM, показавшего наилучшие результаты среди включённых в анализ методов машинного обучения. Для обучения использовалась выборка из 25 млн пикселей, а 3 млн исключённых из обучения пикселей резервировались для оценки точности.

Результат верифицировался по контрольной выборке из 3 млн пикселей, а также по наземным данным, полученным на таксационно-дешифровочных участках ГИЛ, описанных выше в разд. «Данные». Построенная регрессия отличается высоким уровнем точности с коэффициентом детерминации R^2 , равным 0,87, и среднеквадратической ошибкой 0,5 класса бонитета лесов. Сравнение с данными таксационно-дешифровочных участков ГИЛ показывает, что в 83 % случаев предложенный подход позволяет корректно определять бонитет лесов с погрешностью, не превышающей 1-й класс бонитета. В целом проведённый анализ демонстрирует достаточно высокую точность полученных оценок бонитета лесов.

Заключение

Выполненный анализ позволил выбрать оптимальный набор наиболее информативных признаков и наилучший метод машинного обучения для решения задачи оценки продуктивности лесов страны по данным ДЗЗ и построить карту их бонитета.

Анализ множества методов выбора информативных индикаторов показал преимущество подхода на основе случайной перестановки входных признаков, позволившего построить регрессию с высоким уровнем точности при меньшем их количестве. Среди протестированных методов машинного обучения было выявлено преимущество подходов с использованием градиентного бустинга, в частности, его реализации LightGBM.

Стоит отметить, что сделанные по результатам исследования выводы об оптимальных методах и признаках применимы только к задаче оценки бонитета лесов России на основе данных MODIS. В связи с этим в дальнейшем целесообразно провести подобный анализ для других характеристик лесов страны с формированием более универсальных рекомендаций.

Получение спутниковых данных и их обработка были проведены с использованием инфраструктуры Центра коллективного пользования системами архивации, обработки и анализа данных спутниковых наблюдений ИКИ РАН для решения задач изучения и мониторинга окружающей среды (Лупян и др., 2019).

Анализ методов машинного обучения и методов оценки значимости признаков выполнялся с помощью разработанного в ИКИ РАН программного обеспечения FORS-MachLearn (Барталев, Хвостиков, 2023).

Работа выполнена в рамках реализации важнейшего инновационного проекта государственного значения «Разработка системы наземного и дистанционного мониторинга пулов углерода и потоков парниковых газов на территории Российской Федерации, обеспечение создания системы учёта данных о потоках климатически активных веществ и бюджете углерода в лесах и других наземных экологических системах» (рег. № 123030300031-6).

Литература

1. *Барталев С. А., Хвостиков С. А.* FORS-MachLearn. Свидетельство о гос. регистрации программы для ЭВМ № 2023666251 Рег. 27.07.2023.
2. *Барталев С. А., Егоров В. А., Жарко В. О., Лупян Е. А., Плотников Д. Е., Хвостиков С. А., Шабанов Н. В.* Спутниковое картографирование растительного покрова России. М.: ИКИ РАН, 2016. 208 с.
3. *Ворушилов И. И., Барталев С. А., Егоров В. А.* Оценка динамики запаса стволовой древесины нарушенных территорий России // Материалы 20-й Международной конф. «Современные проблемы дистанционного зондирования Земли из космоса». М.: ИКИ РАН, 2022. С. 290. DOI: 10.21046/20DZZconf-2022a.
4. *Лупян Е. А., Прошин А. А., Бурцев 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.
5. *Лупян Е. А., Прошин А. А., Бурцев М. А. и др.* Система «Вега-Science»: особенности построения, основные возможности и опыт использования // Современные проблемы дистанционного зондирования Земли из космоса. 2021. Т. 18. № 6. С. 9–31. DOI: 10.21046/2070-7401-2021-18-6-9-31.
6. *Миклашевич Т. С., Барталев С. А., Плотников Д. Е.* Интерполяционный алгоритм восстановления длинных временных рядов данных спутниковых наблюдений растительного покрова // Современные проблемы дистанционного зондирования Земли из космоса. 2019. Т. 16. № 6. С. 143–154. DOI: 10.21046/2070-7401-2019-16-6-143-154.
7. *Сочилова Е. Н., Сурков Н. В., Ершов Д. В. и др.* Картографирование классов бонитета лесов Приморского края на основе спутниковых изображений и данных о характеристиках рельефа // Современные проблемы дистанционного зондирования Земли из космоса. 2018. Т. 15. № 5. С. 96–109. DOI: 10.21046/2070-7401-2018-15-5-96-109.
8. *Ховратович Т. С.* Показатели горизонтальной структуры лесов и их дистанционная оценка на основе оптических спутниковых данных (лекция) // Материалы 20-й Международной конф. «Современные проблемы дистанционного зондирования Земли из космоса». М.: ИКИ РАН, 2022. С. 505. DOI: 10.21046/20DZZconf-2022a.
9. *Швиденко А. З., Щепашенко Д. Г., Нильссон С., Булуй Ю. И.* Таблицы и модели хода роста и продуктивности насаждений основных лесообразующих пород Северной Евразии (нормативно-справочные материалы), 2-е изд., доп. М.: Федер. агентство лесного хоз-ва, 2008. 886 с.
10. *Belitz K., Stackelberg P. E.* Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models // Environmental Modelling and Software. 2021. V. 139. Article 105006. DOI: 10.1016/j.envsoft.2021.105006.
11. *Bjelanovic I., Comeau P., White B.* High resolution site index prediction in boreal forests using topographic and wet areas mapping attributes // Forests. 2018. V. 9. No. 3. Article 113. DOI: 10.3390/f9030113.
12. *Chen T., Guestrin C.* XGBoost: A scalable tree boosting system // Proc. 22nd ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016. P. 785–794. DOI: 10.1145/2939672.293978.
13. *Erickson N., Mueller J., Shirkov A. et al.* Autogluon-tabular: Robust and accurate AutoML for structured data // <https://arxiv.org/>. arXiv:2003.06505. 2020. 28 p. DOI: 10.48550/arXiv.2003.06505.
14. *Hansen M. C., Potapov P. V., Moore R. et al.* High-resolution global maps of 21st-century forest cover change // Science. 2013. V. 342. No. 6160. P. 850–853. DOI: 10.1126/science.1244693.
15. *Hooker G., Mentch L., Zhou S.* Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance // Statistics and Computing. 2021. V. 31. Article 82. 16 p. DOI: 10.1007/s11222-021-10057-z.
16. *Huuvu I., Wallerman J., Fransson J. E. S., Persson H. J.* Prediction of site index and age using time series of TanDEM-X phase heights // Remote Sensing. 2023. V. 15. No. 17. Article 4195. DOI: 10.3390/rs15174195.
17. *Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y.* LightGBM: A highly efficient gradient boosting decision tree // Advances in neural information processing systems 30: 31th Conf. Neural Information Processing Systems (NIPS 2017). 2017. V. 30. P. 3149–3157.

18. *Khatami R., Mountrakis G., Stehman S. V.* A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research // *Remote Sensing of Environment*. 2016. V. 177. P. 89–100. DOI: 10.1016/j.rse.2016.02.028.
19. *LeDell E., Poirier S.* H₂O AutoML: Scalable automatic machine learning // *Proc. AutoML Workshop at ICML*. ICML, 2020. 16 p.
20. *Maxwell A. E., Warner T. A., Fang F.* Implementation of machine-learning classification in remote sensing: an applied review // *Intern. J. Remote Sensing*. 2018. V. 39. No. 9. P. 2784–2817. DOI: 10.1080/01431161.2018.1433343.
21. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* Scikit-learn: Machine learning in Python // *J. Machine Learning Research*. 2011. V. 12. P. 2825–2830.
22. *Penner M., Woods M., Bilyk A.* Assessing site productivity via remote sensing — age-independent site index estimation in even-aged forests // *Forests*. 2023. V. 14. No. 8. Article 1541. DOI: 10.3390/f14081541.
23. *Sheykhoumousa M., Mahdianpari M., Ghanbari H. et al.* Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review // *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing*. 2020. V. 13. P. 6308–6325. DOI: 10.1109/JSTARS.2020.3026724.
24. *Strobl C., Boulesteix A. L., Zeileis A., Hothorn T.* Bias in random forest variable importance measures: Illustrations, sources and a solution // *BMC Bioinformatics*. 2007. V. 8. No. 1. Article 25. DOI: 10.1186/1471-2105-8-25.
25. *Tachikawa T., Hato M., Kaku M., Iwasaki A.* Characteristics of ASTER GDEM version 2 // 2011 IEEE Intern. Geoscience and Remote Sensing Symposium. Vancouver, BC, Canada: IEEE, 2011. P. 3657–3660. DOI: 10.1109/IGARSS.2011.6050017.
26. *Wang C., Wu Q., Weimer M., Zhu E.* FLAML: A fast and lightweight AutoML library // *Proc. Machine Learning and Systems 3 (MLSys 2021)*. 2021. V. 3. P. 434–447.
27. *Wright M. N., Ziegler A.* ranger: A fast implementation of random forests for high dimensional data in C++ and R // *J. Statistical Software*. 2017. V. 77. Iss. 1. P. 1–17. DOI: 10.18637/jss.v077.i01.

Selection of important variables and best machine learning method for estimation of forest productivity using remote sensing data

S. A. Khvostikov, S. A. Bartalev, V. A. Egorov, E. A. Stytsenko

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

Recent advances in machine learning have resulted in abundance of methods to evaluate vegetation parameters using remote sensing data. The objective of this research was to find best machine learning method and a group of most informative variables to evaluate forest productivity using MODIS data. Big size of training sample (28 million pixels) and use of 669 variables based on remote sensing data necessitated selection of a subgroup of most informative variables. Applying multiple methods of variable importance evaluation led to a conclusion that methods based on data permutations showed best performance on available data, and they were used to select a subgroup of 100 informative remote sensing indicators with negligible loss of accuracy. This reduced subgroup of indicators allowed us to analyze multiple machine learning methods. Of all tested methods, gradient boosting based on LightGBM showed the best performance. A combination of LightGBM and the subgroup of best remote sensing indicators was used to create a productivity map of Russian forests. Analysis of its accuracy on a control sample that was not used to train the LightGBM model showed a coefficient of determination of 0.87 and RMSE of 0.5 classes of productivity. Comparison with field data showed a good match with remote sensing productivity estimates in more than 80 % of cases.

Keywords: forest, forest productivity, remote sensing, variable importance, gradient boosting

Accepted: 05.06.2025

DOI: 10.21046/2070-7401-2025-22-4-64-75

References

1. Bartalev S. A., Khvostikov S. A., *FORS-MachLearn*, Certificate of state registration of software No. 2023666251 (RU), Reg. 27.07.2023 (in Russian).
2. Bartalev S. A., Egorov V. A., Zharko V. O., Loupian E. A., Plotnikov D. E., Khvostikov S. A., Shabanov N. V., *Sputnikovoe katrographirovanie rastitel'nogo pokrova Rossii* (Land cover mapping over Russia using Earth observation data), Moscow: IKI RAN, 2016, 208 p. (in Russian).
3. Vorushilov I. I., Bartalev S. A., Egorov V. A., Evaluation of growing stock dynamic for disturbed territory of Russia, *Materialy 20-i Mezhdunarodnoi konferentsii "Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa"* (Proc. 20th Intern. Conf. "Current Problems in Remote Sensing of the Earth from Space"), Moscow: IKI RAS, 2022, p. 290 (in Russian), DOI: 10.21046/20DZZconf-2022a.
4. 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, V. 16, No. 3, pp. 151–170 (in Russian), DOI: 10.21046/2070-7401-2019-16-3-151-170.
5. Loupian E. A., Proshin A. A., Bourtsev M. A. et al., Vega-Science system: design features, main capabilities and usage experience, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2021, V. 18, No. 6, pp. 9–31 (in Russian), DOI: 10.21046/2070-7401-2019-16-3-151-170.
6. Miklashevich T. S., Bartalev S. A., Plotnikov D. E., Interpolation algorithm for the recovery of long satellite data time series of vegetation cover observation, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2019, V. 16, No. 6, pp. 143–154 (in Russian), DOI: 10.21046/2070-7401-2019-16-6-143-154.
7. Sochilova E. N., Surkov N. V., Ershov D. V. et al., Mapping of forest site index classes in Primorskiy Krai based on satellite images and terrain characteristics, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2018, V. 15, No. 5, pp. 96–109 (in Russian), DOI: 10.21046/2070-7401-2018-15-5-96-109.
8. Khovrativich T. S., Indicators of horizontal forest structure and their remote assessment using optical satellite data (lecture), *Materialy 20-i Mezhdunarodnoi konferentsii "Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa"* (Proc. 20th Intern. Conf. "Current Problems in Remote Sensing of the Earth from Space"), Moscow: IKI RAS, 2022, p. 505 (in Russian), DOI: 10.21046/20DZZconf-2022a.
9. Shvidenko A. Z., Schepaschenko D. G., Nilsson S., Buluy Yu. I., *Tablitsy i modeli khoda rosta i produktivnosti nasazhdenii osnovnykh lesoobrazuyushchikh porod Severnoi Evrazii (normativno-spravochnye materialy)* (Tables and models of growth and productivity of forests of major forest forming species of northern Eurasia (standard and reference materials)), 2nd ed., Moscow: Federal'noe agentstvo lesnogo khozyaistva, 2008, 886 p. (in Russian).
10. Belitz K., Stackelberg P. E., Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models, *Environmental Modelling and Software*, 2021, V. 139, Article 105006, DOI: 10.1016/j.envsoft.2021.105006.
11. Bjelanovic I., Comeau P., White B., High resolution site index prediction in boreal forests using topographic and wet areas mapping attributes, *Forests*, 2018, V. 9, No. 3, Article 113, DOI: 10.3390/f9030113.
12. Chen T., Guestrin C., XGBoost: A scalable tree boosting system, *Proc. 22nd ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, San Francisco, California, USA: ACM, 2016, pp. 785–794, DOI: 10.1145/2939672.293978.
13. Erickson N., Mueller J., Shirkov A. et al., Autogluon-tabular: Robust and accurate AutoML for structured data, <https://arxiv.org/>, arXiv:2003.06505, 2020, 28 p., DOI: 10.48550/arXiv.2003.06505.
14. Hansen M. C., Potapov P. V., Moore R. et al., High-resolution global maps of 21st-century forest cover change, *Science*, 2013, V. 342, No. 6160, pp. 850–853, DOI: 10.1126/science.1244693.
15. Hooker G., Mentch L., Zhou S., Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance, *Statistics and Computing*, 2021, V. 31, Article 82, 16 p., DOI: 10.1007/s11222-021-10057-z.
16. Huuva I., Wallerman J., Fransson J. E. S., Persson H. J., Prediction of site index and age using time series of TanDEM-X phase heights, *Remote Sensing*, 2023, V. 15, No. 17, Article 4195, DOI: 10.3390/rs15174195.
17. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y., LightGBM: A highly efficient gradient boosting decision tree, *Advances in neural information processing systems 30: 31th Conf. Neural Information Processing Systems (NIPS 2017)*, 2017, V. 30, pp. 3149–3157.
18. Khatami R., Mountrakis G., Stehman S. V., A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research, *Remote Sensing of Environment*, 2016, V. 177, pp. 89–100, DOI: 10.1016/j.rse.2016.02.028.
19. LeDell E., Poirier S., H₂O AutoML: Scalable automatic machine learning, *Proc. AutoML Workshop at ICML, ICML*, 2020, 16 p.

20. Maxwell A. E., Warner T. A., Fang F., Implementation of machine-learning classification in remote sensing: an applied review, *Intern. J. Remote Sensing*, 2018, V. 39, No. 9, pp. 2784–2817, DOI: 10.1080/01431161.2018.1433343.
21. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É., Scikit-learn: Machine learning in Python, *J. Machine Learning Research*, 2011, V. 12, pp. 2825–2830.
22. Penner M., Woods M., Bilyk A., Assessing site productivity via remote sensing — age-independent site index estimation in even-aged forests, *Forests*, 2023, V. 14, No. 8, Article 1541, DOI: 10.3390/f14081541.
23. Sheykhmousa M., Mahdianpari M., Ghanbari H. et al., Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review, *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, V. 13, pp. 6308–6325, DOI: 10.1109/JSTARS.2020.3026724.
24. Strobl C., Boulesteix A. L., Zeileis A., Hothorn T., Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics*, 2007, V. 8, No. 1, Article 25, DOI: 10.1186/1471-2105-8-25.
25. Tachikawa T., Hato M., Kaku M., Iwasaki A., Characteristics of ASTER GDEM version 2, *2011 IEEE Intern. Geoscience and Remote Sensing Symposium*, Vancouver, BC, Canada: IEEE, 2011, pp. 3657–3660, DOI: 10.1109/IGARSS.2011.6050017.
26. Wang C., Wu Q., Weimer M., Zhu E., FLAML: A fast and lightweight AutoML library, *Proc. Machine Learning and Systems 3 (MLSys 2021)*, 2021, V. 3, pp. 434–447.
27. Wright M. N., Ziegler A., ranger: A fast implementation of random forests for high dimensional data in C++ and R, *J. Statistical Software*, 2017, V. 77, Iss. 1, pp. 1–17, DOI: 10.18637/jss.v077.i01.