Пространственная классификация преобладающих древесных пород на территории Самарской области по данным Sentinel-2 и таксации леса

А. Ю. Денисова[1], Л. М. Кавеленова[1], Е. С. Корчиков[1], Н. В. Прохорова[1], Д. А. Терентьева[1], В. А. Федосеев[1,2]

[1] *Самарский национальный исследовательский университет им. акад. С. П. Королева, Самара, 443086, Россия*

[2] *Институт систем обработки изображений РАН — филиал ФНИЦ «Кристаллография и фотоника» РАН, Самара, 443001, Россия*

E-mail: vicanfed@gmail.ru

Целью работы являлось исследование возможности построения классификатора преобладающих пород леса, обученного на данных Sentinel-2 за один сезон и обладающего точностью, достаточной для составления предварительной тематической разметки новых участков. Данная задача важна для выявления фрагментарно сохранившихся коренных лесных сообществ с целью придания им охранного статуса, а также для обнаружения ценных и редких в контексте определённой территории древесных пород в протяжённых лесных массивах. Преимущество спутников Sentinel-2 для решения этой задачи заключается в высокой частоте съёмки, а также большой площади покрытия территории. В статье освещаются вопросы подбора спутниковых данных, выбора алгоритма классификации и метода обучения классификатора, а также процедур пространственной обработки. Исследования основывались на данных космосъёмки за 2018 г. территории Красносамарского лесничества Самарской области. Для обучения и проверки классификаторов использовались данные таксации леса за 2013–2014 гг. Проведённые исследования позволили достичь высокой точности (около 0,82) на контрольном фрагменте территории, причём существенная доля ошибок классификации обусловлена изменениями на местности, не учтёнными в данных таксации, что подтвердили результаты наземных обследований. Таким образом, одним из результатов работы является выработанная технология уточнения таксации леса, которая, как показали исследования, имеет определённый практический потенциал.

Ключевые слова: классификация пород леса, Sentinel-2, пространственная классификация, таксация леса, GEM, SVM

Одобрена к печати: 17.05.2019

DOI: 10.21046/2070-7401-2019-16-4-86-101

Введение

Около 46 % площади России покрыта лесной растительностью (Госдоклад..., 2017), что делает леса одним из наиболее представительных типов растительного покрова России. На землях лесного фонда и в лесах, не входивших в лесной фонд ранее, до принятия нового Лесного кодекса, проводилось плановое периодическое лесоустройство, которое создавало базовую учётную информацию о лесах: таксационное описание и планшет. Системообразующим видом лесоучётных работ являлось лесоустройство по первому, второму и третьему разрядам точности, которое дополнялось лесоинвентаризацией, проводимой на основе аэрокосмических методов (Мальков, Закамский, 2006; Таранков, 2006). С 2006 г. в части оценки состояния лесов страны лесоустройство заменила Государственная инвентаризации лесов (ГИЛ) (Госдоклад..., 2017). Однако данные ГИЛ, как и ранее полученные данные периодического лесоустройства, являются основой ведения Государственного лесного реестра страны (ГЛР), разработки лесных планов субъектов РФ, лесохозяйственных регламентов лесничеств/лесопарков и проектов освоения лесов (Галеева..., 2016; Положение..., 1995), а значит, составление таксационных описаний по-прежнему играет важную роль в управлении лесным хозяйством.

Таксационное описание содержит исчерпывающую информацию о лесной территории, например: преобладающие и сопутствующие древесные породы, границы участков, имеющих однородный породный состав, характеристики лесорастительных условий и состояние древесных пород (Троц и др., 2015). Однако получение таксационного описания продолжает оставаться дорогостоящей и, как следствие, редко проводимой процедурой при использовании как традиционных методов, так и данных аэрофотосъёмки (Любимов и др., 2018). Поэтому при анализе таксационных данных целесообразно применение данных дистанционного зондирования Земли (ДЗЗ), позволяющих актуализировать информацию для сравнительно больших участков территории с меньшими затратами.

Современная тенденция в классификации породного состава лесов по данным ДЗЗ состоит в использовании разнородных систем и сенсоров, например гиперспектральных данных ДЗЗ и данных авиационного лазерного сканирования (Dalponte et al., 2014), а также беспилотных летательных аппаратов (Puliti et al., 2015). Однако более доступным вариантом является использование данных среднего и низкого пространственного разрешения в силу меньшей стоимости и большего охвата территории (Corbane et al., 2015). Например, применение данных низкого разрешения сенсора MODIS космических аппаратов (далее — КА) Terra/Aqua для классификации лесов глубоко исследовано в работах (Барталев, 2007; Жарко, Барталев, 2014; Waring et al., 2006 и др.). Примеры успешного использования данных среднего разрешения (КА Landsat и SPOT) для задачи классификации лесов описаны в публикациях (Барталев и др., 1995; Курбанов и др., 2013; Перепечина и др., 2016; Karlson et al., 2015; Xiao et al., 2002 и др.), где показаны существенные возможности снимков среднего разрешения для задачи классификации лесов.

С запуском спутников Sentinel-2A/2B в 2015–2017 гг. появилась возможность использовать новый информационный ресурс для классификации лесов. Данные Sentinel-2 являются открытыми и содержат 13 спектральных каналов с разрешением от 10 до 60 м, что в сочетании с высокой периодичностью съёмки (не реже 5 дней) и широкой полосой обзора делает их ценным источником информации (Drusch et al., 2012).

Решение задачи классификации леса по данным Sentinel-2 рассматривалось в ряде работ последних лет (Кедров, Тарасов, 2017; Перепечина и др., 2016; Immitzer et al., 2016). Однако в них авторы ограничивались анализом отдельных снимков, а также, как правило, использовали ограниченный набор каналов (ближний инфракрасный, зелёный и синий) для классификации. В результате влияние сезонных изменений характеристик лесов в различных спектральных каналах на точность классификации не было исследовано.

В отличие от упомянутых выше работ, в настоящей статье рассматриваются различные периоды съёмки и состав спектральных каналов данных Sentinel-2 для классификации породного состава леса, а для обучения используются результаты таксационных обследований. Для этого составлялось многоканальное изображение (далее — сезонный композит), представлявшее собой стек (совокупность) разновременных и разноканальных изображений, сформированных из безоблачных для тестового участка снимков Sentinel-2. В качестве метода классификации по многомерному сезонному композиту снимков в статье рассматриваются различные алгоритмы и приёмы пространственной классификации, представленные в работах (Борзов, Потатуркин, 2014; Fedoseev, 2018) и использованные в оригинальных статьях для анализа гиперспектральных данных.

Описание исходных данных

Для исследования в нашей работе была выбрана территория Красносамарского лесничества, расположенного в Кинельском районе Самарской области. Как отмечено выше, исходными данными ДЗЗ для классификации древесных пород служили композиты, составленные из снимков Sentinel-2 за период с апреля по октябрь 2018 г. Для разметки территории использовались данные таксации леса за 2013–2014 гг., доступные в виде векторной карты полигональных объектов в ГИС «ИнГео».

Для подготовки композита отбирались безоблачные снимки уровня обработки L2A (с атмосферной коррекцией) и L1C (без атмосферной коррекции). Использовались четыре 10-метровых канала (Band 2–4 и 8: Blue, Green, Red, NIR), а также четыре 20-метровых канала инфракрасного диапазона (Band 5–7 и 8A) (Drusch et al., 2012), для которых выполнялось повышение разрешения до 10 м с помощью пакета Scanex Image Processor 4.2. Таким образом, полный композит за весь вегетационный сезон включал 136 каналов и был составлен из 17 снимков (даты снимков указаны ниже в *табл. 4*). Также в некоторых исследованиях использовались данные NDVI (Normalized Difference Vegetation Index — нормализованный относительный индекс растительности) (вместе с основными каналами или вместо них). Значения NDVI рассчитывались по каналам Band 4 и Band 8A в соответствии с (Zhang et al., 2017).

Для расчёта векторов признаков исходные каналы композитов преобразовывались к главным компонентам с возможностью дальнейшего сокращения размерности, после чего для улучшения процедуры обучения выполнялась дополнительная нормализация значений к диапазону [-1; 1]. При обучении модели использовалось фиксированное число эталонных экземпляров (пикселей) на каждый класс по 2000 пикселей, отбиравшихся случайным образом из северной части лесничества, включавшей 647 151 пикселей. Тем самым мы умышленно не учитывали различную вероятность появления различных классов, чтобы исследовать количественные характеристики применения обученного классификатора на неизвестной территории, где априорные вероятности неизвестны. Для оценки точности классификатора использовалось 118 284 пикселей из южной части лесничества.

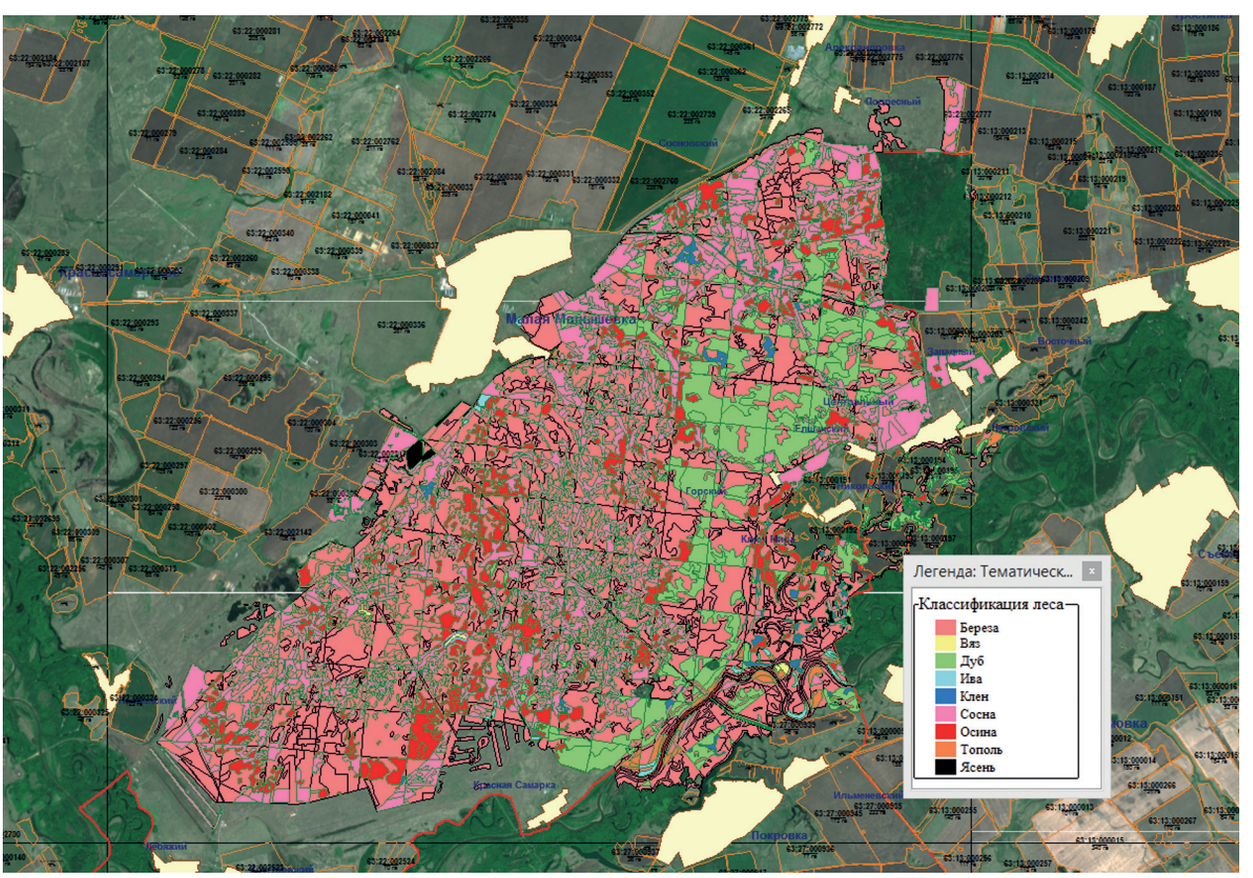


Рис. 1. Тематическая карта данных таксации Красносамарского лесничества, использованных в исследовании

Из таксационных данных посредством семантических запросов были отобраны участки леса, содержащие не менее 8/10 доли одной породы. Отобранные участки сформировали

9 классов пород: берёза, вяз, дуб, ива, клён, осина, сосна, тополь, ясень (*рис. 1*). Далее полученные участки были переведены в растровую форму (маску классов пород леса) и впоследствии использовались для обучения и тестирования классификатора. Пространственное разрешение маски классов соответствовало разрешению используемого композита снимков. Отметим также, что тестовая территория содержала 46 % выделов сосны, 21 % — дуба, 19 % — осины, 10 % — берёзы и около 4 % в сумме приходилось на остальные породы.

Краткое описание используемых методов анализа данных

На *рис. 2* наглядно представлена схема используемой в настоящей статье процедуры классификации, учитывающей пространственные и спектральные связи в анализируемых данных (далее будем называть её коротко процедурой пространственной классификации) и построенной на основе результатов работы (Fedoseev, 2018). Данная процедура включает следующие этапы: пространственная предобработка, кластеризация, сегментация, поэлементная классификация, агрегация результатов классификации и сегментации, пространственная постобработка.

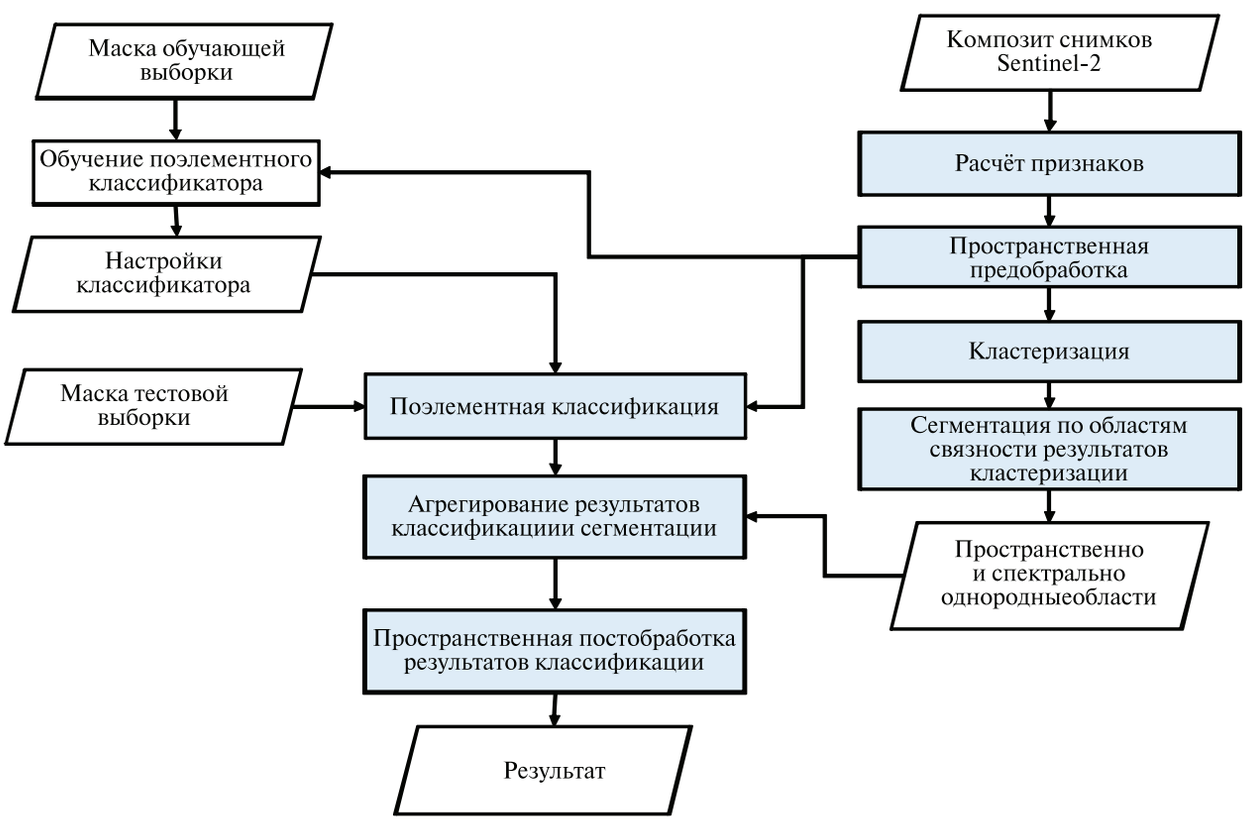


Рис. 2. Схема процедуры классификации

Для пространственной предобработки, имеющей целью сокращение уровня шумов в исходных данных, использовался медианный фильтр с небольшим окном (3×3, 5×5).

Далее осуществлялась кластеризация для выделения из исходных данных подмножеств, сходных по признаковому описанию. Рассматривались два алгоритма кластеризации: kmeans++ (Arthur, Vassilvitskii, 2007), хорошо показавший себя в работах (Зимичев и др., 2014; Fedoseev, 2018), а также обобщённый EM-алгоритм (Generalized Expectation Maximization — GEM), предложенный в работе (Denisova, Sergeyev, 2018). Поскольку получаемые в результате кластеризации области учитывают характер взаимосвязей только в признаковом пространстве

без учёта пространственных зависимостей, то на этапе сегментации после кластеризации выполнялось разбиение кластеров на отдельные компоненты связности (сегменты).

Поэлементная классификация, являющаяся неотъемлемой частью рассматриваемой процедуры пространственной классификации, выполнялась при помощи метода опорных векторов с радиальными базисными функциями (SVM-RBF). Это обусловлено тем, что данная модель показала себя одной из наиболее эффективных для классификации растительности по многоканальным снимкам ДЗЗ в ряде работ (Борзов, Потатуркин, 2014; Зимичев и др., 2014; Кузнецов, Мясников, 2014). Многоклассовая модель SVM-RBF сформирована из множества бинарных классификаторов для каждой пары пород леса. Для обучения каждого из них перебирались несколько вариантов значений параметра регуляризации C и параметра ядра RBF γ (Cristianini, Shawe-Taylor, 2000).

Результатом последующего этапа агрегации (объединения) результатов и сегментации является маска классов, однородная по сегментам (т. е. маска, в которой всем пикселям каждого из сегментов присвоена одна и та же метка). Исследовались два способа агрегации: голосование большинства и классификация медиан. В первом способе (Зимичев и др., 2014) поэлементная классификация применялась ко всем пикселям тестового изображения, а далее метка класса выбиралась для каждого сегмента голосованием. Во втором способе (Борзов, Потатуркин, 2014) рассчитывались вектора признаков, являющиеся медианами по каждому из сегментов, и уже для них выполнялась процедура классификации методом SVM-RBF.

Заключительным этапом является пространственная постобработка результатов классификации, предназначенная для коррекции единичных неправильно проклассифицированных пикселей. Она осуществлялась путём обработки скользящим окном по следующему принципу: если количество пикселей окна, относящихся к доминирующему классу, превышает порог T и номер класса для центрального пикселя окна не совпадает с доминирующим классом, то значение центрального пикселя окна заменяется номером доминирующего класса. В экспериментах использовался размер окна 3×3 и $T = 5$.

Рассмотрим чуть подробнее алгоритм кластеризации GEM (Denisova, Sergeyev, 2018). Как и EM-алгоритм, он использует предположение, что в каждом кластере данные распределены по нормальному закону и глобальное распределение описывается смесью гауссовых распределений. При этом число и параметры кластеров полагаются неизвестными и определяются в процессе работы алгоритма, а метки кластеров присваиваются по максимуму условной вероятности попадания в кластер.

GEM итерационно уточняет количество кластеров и их параметры с использованием глобальной гистограммы изображения, представленной в виде иерархической структуры данных (гистограмма-дерево). Начальное количество кластеров оценивается по приближённому списку мод гистограммы, извлекаемому из высших уровней иерархии. Затем выполняются EM-итерации с использованием гистограммы-дерева. Кластеры с вырожденной корреляционной матрицей удаляются из набора кластеров. Гистограмма-дерево позволяет выполнять итерации алгоритма в многомерном случае значительно быстрее, чем для его стандартной реализации. Критерием останова является стабилизация количества и положения центров кластеров.

Основным параметром алгоритма является минимальная частота попадания пикселей в ячейку гистограммы (Minimum Frequency Value — MFV), которая определяет точность используемой аппроксимации глобальной гистограммы: ячейки с частотой меньше MFV полагаются маловероятными и отбрасываются. На основе исследований, проведённых в работе (Bavrina et al., 2018), значение MFV выбиралось равным $0{,}0005N$, где N — общее число пикселей территории. Отметим также, что для повышения быстродействия кластеризация выполнялась по 17-канальному композиту, составленному только из значений NDVI по каждому из снимков.

Экспериментальные исследования

Сравнительный анализ точности классификации в зависимости от параметров алгоритма, а также выбора и способа подготовки входных данных

Экспериментальные исследования, имевшие целью выбор параметров алгоритма классификации, а также данных и способов их подготовки, проводились в несколько этапов. Так, эксперимент 1 был посвящён выбору параметров процедуры пространственной классификации, наиболее подходящих для используемых данных. Для этого в качестве векторов признаков использовался полный набор из 136 главных компонент, рассчитанных по 136 каналам композита продуктов уровня L2A за весь сезон с 12.04.2018 по 19.10.2018. Сокращение размерности в данном эксперименте не проводилось, чтобы избежать потери потенциально значимой информации, а также для того, чтобы затем отдельно исследовать изменение точности классификации при сокращении пространства признаков (эксперимент 4). Перед расчётом главных компонент выполнялась предварительная медианная фильтрация окном 3×3.

Результаты эксперимента отражены в *табл. 1*. Число сегментов и точность классификации являются рассчитываемыми значениями, в то время как остальные столбцы содержат варьируемые параметры. При этом точность классификации во всех экспериментах оценивалась на уровне пикселей как общая вероятность правильной классификации по всем классам. Варианты 1 и 2 показывают результаты поэлементной классификации, при которой не выполнялся анализ сегментов. Приведённые результаты показали преимущество кластеризации GEM, метода голосования большинства при классификации сегментов, а также целесообразность оконной постобработки результатов классификации. Точность наилучшего результата (вариант 5) на 0,0743 превысила точность базового поэлементного классификатора.

Таблица 1. Сравнительный анализ различных параметров пространственной классификации

Номер варианта	Метод кластеризации	Число кластеров	Число сегментов	Объединение результатов по сегменту	Оконная постобработка	Точность на тестовом фрагменте
1	–	–	–	–	–	0,7473
2	–	–	–	–	+	0,7607
3	GEM	9	2495	Голосование	–	0,8187
4	GEM	9	2495	Медиана	–	0,8158
5	GEM	9	2495	Голосование	+	0,8216
6	GEM	9	2495	Медиана	+	0,8187
8	kmeans++	5	1897	Голосование	+	0,7931
9	kmeans++	9	4300	Голосование	+	0,7914

Следует отметить, что точность поэлементного классификатора, рассчитанная по обучающей выборке методом K-fold при $K = 10$, составила 0,85, т. е. существенно больше, чем на тестовом фрагменте. Это обстоятельство свидетельствует об изменении спектральных характеристик одних и тех же классов растительности в различных фрагментах территории даже в пределах одного лесничества, что, в свою очередь, повышает практическую значимость результатов исследования, полученных на отдельном тестовом фрагменте.

Выявленные в *табл. 1* предпочтительные параметры процедуры пространственной классификации показали себя наилучшим образом и при других параметрах предобработки, а также при другом составе данных. Так, на стадии предобработки тестировался вариант окна 5×5 медианной фильтрации (наивысшая точность 0,8136), а также вариант без использования предварительной обработки (точность 0,8103). Увеличение точности при использовании

предварительной фильтрации, предположительно, объясняется наличием шумов в исходных снимках, которые могут быть вызваны погрешностями регистрации изображения, а также применением процедур атмосферной коррекции и повышения пространственного разрешения ряда каналов. Кроме того, нельзя исключать и фактор возможного наличия на таксационных выделах отдельных неучтённых участков малой площади, отличных по породному составу.

Использование набора из 136 главных компонент вместо 136 исходных каналов в данном эксперименте не является принципиальным. Сделано это для согласования полученных результатов с результатами эксперимента 4 по сокращению числа признаков, который будет описан ниже. Отметим, что в ходе наших исследований эксперимент 1 проводился также на наборе из исходных каналов. Это не повлияло на ранжирование различных вариантов процедуры классификации, но привело к небольшому снижению абсолютных показателей точности. Так, наилучший вариант 5 в этом случае показал точность, равную 0,8135. По причине эмпирически обнаруженного снижения точности при использовании исходных каналов, а также для единообразия в экспериментах 2–3 использовались главные компоненты.

Эксперимент 2 был посвящён тестированию алгоритма классификации при ином составе данных. Помимо L2A рассматривалась также возможность использования данных уровня L1C и признаков NDVI. Результаты эксперимента отражены в *табл. 2*. Они свидетельствуют о том, что атмосферная коррекция снимков Sentinel-2 может существенно повысить точность классификации, следовательно, ею не следует пренебрегать на практике. Что касается признаков NDVI, то они обладают высокой информативностью (что отражает последняя строка таблицы) и потому их использование уместно для сокращения размерности на стадии кластеризации, но дополнение ими полного набора каналов не представляется целесообразным.

Таблица 2. Влияние состава композита и атмосферной коррекции на точность классификации

Состав композита (использовались все снимки с 12 апреля по 19 октября)	Число признаков	Атмосферная коррекция данных	Точность на тестовом фрагменте
Band 2–8A	136	+	0,8216
Band 2–8A	136	–	0,7798
Band 2–8A + NDVI	153	+	0,8158
NDVI	17	+	0,7535

Таблица 3. Влияние сезона и пространственного разрешения каналов на точность классификации

Состав композита	Период съёмки	Число признаков	Точность на тестовом фрагменте
Все (Band 2–8A)	Весь сезон (12 апреля – 19 октября)	136	0,8216
Только 10 м (Band 2–4, 8)	Весь сезон (12 апреля – 19 октября)	68	0,8162
Только 20 м (Band 5–7, 8A)	Весь сезон (12 апреля – 19 октября)	68	0,8000
Все (Band 2–8A)	Весна (17 апреля – 27 мая)	48	0,8261
Все (Band 2–8A)	Весна + осень (17 апреля – 27 мая, 1 сентября – 19 октября)	80	0,7585
Все (Band 2–8A)	Осень (1 сентября – 19 октября)	32	0,7560
Все (Band 2–8A)	Лето (3 июня – 20 августа)	40	0,7329

Эксперимент 3 был посвящён исследованию влияния сезона съёмки и групп каналов равного пространственного разрешения на точность классификации. Его результаты, приведённые в *табл. 3*, подтвердили, что четыре канала с разрешением 10 м являются более зна-

чимыми, нежели четыре канала с разрешением 20 м, однако последние (в отличие от NDVI) хорошо дополняют каналы 10-метрового разрешения, обеспечивая более высокую точность классификации при использовании всех восьми каналов. Анализ данных по различным сезонам показал безусловное преимущество весенних снимков. Более того, изъятие из рассмотрения всех снимков, за исключением весенних, позволило повысить точность в сравнении с данными за весь сезон. Тем не менее при условии недостаточного количества безоблачных весенних снимков для конкретной территории использование всех имеющихся изображений за полный сезон представляется авторам разумной стратегией, что подтверждают данные в *табл. 3*.

Далее в эксперименте 4 мы провели анализ числа главных компонент, достаточных для достижения приемлемой точности классификации. Кумулятивная сумма собственных значений достигает 99 % от полной дисперсии уже при использовании трёх компонент и 99,9 % — при пятнадцати, однако точность классификации при такой длине вектора признаков оказалась невысокой. Поэтому были экспериментально получены значения точности классификации при разном числе главных компонент, использованных в качестве признаков. В качестве исходных данных полной размерности были взяты 136-канальный композит за весь сезон и хорошо показавший себя в предыдущем эксперименте 48-канальный композит с весенними снимками. Результирующие графики на *рис. 3*, за исключением небольших локальных максимумов, являются возрастающими, и глобальный максимум для обоих композитов достигается при использовании полного вектора признаков. Эмпирически можно прийти к выводу, что для снижения вычислительной сложности при обучении и классификации можно использовать 40 компонент в обоих случаях. При этом значении точность классификации по весеннему композиту составила 0,8259, по полному — 0,816.

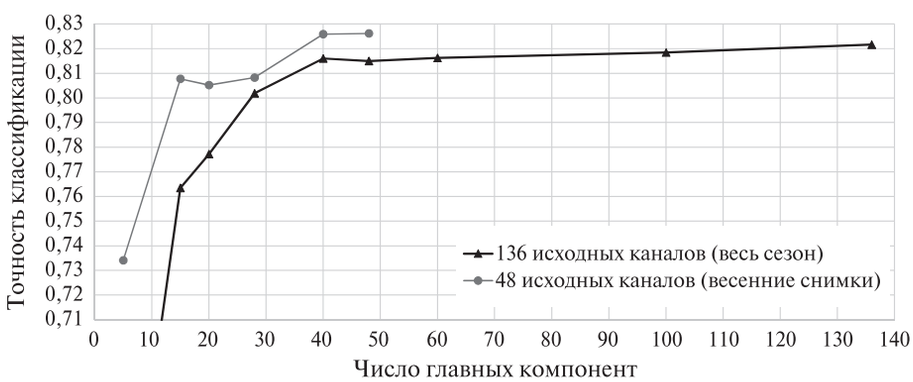


Рис. 3. Зависимость качества классификации от числа используемых главных компонент при различном составе исходных композитов

Для исследования информативности отдельных каналов и снимков в продолжение исследования, отражённого в *табл. 3*, был проведён отдельный эксперимент 5, в котором по полному композиту, состоящему из 136 исходных каналов (расчёт главных компонент на этот раз не проводился), отыскивались 60 наиболее информативных признаков. Для этого использовался метод последовательного добавления признаков, который не гарантирует оптимальность найденного решения. В качестве критерия при выборе очередного признака использовалась усреднённая ошибка классификации при применении процедуры 4-fold на выборке, включающей по 1000 элементов каждого класса. Пространственное уточнение результатов классификации в данном эксперименте не выполнялось. Число 60 было выбрано, поскольку при близких к нему значениях количества признаков достигался минимум ошибки классификации при различных запусках расчётной процедуры.

Результаты эксперимента 5 отражены в *табл. 4*. Числа от 1 до 60 в ячейках отражают номер, под которым был добавлен данный признак. На основе этих значений были рассчитаны показатели информативности отдельных снимков и каналов (последний столбец и последняя

строка соответственно), равные среднему значений в соответствующих строке или столбце (при этом значения в пустых ячейках принимались равными 61). Таким образом, меньшие значения показателя информативности в последних столбце и строке означают большую значимость соответствующего снимка или канала.

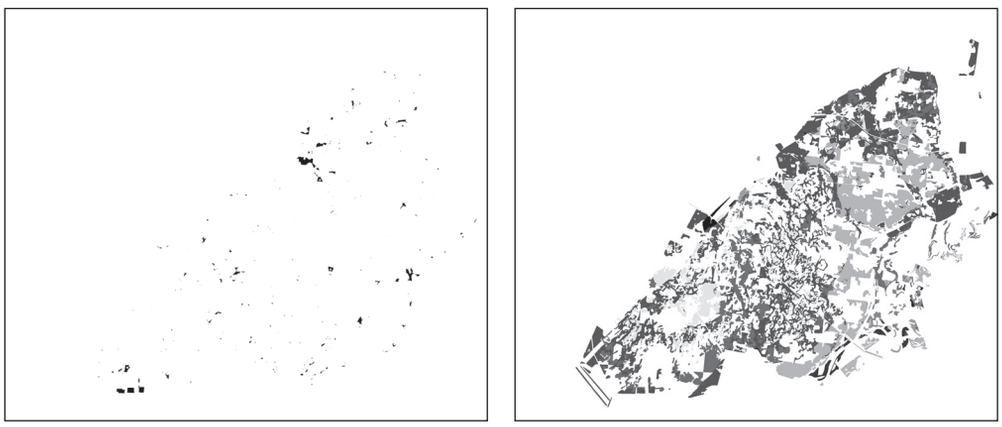
Таблица 4. Порядковая статистика информативности отдельных признаков 136-канального композита и суммарная информативность различных каналов и снимков

Дата	Канал								
	2 (Blue)	3 (Green)	4 (Red)	8 (NIR)	5 (VRE)	6 (VRE)	7 (VRE)	8A (NNIR)	Среднее по датам
12 апреля							7		54
17 апреля	42		48	20		44	60	14	44
2 мая	1	31					51	3	41
4 мая	6	24	23	52	47	8	45	15	28
12 мая	36	27	59	54				10	46
22 мая			18		11				49
27 мая		49	56		28			17	49
23 июня	57	35			5		30		46
3 июля	29	21			40	41		9	40
31 июля	26				34				53
5 августа	37		2						51
7 августа	55		43						58
20 августа	50								60
1 сентября	19	32	38				22	46	43
4 сентября	13		4	25	53				42
16 сентября	33	16	39					12	43
19 октября				58					61
Среднее по каналам	38	46	45	55	49	56	52	40	

По результатам эксперимента, наибольшую информативность с сильным преимуществом показали каналы Band 2 (Blue) и Band 8A (NNIR), далее следуют Band 4 (Red) и Band 3 (Green). Примечательно, что 10-метровый канал Band 8 (NIR) в силу широкой полосы спектра оказался в числе наименее информативных. Среди отдельных снимков наиболее значимый — снимок от 4 мая. Помимо него выделяются все снимки с 17 апреля до 3 июля, а также сентябрьские. Неинформативными ожидаемо оказались безлистные изображения от 12 апреля и 19 октября, а также снимки второй половины лета.

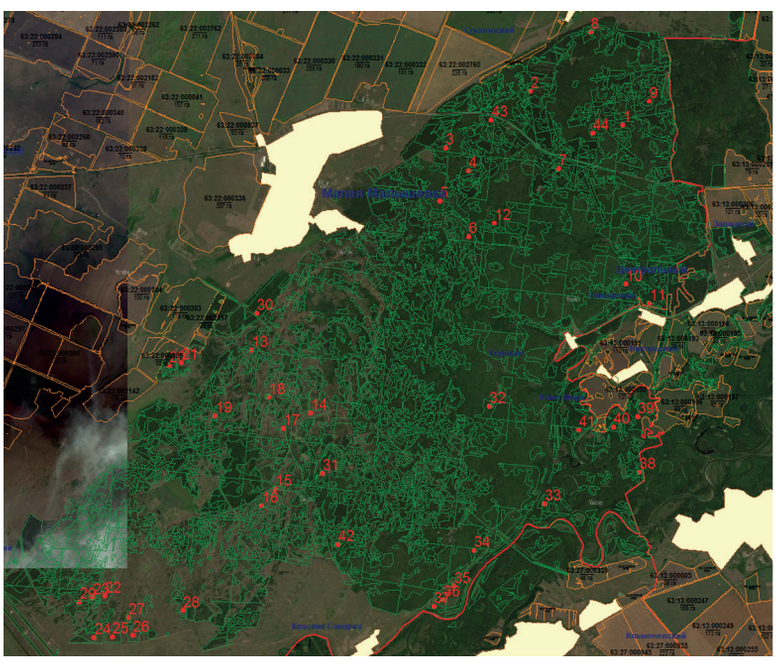
Практическое использование полученного классификатора для актуализации данных таксации по выбранному лесничеству

В результате анализа исходных данных и выбора параметров алгоритма пространственной классификации была получена модель классификатора, которая далее использовалась для решения задачи уточнения (актуализации) таксации леса, которая, как отмечалось выше, была составлена в 2013–2014 гг. и могла устареть. Для этого выбранный классификатор был обучен по выборке с тем же числом экземпляров на каждый класс (2000), но включающей примеры с территории всего лесничества. Далее была классифицирована полная территория лесничества (точность составила 0,8535) и построена маска ошибок классификации (*рис. 4а*, см. с. 95).

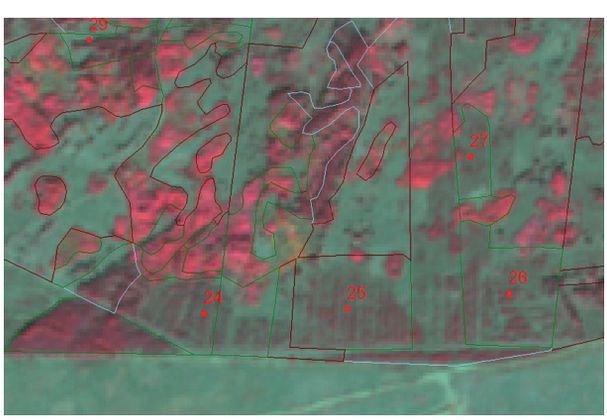


а *б*

Рис. 4. Ошибки первичной классификации Красносамарского лесничества (*а*); итоговый результат классификации после уточнения 8 участков (*б*)



а



б

Рис. 5. Центры 45 ошибочно классифицированных участков лесничества на карте (*а*); увеличенный фрагмент, содержащий выделы молодого сосняка, которые мало отличаются от степной растительности (*б*)

Затем на полученной маске при помощи ранговой фильтрации были выделены 45 относительно крупных локализованных участка, которые вслед за этим нанесли на карту (*рис. 5а*, см. с. 95). В октябре 2018 г. были проведены маршрутные полевые обследования 13 из них, в ходе которых оценивался актуальный статус экосистем и для лесных насаждений — характер древостоя. Обследования показали, что для 9 участков в действительности данные таксации не являются достоверными. Причины этого самые разные, а именно:

1) сукцессионные изменения в выделах в результате замещения ранее доминировавших видов другими древесными породами;
2) выгорание древесного яруса в результате пожаров в последние годы, причём для части наиболее старых горельников началось зарастание преимущественно осиной и берёзой;
3) неточности в исходных данных, связанные с возрастом лесокультур (молодые посадки сосны не отличаются на снимках от степной растительности (*рис. 5б*));
4) неточности исходных данных в определении границ между выделами.

Из четырёх оставшихся участков в двух случаях ошибки объясняются низкой сомкнутостью крон (0,5 и ниже). К сожалению, данные, использованные для обучения, не включали этот важный параметр.

Таблица 5. Матрица ошибок при классификации всей территории лесничества

		Ответ классификатора									Сумма	Ошибка первого рода, %
		Берёза	Вяз	Дуб	Ива	Клён	Осина	Сосна	Тополь	Ясень		
Данные таксации	Берёза	58423	0	858	0	88	4810	12 006	0	783	76 968	24,1
	Вяз	242	384	0	0	0	300	313	220	0	1459	73,7
	Дуб	283	0	148 242	13	1225	7662	4650	67	0	162 142	8,6
	Ива	16	0	17	2727	212	57	682	808	0	4519	39,7
	Клён	137	0	2295	62	5340	696	2405	0	0	10 935	51,2
	Осина	4036	0	11 910	329	1075	112 376	14 345	9	0	144 080	22,0
	Сосна	13 055	0	2429	59	1381	19 902	315 238	610	216	352 890	10,7
	Тополь	100	1	722	172	204	102	171	6720	0	8192	18,0
	Ясень	1	0	116	0	107	39	162	0	3825	4250	10,0
Сумма		76 293	385	166 589	3362	9632	145 944	349 972	8434	4824		
Ошибка второго рода, %		23,4	0,3	11,0	18,9	44,6	23,0	9,9	20,3	20,7		

Для найденных участков была изменена семантическая информация, после чего получена новая маска истинной разметки классов (*рис. 4б*). Точность классификации на новой маске немного повысилась (до 0,8551). В *табл. 5* приведена итоговая матрица ошибок, а также отражены доли ошибок первого и второго рода для каждой породы леса в общей площади размеченной части лесничества. Оставшиеся участки будут проверены в новом полевом сезоне, но уже полученные результаты позволяют сделать вывод о потенциальной работоспособности выработанной методики уточнения данных таксации на основе классификации композитов данных Sentinel-2.

Заключение

Результаты проведённого исследования показали, что для классификации породного состава лесных территорий по сезонным композитам снимков Sentinel-2 и данным таксации могут эффективно применяться приёмы управляемой пространственной классификации, разрабо-

танные для гиперспектральных данных ДЗЗ. В ходе экспериментальных исследований был выбран оптимальный по показателю точности вариант реализации пространственной обработки результатов классификации, исследованы различные варианты формирования композитов снимков Sentinel-2, используемых для классификации, а также произведён анализ ошибок классификации с использованием наземных обследований.

По результатам исследований различных вариантов реализации пространственной обработки и сезонов съёмки, используемых для составления композита данных Sentinel-2, было получено, что пространственная обработка при классификации позволяет повысить её точность относительно поэлементной на 2–10 % и даже выше. Наивысший экспериментально полученный показатель точности классификации с использованием пространственной обработки составил 0,8261 и был получен для набора из 48 весенних снимков с атмосферной коррекцией.

В процессе изучения было выявлено, что в рассмотренной технологии индексы NDVI имеет смысл использовать только на этапе пространственной обработки для кластеризации. Кроме того, исследования показали, что дополнение 10-метровых каналов 20-метровыми вполне целесообразно, поскольку повышает точность классификации. В результате дальнейшего анализа информативности спектральных каналов Sentinel-2 было получено, что самыми информативными являются каналы B, NNIR (20 м), R и G.

В ходе исследования ошибок классификации с привлечением результатов наземных обследований было подтверждено предположение, что ошибки классификации при обучении по данным таксации нередко вызваны изменениями состояния участков лесных экосистем вследствие сукцессионных изменений и пожаров, произошедшими после формирования данных таксации, а также неточностями данных о возрасте древостоев. Таким образом, исследования показали, что рассмотренная процедура классификации может быть применена для уточнения таксационных данных с использованием сезонных композитов снимков Sentinel-2.

Исследование выполнено при поддержке Российского фонда фундаментальных исследований (проекты № 16-29-09494 офи_м, 18-07-00748 а, 19-07-00357 а) и Министерства науки и высшего образования РФ (соглашение № 007-ГЗ/Ч3363/26).

Литература

1. *Барталев С. А.* Разработка методов оценки состояния и динамики лесов на основе данных спутниковых наблюдений: автореф. дис. ... д-ра техн. наук. М.: ИКИ РАН, 2007. 48 с.
2. *Барталев С. А.*, *Жирин В. М.*, *Ершов Д. В.* Сравнительный анализ данных спутниковых систем «Космос-1939», SPOT и Landsat-TM при изучении бореальных лесов // Исследование Земли из космоса. 1995. Т. 1. С. 101–114.
3. *Борзов С. М.*, *Потатуркин О. И.* Классификация типов растительного покрова по гиперспектральным данным дистанционного зондирования Земли // Вестн. Новосибирского гос. ун-та. Сер.: Информационные технологии. 2014. Т. 12. № 4. С. 13–22.
4. *Галеева Л. П.* Мониторинг лесных земель: учебно-метод. пособие. Новосибирск: Изд-во НГАУ «Золотой колос», 2016. 147 с.
5. Госдоклад «О состоянии и об охране окружающей среды Российской Федерации в 2016 году». Москва: Минприроды России; НИА-Природа, 2017. 760 с.
6. *Жарко В. О.*, *Барталев С. А.* Оценка распознаваемости древесных пород леса на основе спутниковых данных о сезонных изменениях их спектрально-отражательных характеристик // Современные проблемы дистанционного зондирования Земли из космоса. 2014. Т. 11. № 3. С. 159–170.
7. *Зимичев Е. А.*, *Казанский Н. Л.*, *Серафимович П. Г.* Пространственная классификация гиперспектральных изображений с использованием метода кластеризации k-means++ // Компьютерная оптика. 2014. Т. 38. № 2. С. 281–286.
8. *Кедров А. В.*, *Тарасов А. В.* Классификация лесной растительности методом нейронных сетей // Вестн. ПНИПУ. 2017. Т. 22. С. 44–54.
9. *Кузнецов А. В.*, *Мясников В. В.* Сравнение алгоритмов управляемой поэлементной классификации гиперспектральных изображений // Компьютерная оптика. 2014. Т. 38. № 3. С. 494–502.

10. *Курбанов Э. А.*, *Воробьев О. Н.*, *Незамаев С. А.*, *Губаев А. В.*, *Лежнин С. А.*, *Полевщикова Ю. А.* Тематическое картирование и стратификация лесов марийского Заволжья по спутниковым снимкам Landsat // Вестн. Поволжского гос. технолог. ун-та. 2013. Т. 3. № 19. С. 82–92.
11. *Любимов А. В.*, *Селиванов А. А.*, *Крючков А. Н.*, *Номалунгело К. Н.*, *Тхин Ч. Х.*, *Саксонов С. В.* Анализ признаков дешифрования таксационных показателей лесов с использованием вероятностных методов // Изв. Самарского науч. центра Российской акад. наук. 2018. Т. 20. № 3. С. 85–90.
12. *Мальков Ю. Г.*, *Закамский В. А.* Мониторинг лесных экосистем: учеб. пособие. Йошкар-Ола: Марийский гос. техн. ун-т, 2006. 212 с.
13. *Перепечина Ю. И.*, *Глушенков О. И.*, *Корсиков Р. С.* Учет и оценка лесов, возникших на сельскохозяйственных землях с использованием данных дистанционного зондирования Земли // Лесной журн. 2016. № 4. С. 71–80.
14. Положение о лесном мониторинге. Федеральная служба лесного хозяйства России. Письмо от 29.11.95 года № МГ-1-17-6/287.
15. *Таранков В. И.* Мониторинг лесных экосистем. СПб.: Лань, 2006. 299 с.
16. *Троц В. Б.*, *Валиуллина А. Т.*, *Моисеева И. С.* Таксация леса: Методические указания для выполнения курсовой работы по теме «Таксация и материально-денежная оценка делянки». Кинель: РИЦ СГСХА, 2015. 392 с.
17. *Arthur D.*, *Vassilvitskii S.* k-means++: The Advantages of Careful Seeding // Proc. 18th Annual ACM-SIAM Symp. Discrete Algorithms. 2007. P. 1027–1035.
18. *Bavrina A.*, *Denisova A.*, *Kavelenova L.*, *Korchikov E.*, *Kuzovenko O.*, *Prokhorova N.*, *Terentyeva D.*, *Fedoseev V.* Some problems of regional reference plots system for ground support of remote sensing materials processing // Information Technologies in the Research of Biodiversity: Springer Proc. Earth and Environmental Sciences. 2018. P. 131–143.
19. *Corbane C.*, *Lang S.*, *Pipkins K.*, *Alleaume S.*, *Deshayes M.*, *García Millán V. E.*, *Michael F.* Remote sensing for mapping natural habitats and their conservation status – New opportunities and challenges // Intern. J. Applied Earth Observation and Geoinformation. 2015. V. 37. P. 7–16.
20. *Cristianini N.*, *Shawe-Taylor J.* An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000. 216 p.
21. *Dalponte M.*, *Ørka H. O.*, *Ene L. T.*, *Gobakken T.*, *Næsset E.* Tree crown delineation and tree species classification in boreal forests using hyperspectral and ALS data // Remote Sensing of Environment. 2014. V. 140. P. 306–317.
22. *Denisova A. Y.*, *Sergeyev V. V.* EM clustering algorithm modification using multivariate hierarchical histogram in the case of undefined cluster number // Proc. SPIE. 2018. V. 10806. P. 108064H.
23. *Drusch M.*, *Del Bello U.*, *Carlier S.*, *Colin O.*, *Fernandez V.*, *Gascon F.*, *Bargellini P.* Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services // Remote Sensing of Environment. 2012. V. 120. P. 25–36.
24. *Fedoseev V.* Hyperspectral satellite image classification using small training data from its samples // J. Physics: Conf. Series. 2018. V. 1096. P. 012042.
25. *Immitzer M.*, *Vuolo F.*, *Atzberger C.* First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe // Remote Sensing. 2016. V. 8. No. 3. P. 166.
26. *Karlson M.*, *Ostwald M.*, *Reese H.*, *Sanou J.*, *Tankoano B.*, *Mattsson E.* Mapping tree canopy cover and aboveground biomass in sudano-sahelian woodlands using Landsat 8 and Random Forest // Remote Sensing. 2015. V. 7. No. 8. P. 10017–10041.
27. *Puliti S.*, *Ørka H.*, *Gobakken T.*, *Næsset E.*, Inventory of Small Forest Areas Using an Unmanned Aerial System // Remote Sensing. 2015. V. 7. No. 8. P. 9632–9654.
28. *Waring R. H.*, *Coops N. C.*, *Fan W.*, *Nightingale J. M.* MODIS enhanced vegetation index predicts tree species richness across forested ecoregions in the contiguous U.S.A. // Remote Sensing of Environment. 2006. V. 103. No. 2. P. 218–226.
29. *Xiao X.*, *Boles S.*, *Liu J.*, *Zhuang D.*, *Liu M.* Characterization of forest types in Northeastern China, using multi-temporal SPOT-4 vegetation sensor data // Remote Sensing of Environment. 2002. V. 82. No. 2–3. P. 335–348.
30. *Zhang T.*, *Su J.*, *Liu C.*, *Chen W.-H.*, *Liu H.*, *Liu G.* Band selection in Sentinel-2 satellite for agriculture applications // 23rd Intern. Conf. Automation and Computing (ICAC). Huddersfield, United Kingdom: IEEE, 2017. P. 1–6.

Tree species classification in Samara Region using Sentinel-2 remote sensing images and forest inventory data

A. Yu. Denisova[1], L. M. Kavelenova[1], E. S. Korchikov[1], N. V. Prokhorova[1], D. A. Terentyeva[1], V. A. Fedoseev[1,2]

[1] *Samara National Research University, Samara 443086, Russia*
[2] *Image Processing Systems Institute, Branch of Federal Scientific Research Centre "Crystallography and Photonics" RAS, Samara 443001, Russia*
E-mail: vicanfed@gmail.ru

Tree species information is required in many ecological management applications, for example, for conservation status assignment or human activity management for particular forest sites. However, governmental forest inventory is very expensive. Therefore, the inventory is very slowly updated, approximately once a decade. That is why more operative and independent data sources, such as remote sensing systems, are required to refine latest forest inventory data. This paper presents an investigation on tree species classification using seasonal Sentinel-2 data and the latest forest inventory data. The advantage of Sentinel-2 satellites for solving this problem lies in short revisiting time for the territory and large field of view that is important for large territory analysis. The forest inventory data were used for training and classification with further ground survey of misclassified regions. The classification was organized as a comprehensive supervised spectral-spatial classification procedure based on combination of different spectral and spatial image processing algorithms. The paper addresses the issues of optimal image data selection and processing, classification procedure configuration and training method. The studies were carried out using the territory of the Krasnosamarskoe forestry in Samara Region. For training and verification of classifiers, latest available forest inventory data (2013–2014) and seasonal Sentinel-2 data (2018) were used. The experiments showed that proper image data selection and classification procedure configuration result in high classification accuracy (about 0.82) for the control fragment of the territory. Moreover, the performed ground survey partially confirmed that classification errors are related to the changes in tree species concentration and age that were not taken into account in forestry inventory data. Thus, Sentinel-2 data are practically valuable for forest inventory data refinement.

Keywords: tree species classification, Sentinel-2, spatial classification, forest inventory data, EM clustering, SVM

Accepted: 17.05.2019
DOI: 10.21046/2070-7401-2019-16-4-86-101

References

1. Bartalev S. A., *Razrabotka metodov otsenki sostoyaniya i dinamiki lesov na osnove dannykh sputnikovykh nablyudenii: Avtoref. dis. dokt. tekhn. nauk* (Developing of methods for forest state assessment and forest dynamics based on satellite observation data. Ext. abstract Doct. tech. sci. thesis), Moscow: IKI RAN, 2007, 48 p.
2. Bartalev S. A., Zhirin V. M., Ershov D. V., Sravnitel'nyi analiz dannykh sputnikovykh sistem "Kosmos-1939", SPOT i "Landsat-TM" pri izuchenii boreal'nykh lesov (The comparative analysis of "Kosmos-1939", SPOT and Landsat-TM satellite systems data for boreal forests study), *Issledovanie Zemli iz kosmosa*, 1995, Vol. 1, pp. 101–114.
3. Borzov S. M., Potaturkin O. I., Klassifikatsiya tipov rastitel'nogo pokrova po giperspektral'nym dannym distantsionnogo zondirovaniya Zemli (Vegetative cover type classification using hyperspectral remote sensing), *Vestnik NGU. Seriya: Informatsionnye tekhnologii*, 2014, Vol. 12, No. 4, pp. 13–22.
4. Galeeva L. P., *Monitoring lesnykh zemel'* (Monitoring of forest lands), Novosibirsk: Izd. NGAU "Zolotoi kolos", 2016, 147 p.
5. *State report* "*On the state and protection of the environment of the Russian Federation in 2016*", Moscow: Minprirody of Russia, NIA-Priroda, 2017, 760 p. (In Russ).
6. Zharko V. O., Bartalev S. A., Otsenka raspoznavaemosti drevesnykh porod lesa na osnove sputnikovykh dannykh o sezonnykh izmeneniyakh ikh spektral'no-otrazhatel'nykh kharakteristik (Forest tree species recognizability assessment based on satellite data on their spectral reflectance seasonal changes), *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2014, Vol. 11, No. 3, pp. 159–170.

7. Zimichev E. A., Kazanskiy N. L., Serafimovich P. G., Prostranstvennaya klassifikatsiya giperspektral'nykh izobrazhenii s ispol'zovaniem metoda klasterizatsii k-means++ (Spectral-spatial classification with k-means++ particional clustering), *Computer optics*, 2014, Vol. 38, No. 2, pp. 281–286.
8. Kedrov A. V., Tarasov A. V., Klassifikatsiya lesnoi rastitel'nosti metodom neironnykh setei (Classification forest vegetation with neural network), *Vestnik PNIPU*, 2017, Vol. 22, pp. 44–54.
9. Kuznetsov A. V., Myasnikov V. V., Sravnenie algoritmov upravlyaemoi poelementnoi klassifikatsii giperspektral'nykh izobrazhenii (A comparison of algorithms for supervised classification using hyperspectral data), *Computer optics*, 2014, Vol. 38, No. 3, pp. 494–502.
10. Kurbanov E. A., Vorob'ev O. N., Nezamaev S. A., Gubaev A. V., Lezhnin S. A., Polevshchikova Yu. A., Tematicheskoe kartirovanie i stratifikatsiya lesov mariiskogo Zavolzh'ya po sputnikovym snimkam Landsat (Thematic mapping and stratification of forests in middle zavolsgie by landsat satelite images), *Vestnik Povolzhskogo gosudarstvennogo tekhnologicheckogo universiteta*, 2013, Vol. 3, No. 19, pp. 82–92.
11. Lyubimov A. V., Selivanov A. A., Kryuchkov A. N., Nomalungelo K. N., Tkhin Ch. Kh., Saksonov S. V., Analiz priznakov deshifrovaniya taksatsionnykh pokazatelei lesov s ispol'zovaniem veroyatnostnykh metodov (Analyzes of the photointerpretation parameters for the forest inventory on the probability methods), *Izvestiya Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk*, 2018, Vol. 20, No. 3, pp. 85–90.
12. Mal'kov Yu. G., Zakamskii V. A., *Monitoring lesnykh ekosistem* (Monitoring of forest ecosystems), Yoshkar-Ola: Mari State Technical University, 2006, 212 p.
13. Perepechina Yu. I., Glushenkov O. I., Korsikov R. S., Uchet i otsenka lesov, voznikshikh na sel'skokhozyaistvennykh zemlyakh s ispol'zovaniem dannykh distantsionnogo zondirovaniya Zemli (Forest Inventory and Assessment in the Agricultural Lands Using the Earth Remote Sensing Data), *Lesnoi zhurnal*, 2016, No. 4, pp. 71–80.
14. Federal Forestry Agency of the Russian Federation: Letter No. MG-1-17-6/287, Date 29.11.1995.
15. Tarankov V. I., *Monitoring lesnykh ekosistem* (Monitoring of forest ecosystems), Saint Petersburg: Lan', 2006, 299 p.
16. Trots V. B., Valiullina A. T., Moiseeva I. S., *Taksatsiya lesa* (Forest inventories), Kinel': RITs SGSKhA, 2015, 392 p.
17. Arthur D., Vassilvitskii S., k-means++: The Advantages of Careful Seeding, *Proc. 18th Annual ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035.
18. Bavrina A., Denisova A., Kavelenova L., Korchikov E., Kuzovenko O., Prokhorova N., Terentyeva D., Fedoseev V., Some problems of regional reference plots system for ground support of remote sensing materials processing, *Information Technologies in the Research of Biodiversity: Springer Proc. Earth and Environmental Sciences*, 2018, pp. 131–143.
19. Corbane C., Lang S., Pipkins K., Alleaume S., Deshayes M., García Millán V. E., Michael F., Remote sensing for mapping natural habitats and their conservation status — New opportunities and challenges, *Intern. J. Applied Earth Observation and Geoinformation*, 2015, Vol. 37, pp. 7–16.
20. Cristianini N., Shawe-Taylor J., *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000, 216 p.
21. Dalponte M., Ørka H. O., Ene L. T., Gobakken T., Næsset E., Tree crown delineation and tree species classification in boreal forests using hyperspectral and ALS data, *Remote Sensing of Environment*, 2014, Vol. 140, pp. 306–317.
22. Denisova A. Y., Sergeyev V. V., EM clustering algorithm modification using multivariate hierarchical histogram in the case of undefined cluster number, *Proc. SPIE*, 2018, Vol. 10806, p. 108064H.
23. Drusch M., Del Bello U., Carlier S., Colin O., Fernandez V., Gascon F., Bargellini P., Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services, *Remote Sensing of Environment*, 2012, Vol. 120, p. 25–36.
24. Fedoseev V., Hyperspectral satellite image classification using small training data from its samples, *J. Physics: Conf. Series*, 2018, Vol. 1096, p. 012042.
25. Immitzer M., Vuolo F., Atzberger C., First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe, *Remote Sensing*, 2016, Vol. 8, No. 3, p. 166.
26. Karlson M., Ostwald M., Reese H., Sanou J., Tankoano B., Mattsson E., Mapping tree canopy cover and aboveground biomass in sudano-sahelian woodlands using Landsat 8 and Random Forest, *Remote Sensing*, 2015, Vol. 7, No. 8, pp. 10017–10041.
27. Puliti S., Ørka H., Gobakken T., Næsset E., Inventory of Small Forest Areas Using an Unmanned Aerial System, *Remote Sensing*, 2015, Vol. 7, No. 8, pp. 9632–9654.
28. Waring R. H., Coops N. C., Fan W., Nightingale J. M., MODIS enhanced vegetation index predicts tree species richness across forested ecoregions in the contiguous U. S.A., *Remote Sensing of Environment*, 2006, Vol. 103, No. 2, pp. 218–226.

29. Xiao X., Boles S., Liu J., Zhuang D., Liu M., Characterization of forest types in Northeastern China, using multi-temporal SPOT-4 vegetation sensor data, *Remote Sensing of Environment*, 2002, Vol. 82, No. 2–3, pp. 335–348.
30. Zhang T., Su J., Liu C., Chen W.-H., Liu H., Liu G., Band selection in Sentinel-2 satellite for agriculture applications, *23rd Intern. Conf. Automation and Computing* (*ICAC*), Huddersfield, United Kingdom: IEEE, 2017, pp. 1–6.