Использование методов машинного обучения для оценки характеристик лесов России с помощью данных ДЗЗ

Хвостиков С.А., Барталев С.А.
khvostikov@d902.iki.rssi.ru

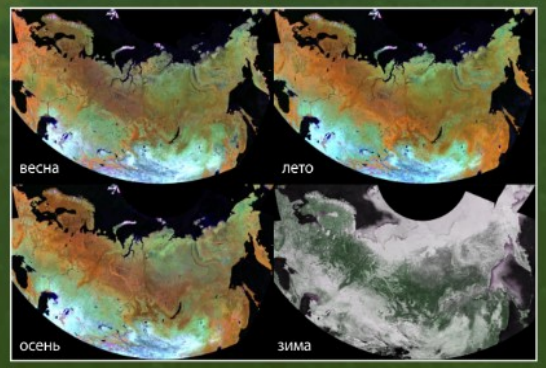
Институт Космических Исследований РАН

Представленные в докладе продукты разработаны коллективом авторов лаборатории спутникового мониторинга наземных экосистем:

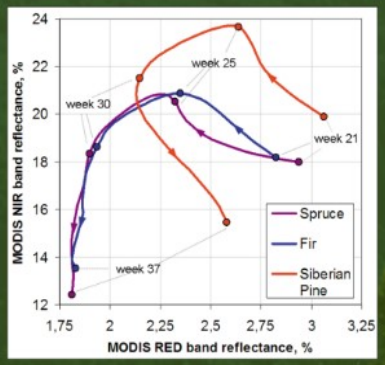
Барталев С.А., Ворушилов И.И., Егоров В.А., Жарко В.О., Михайлов Н.В., Сайгин И.А., Стыценко Ф.В., Хвостиков С.А., Ховратович Т.С.

Двадцать третья международная конференция
«Современные проблемы дистанционного зондирования Земли из космоса»
ИКИ РАН 10 – 14 ноября 2025 г.

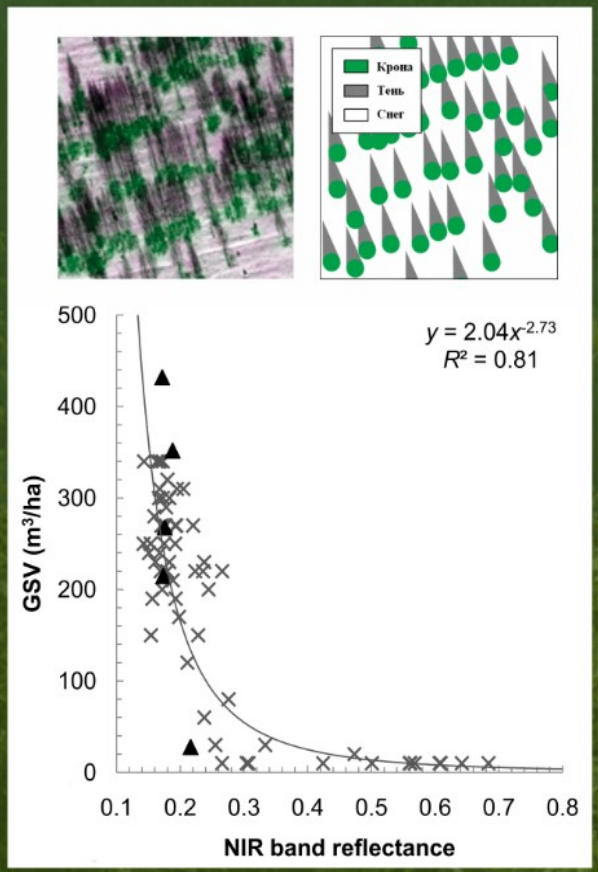
МЕТОДЫ СПУТНИКОВОГО КАРТОГРАФИРОВАНИЯ ЛЕСОВ



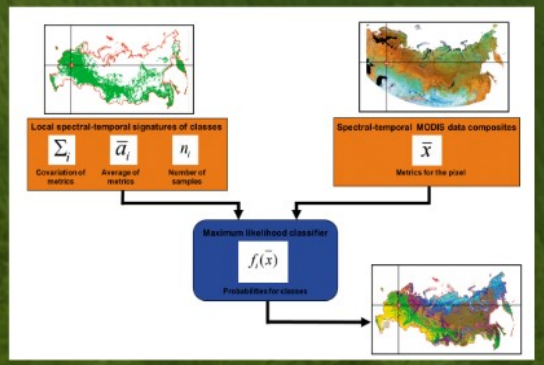
многолетние временные ряды
очищенных от влияния облаков
разносезонных композитных
изображений



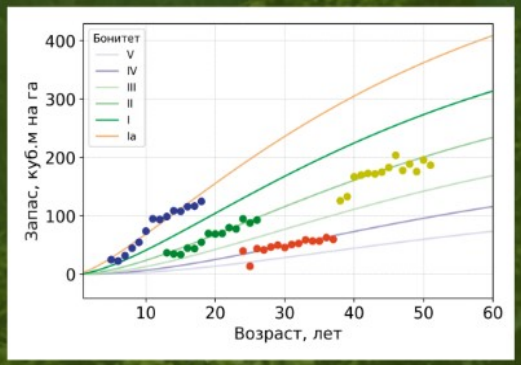
распознавание
древесных пород лесов
по их спектрально-временным
признакам



определение запаса лесов
по данным съемки в зимнее время
при наличии снежного покрова
на земной поверхности



алгоритмы локально-
адаптивной классификации
и оценивания характеристик
земного покрова

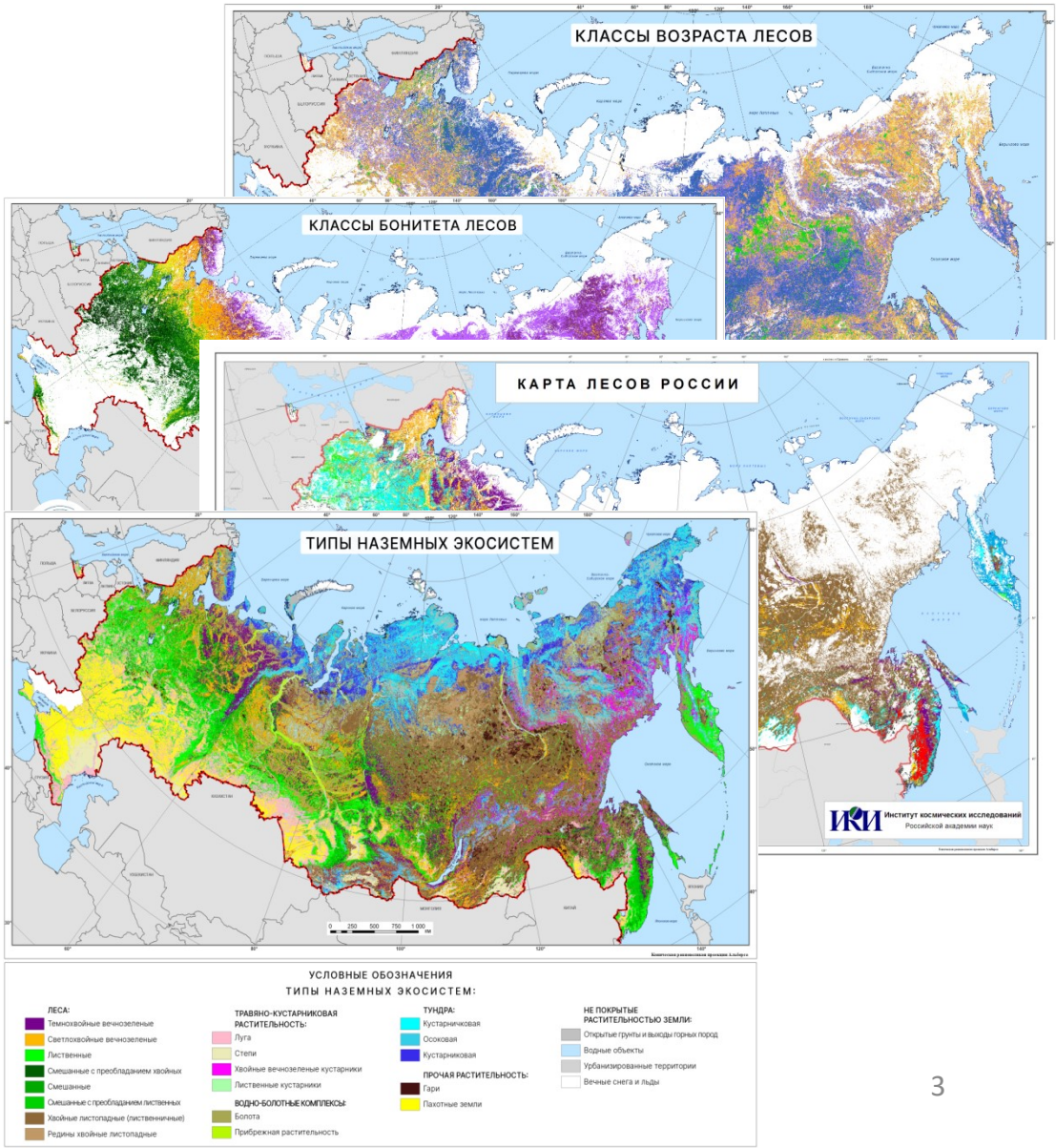


определение бонитета и возраста лесов
с на основе моделей их динамики
и ежегодных измерений запаса
по данным ДЗЗ

Этапы построения карт на основе машинного обучения

Основные этапы:

1. Постановка задачи – определить область, классы
2. Выбрать индикаторы-признаки
3. Собрать выборку для обучения/проверки точности
4. Обучить метод машинного обучения и построить карту
5. Оценить точность



УСЛОВНЫЕ ОБОЗНАЧЕНИЯ

ТИПЫ НАЗЕМНЫХ ЭКОСИСТЕМ:

ЛЕСА:
- Темнохвойные вечнозеленые
- Светлохвойные вечнозеленые
- Лиственные
- Смешанные с преобладанием хвойных
- Смешанные
- Смешанные с преобладанием лиственных
- Хвойные листопадные (лиственничные)
- Редины хвойные листопадные

ТРАВЯНО-КУСТАРНИКОВАЯ РАСТИТЕЛЬНОСТЬ:
- Луга
- Степи
- Хвойные вечнозеленые кустарники
- Лиственные кустарники

ВОДНО-БОЛОТНЫЕ КОМПЛЕКСЫ:
- Болота
- Прибрежная растительность

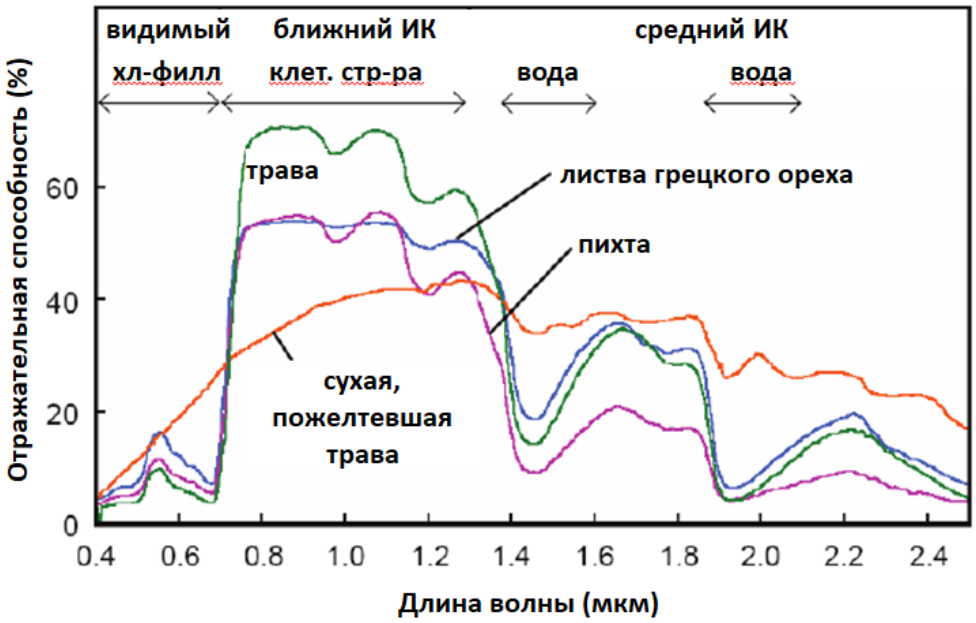
ТУНДРА:
- Кустарничковая
- Осоковая
- Кустарниковая

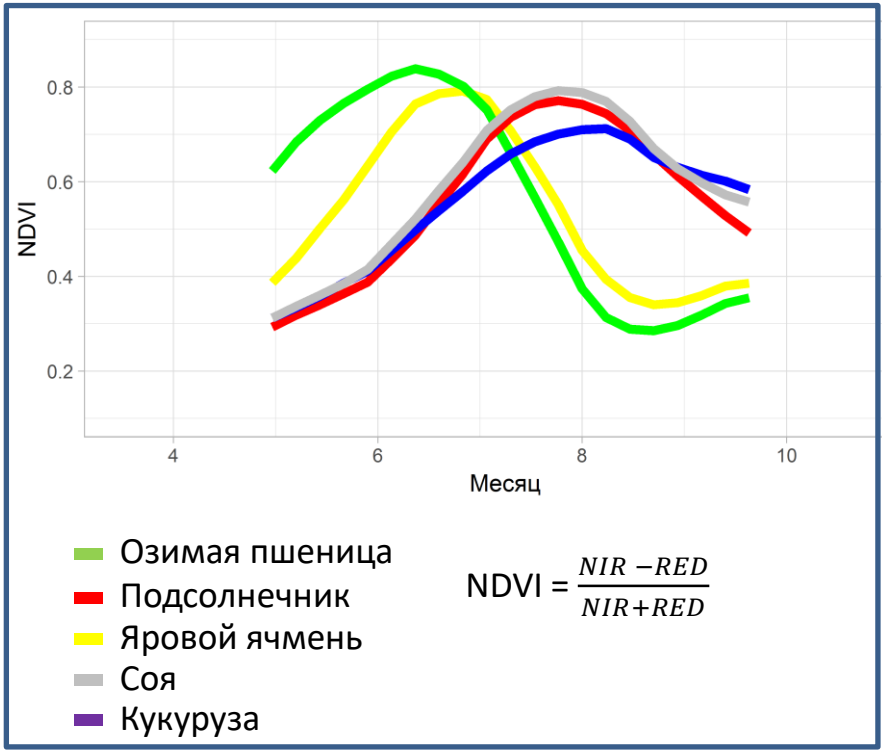
ПРОЧАЯ РАСТИТЕЛЬНОСТЬ:
- Гари
- Пахотные земли

НЕ ПОКРЫТЫЕ РАСТИТЕЛЬНОСТЬЮ ЗЕМЛИ:
- Открытые грунты и выходы горных пород
- Водные объекты
- Урбанизированные территории
- Вечные снега и льды

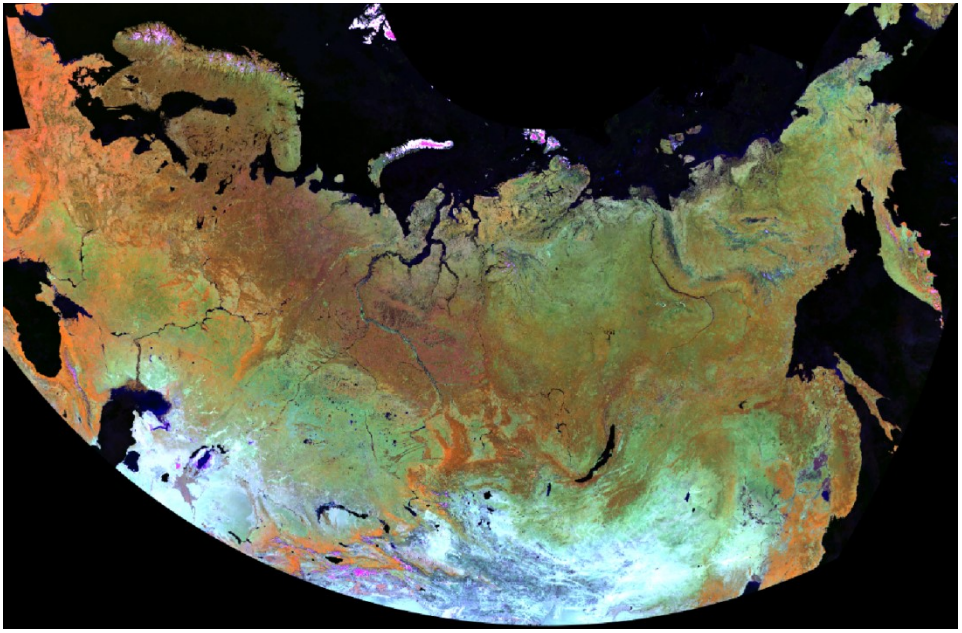
Выбор признаков

Для классификации необходимо выбрать набор признаков, значения которых отличаются для разных классов растительности и позволяют однозначно идентифицировать их.

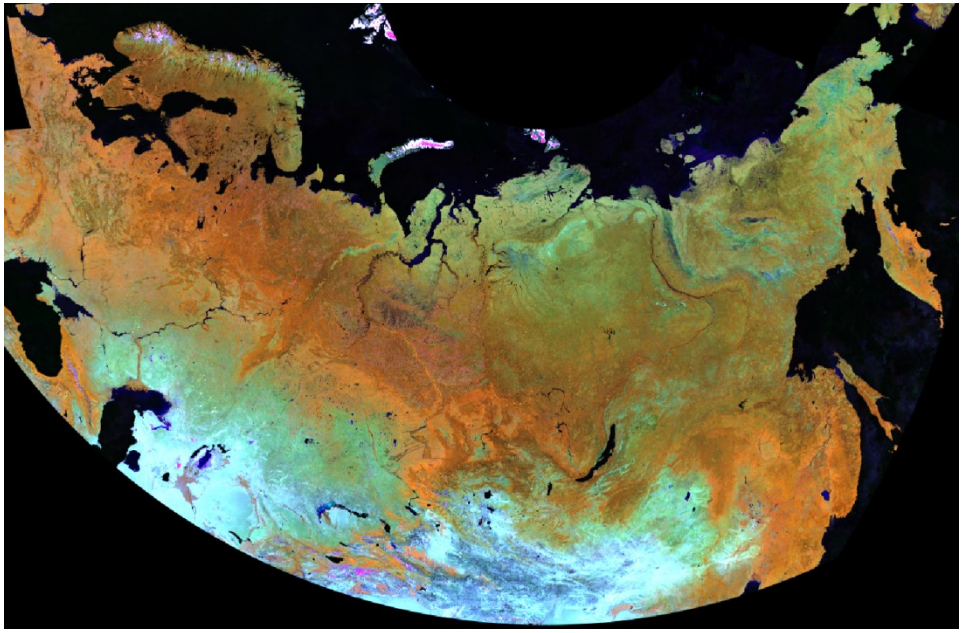




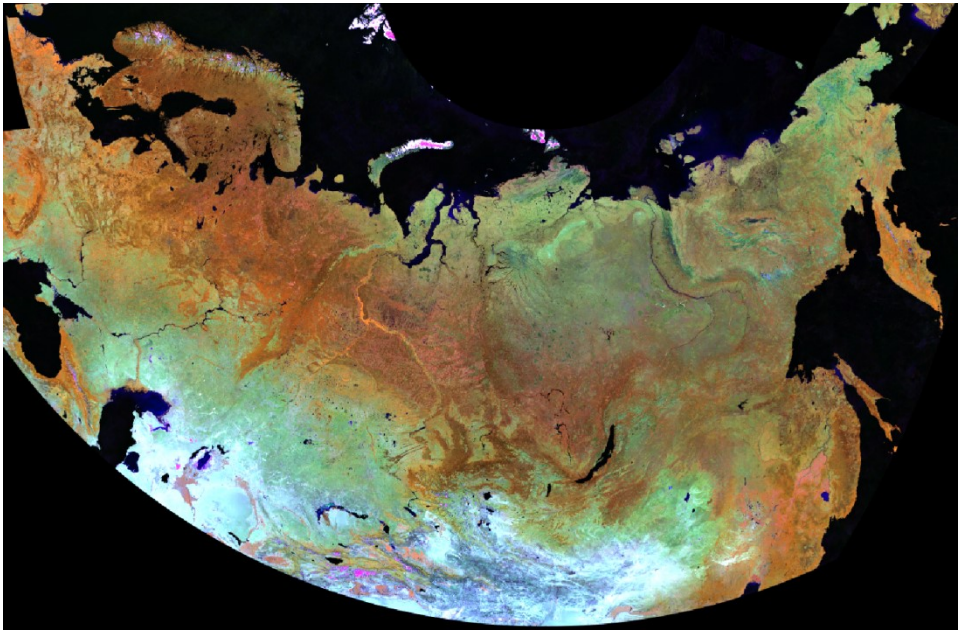
Сезонные композитные изображения MODIS



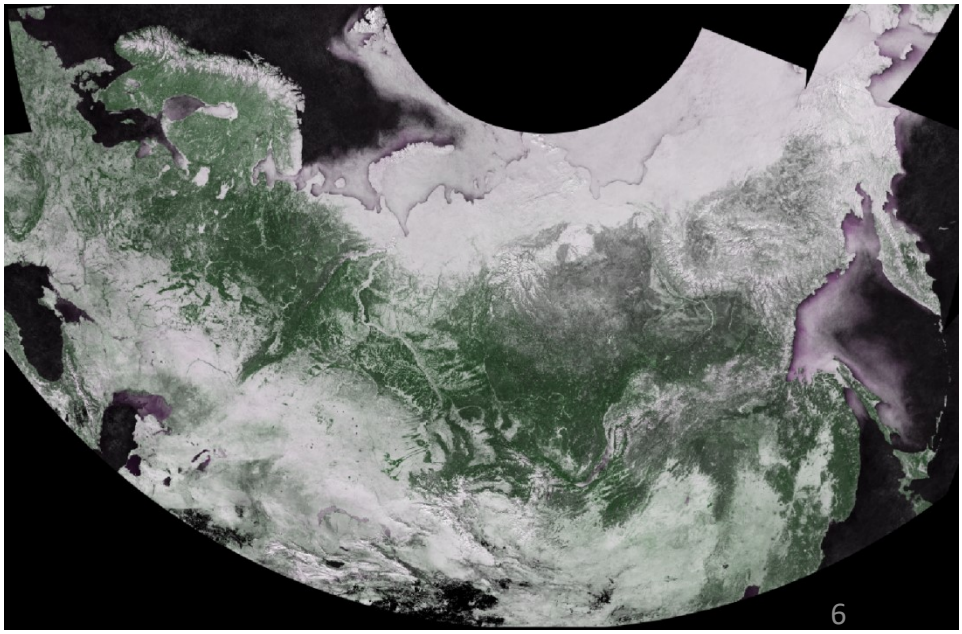
весна



лето



осень



зима

6

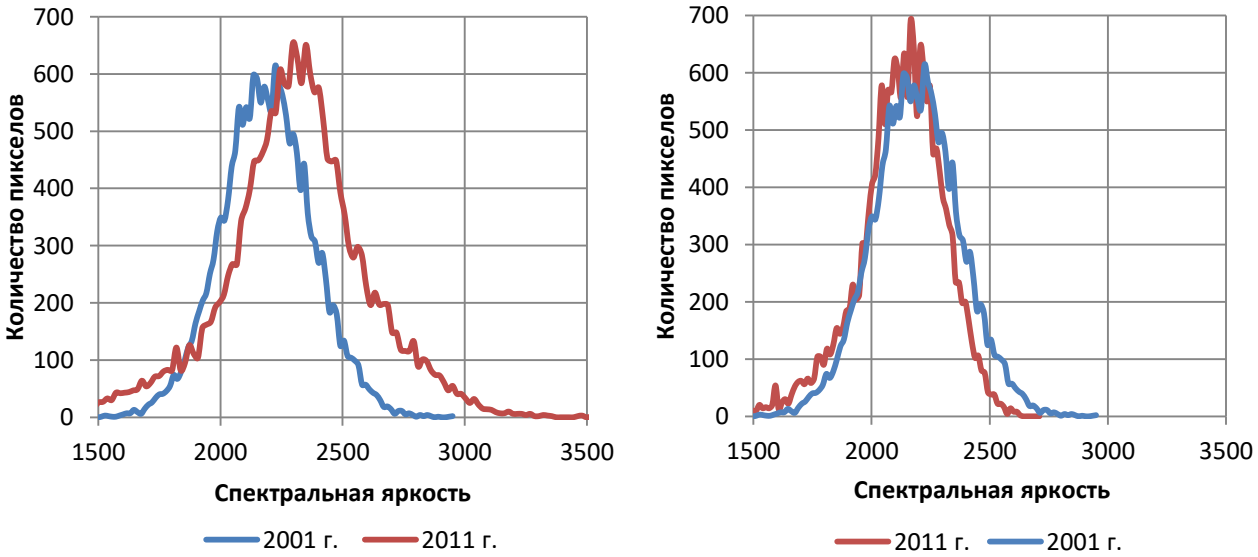
Обучающая выборка

Обучающая выборка – набор точек/пикселов для которых известен их класс или значение целевой характеристики. Ее качество определяет качество итоговой карты.

Выборку можно получить с помощью наземных исследований или на основе экспертной оценки спутниковых данных.

Исходная выборка карты растительности строилась на основе других карт landcover и вспомогательной справочной информации. Но использование одной выборки для классификации нескольких лет приводит к проблеме, потребовавших разработки метода переноса выборки между годами.

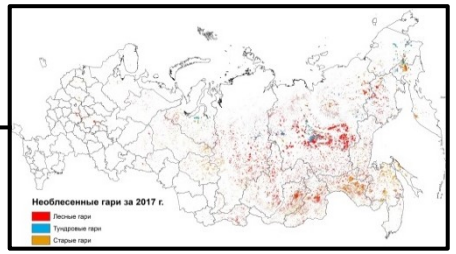
Примеры гистограммы яркостей светлохвойного леса для NIR канала летом



Формирование предварительной обучающей выборки

Опорная карта растительности предыдущего года

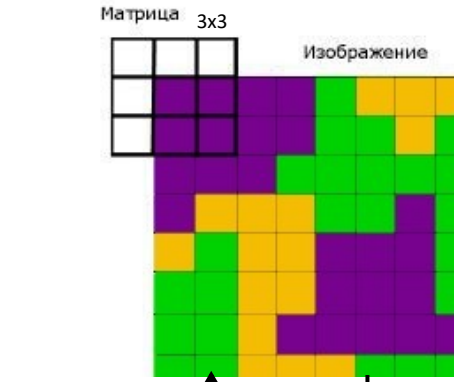
Внешние продукты (гари, с/х, урбанизированная территория, вечные снега)

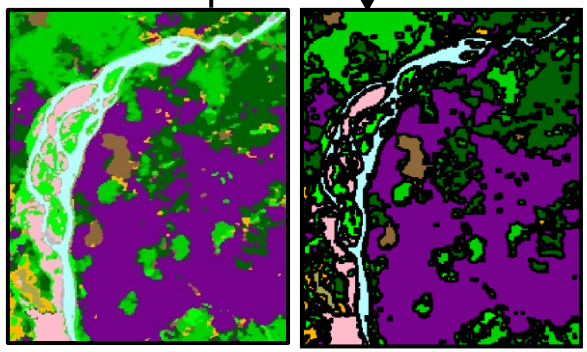


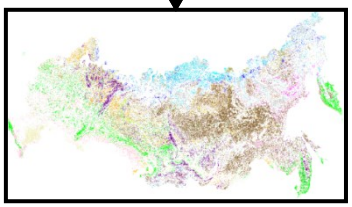
Фильтрация на основе внешних продуктов

Фильтрация граничных условий

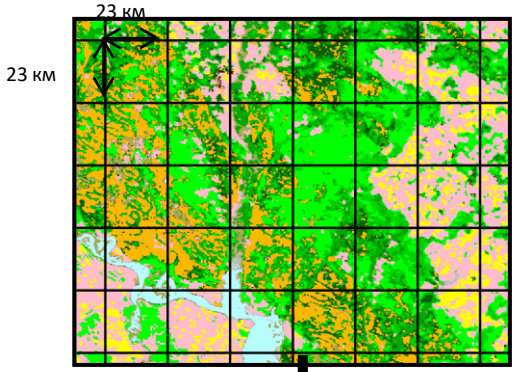
Статистическая фильтрация

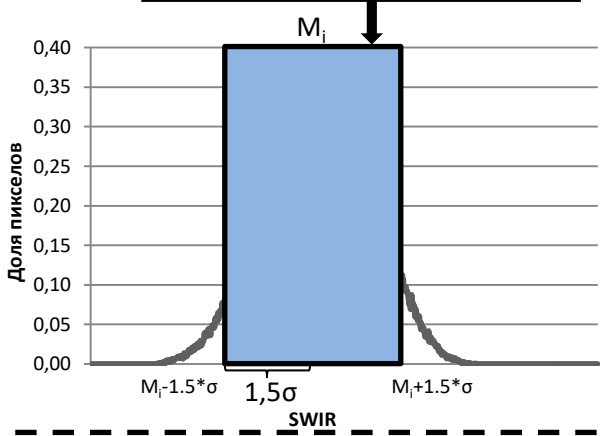






Выборка

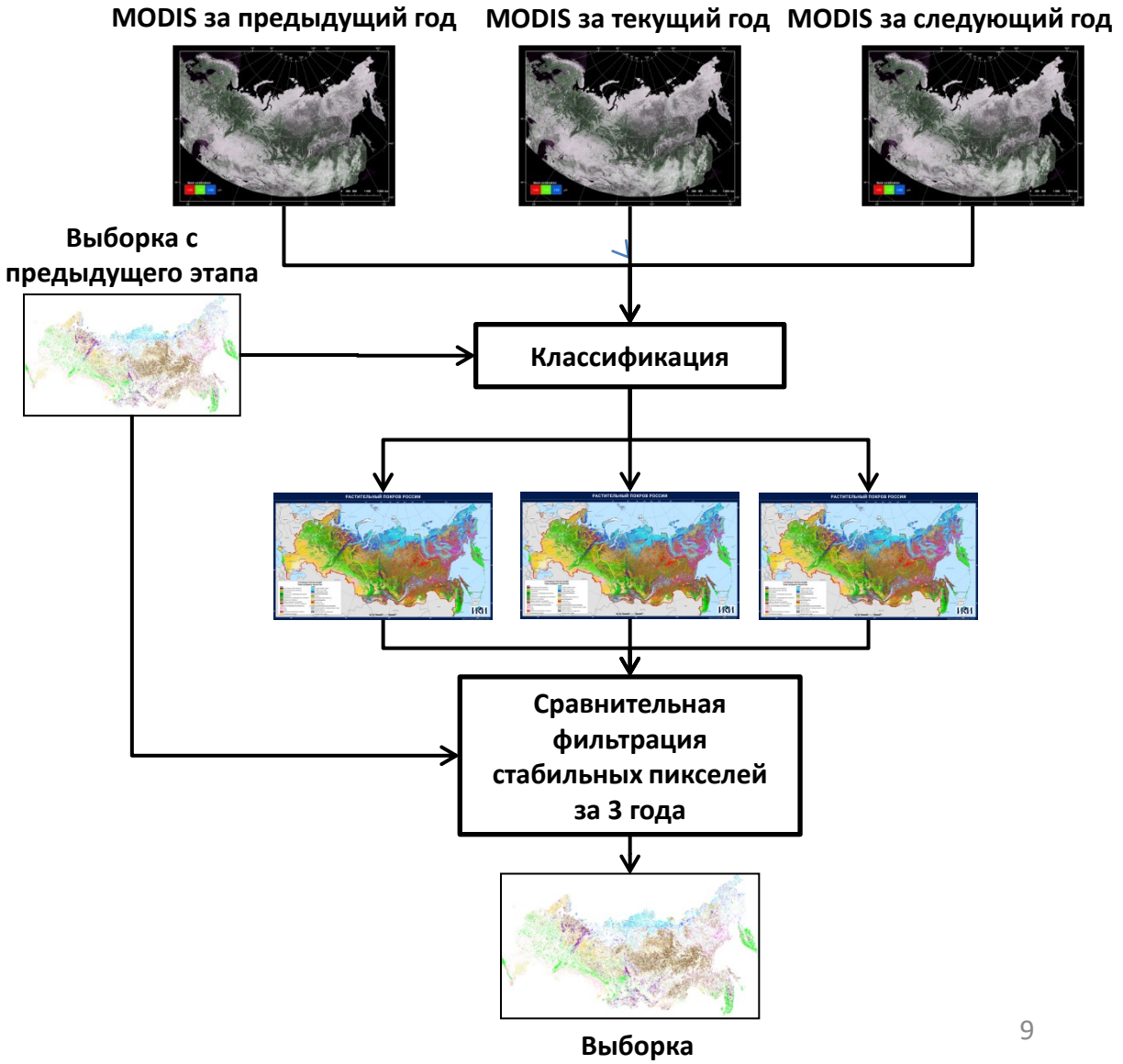




Сравнительная фильтрация обучающей выборки для классов покрытых лесом земель

Полученная выборка, а также данные ДЗЗ за несколько лет использовались для построения набора карт.

В итоговую выборку попадали только пиксели, стабильно классифицируемые как один и тот же класс по данным за несколько лет.

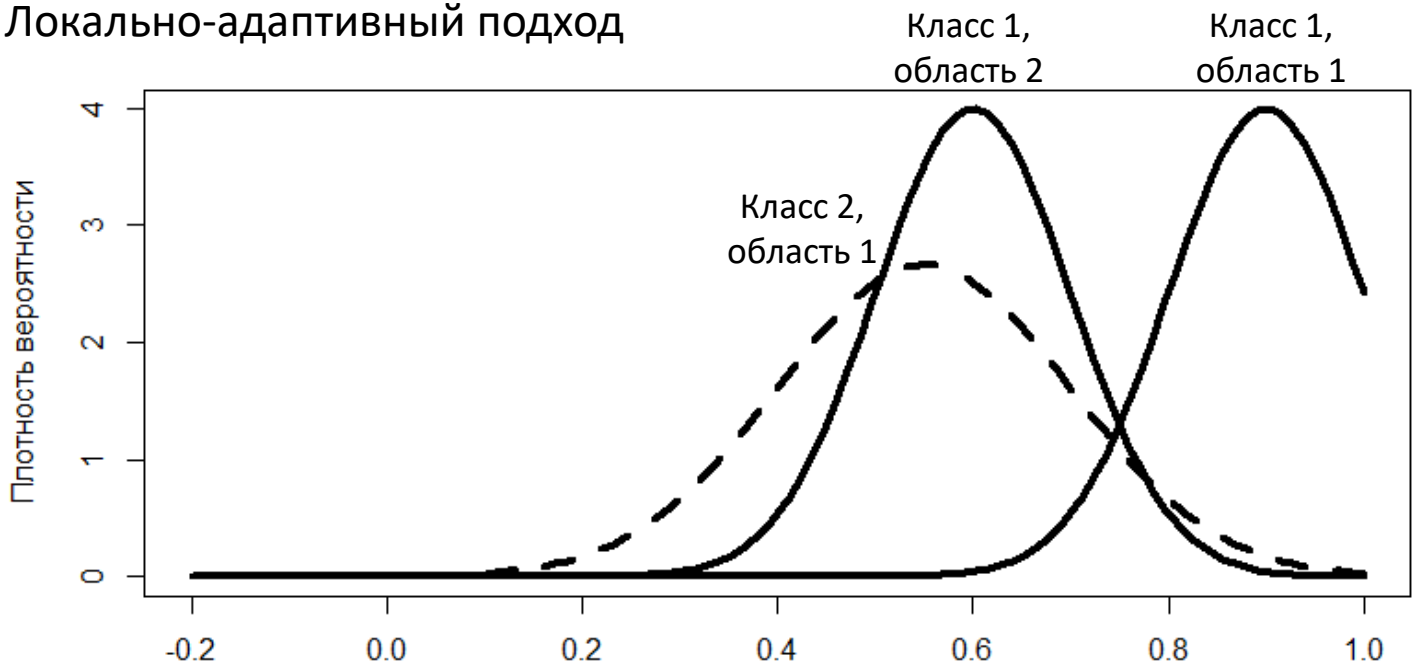


Крупномасштабная классификация

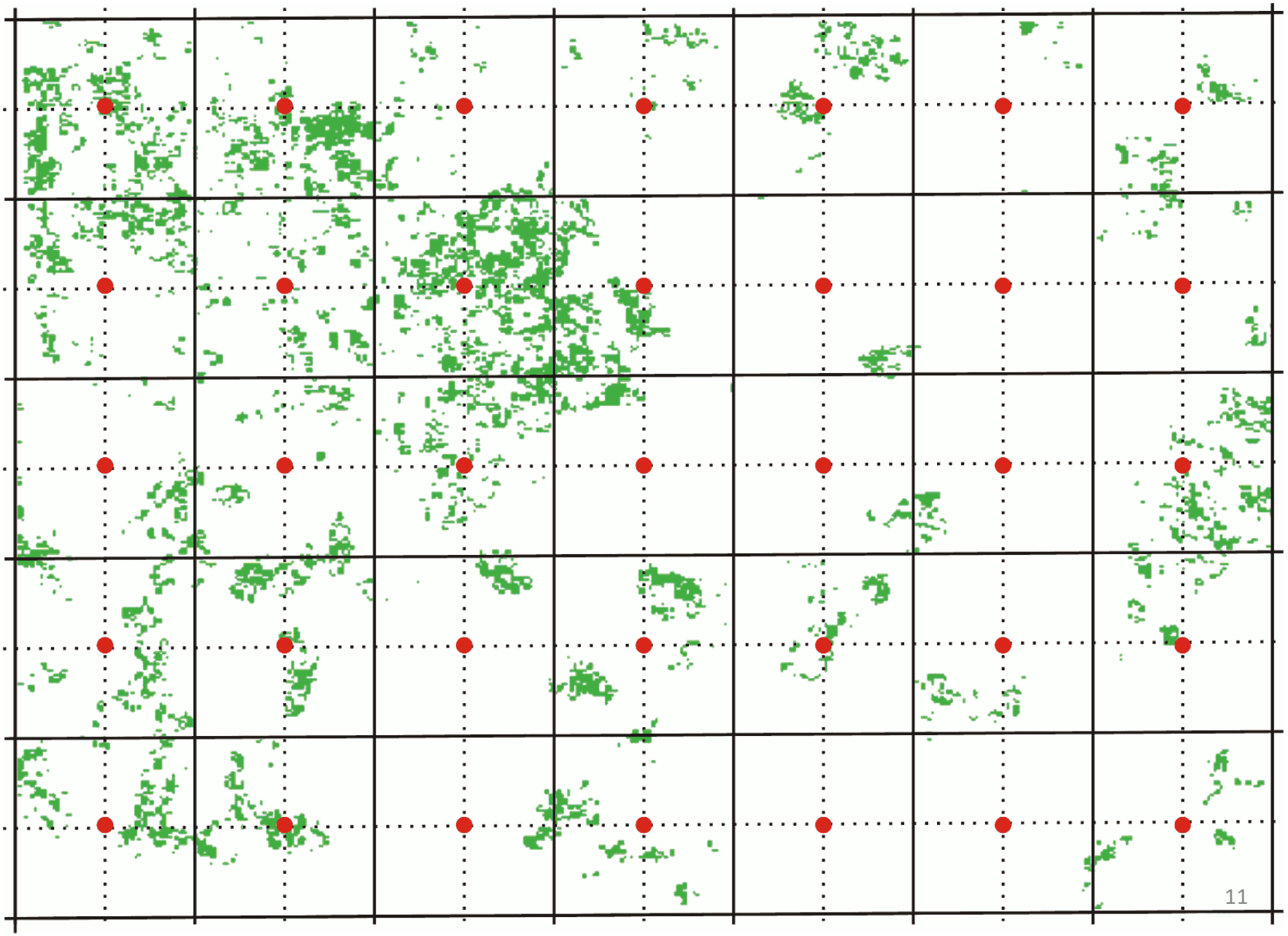
При классификации на больших масштабах возникает проблема пространственной изменчивости спектрально-отражательных характеристик растительного покрова. Один и тот же класс может иметь разные значения признаков в разных местах, и потенциально пересекаться с признаками других классов;

Можно решить с помощью:

- Разбиения классов (гиперкластеризация)
- Разбиения на области (стратификация)
- Локально-адаптивный подход



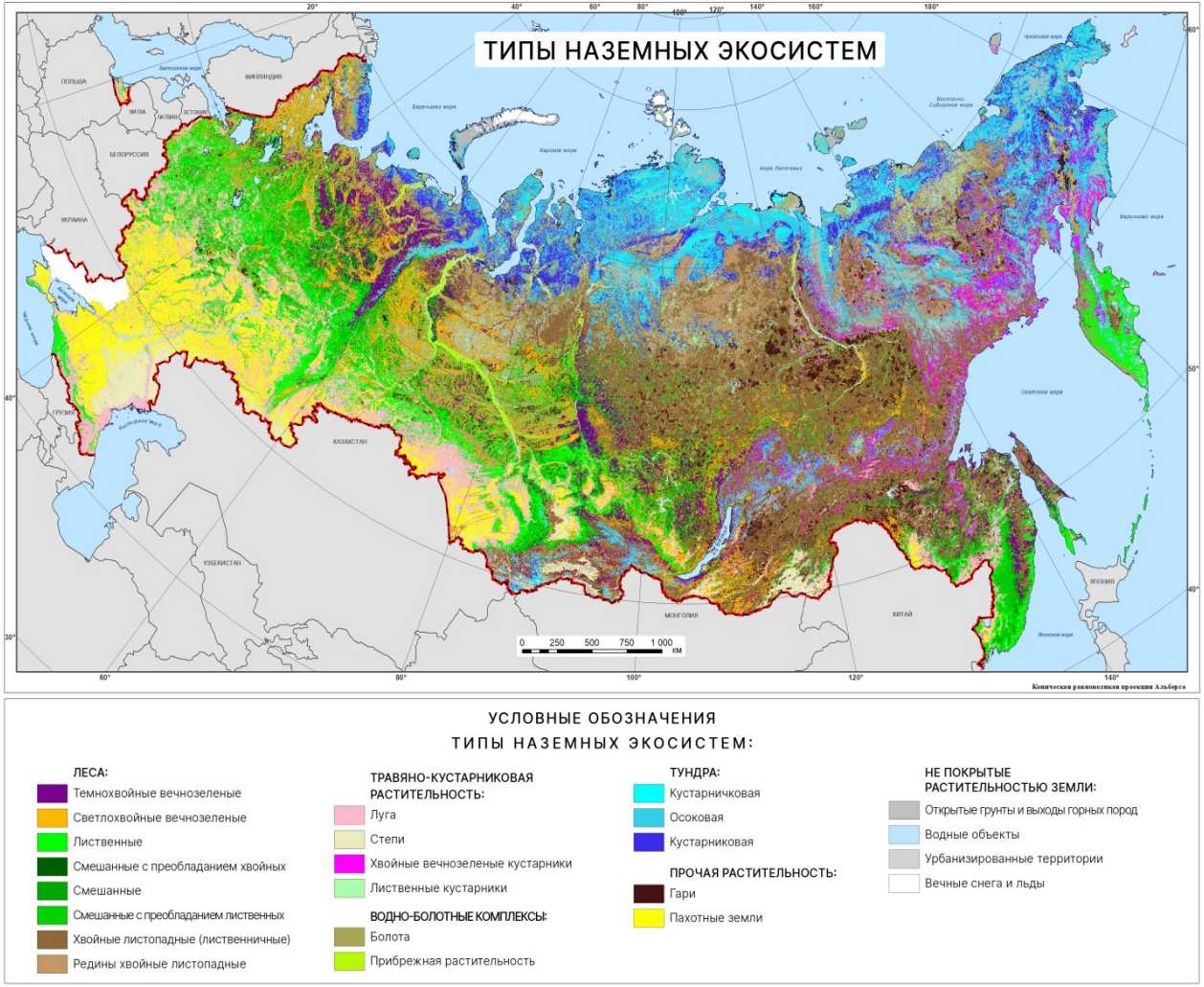
Локально-адаптивный подход LAGMA



Локально-адаптивный подход LAGMA

Локально-адаптивный подход LAGMA (Locally Adaptive Global Mapping Algorithm) позволяет решить проблему пространственной изменчивости классов растительности;

Применение данного подхода значительно увеличивает точность крупномасштабной классификации (около 15% прироста по данным работы Bartalev S. A. et al. A new locally-adaptive classification method LAGMA for large-scale land cover mapping using remote-sensing data //Remote Sensing Letters. – 2014. – Т. 5. – №. 1. – С. 55-64.)



Больше про текущее развитие продукта можно узнать 13 ноября (четверг) на 4 заседании секции F в 10:20 на докладе:

Построение карты наземных экосистем России на основе данных прибора VIIRS
Сайгин И.А., Стыценко Ф.В., Барталев С.А.


КАРТА ЛЕСОВ РОССИИ
УСЛОВНЫЕ ОБОЗНАЧЕНИЯ
Не покрытые лесом территории
Сосна
Ель
Пихта
Лиственница
Кедр
Дуб
Бук
Береза каменная
Береза
Осина
Липа
Клен
0 250 500 1 000 1 500 2 000 2 500
km
ИКИ
Институт космических исследований
Российской академии наук

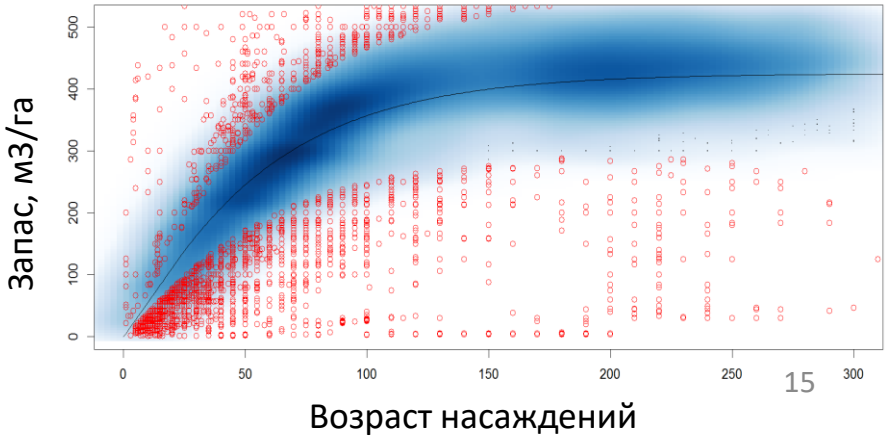
Фильтрация обучающей выборки

Обучающая выборка основана на данных таксационных выделов, огрубленных до уровня пикселов MODIS (230 м);

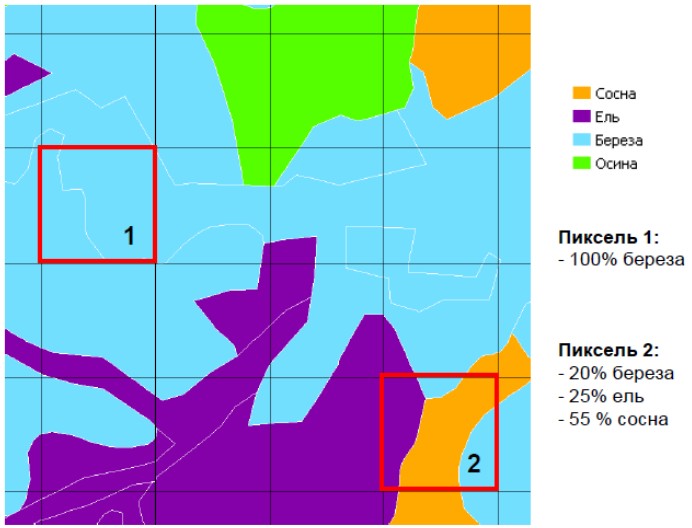
Выполнялась фильтрация опорных данных на уровне таксационных выделов на основе построения моделей хода роста насаждений по запасу древесины

Проводилась фильтрация неоднородных пикселов, пикселов не согласующихся с другими спутниковыми продуктами или на которых наблюдались нарушения;

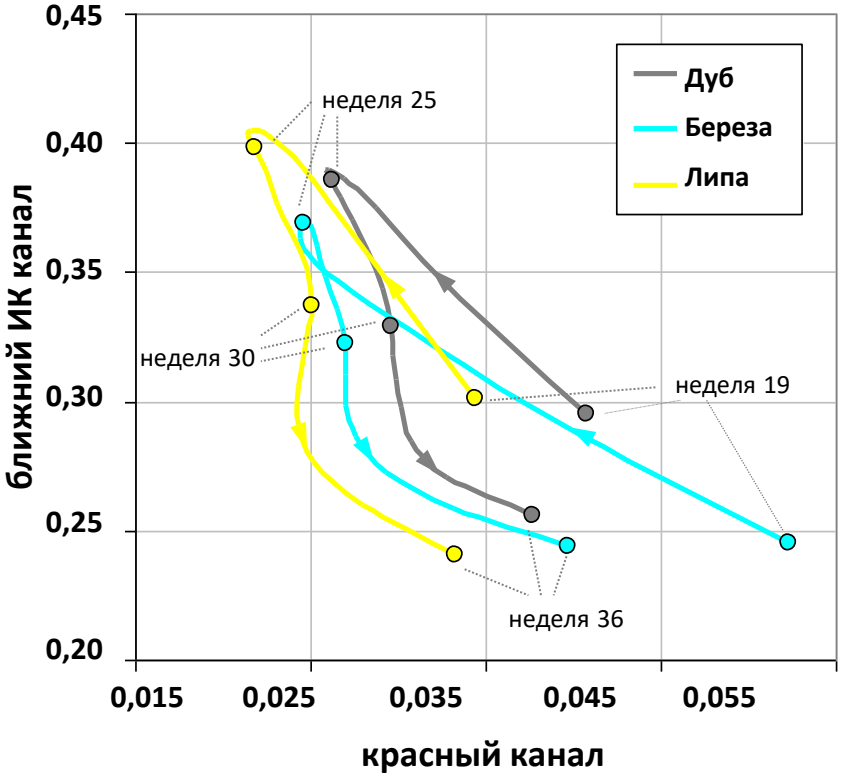
Фильтрация выборки на основе моделей хода роста и данных о возрасте, бонитете, породе и запасе насаждений

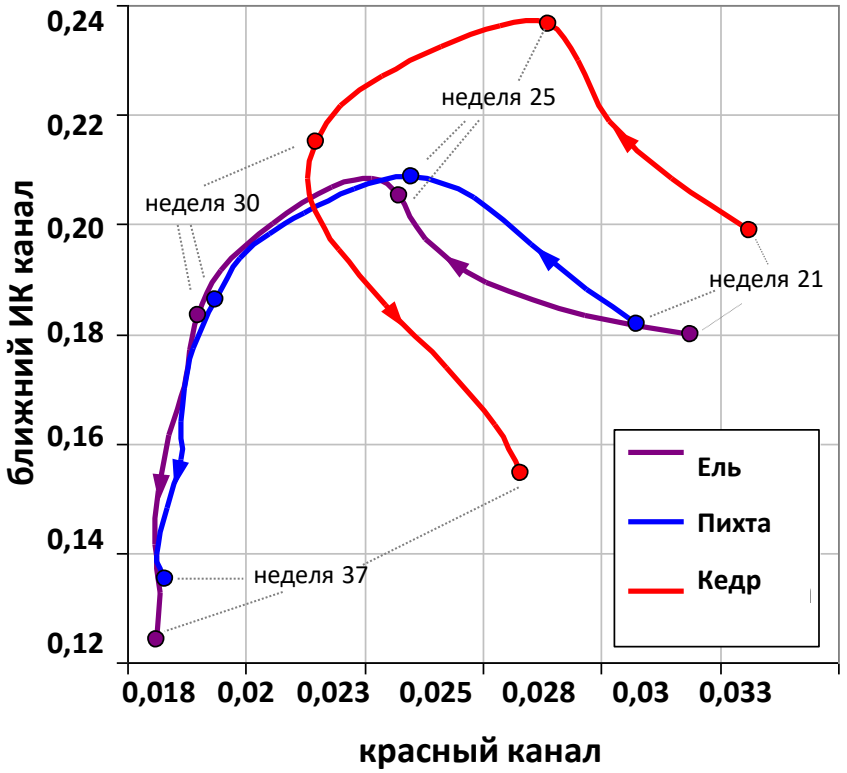


Фильтрация неоднородных пикселов MODIS



Анализ фенологической динамики отражательной способности лесного покрова





Фазовые диаграммы для участков лесного покрова с преобладанием различных пород в пространстве еженедельных измерений отражательной способности в красном (0,62-0,67 мкм) и ближнем ИК (0,84-0,88 мкм) каналах MODIS; движение вдоль кривой соответствует изменению средних значений отражательной способности в течение вегетационного сезона

Оценка точности в задаче классификации

Оценка точности обычно основывается на проверке точности классификации на отложенной части выборки. Возможно простое разделение выборки на 2 части, или использование кросс-валидации.

В задаче классификации для оценки точности используется матрица ошибок, на основе которой рассчитывается ряд метрик.

$$Precision_k = \frac{TP_k}{TP_k + FN_k}$$

$$Recall_k = \frac{TP_k}{TP_k + FP_k}$$

$$F1score_k = \frac{2 * Precision_k * Recall_k}{Precision_k + Recall_k}$$

$$MacroF1 = \sum_k F1score_k$$

$$Accuracy = \frac{\sum_k TP_k}{\sum_k (TP_k + FP_k)}$$

		Прогноз – результат классификации							
		Сосна	Ель	Пихта	Листв.	Кедр	Другие	Precision	F1 score
Фактический класс	Сосна	597 930	14 174	3 708	62 360	16 453	17 018	84.0%	87.9%
	Ель	9 900	523 357	26 295	26 466	45 444	12 663	81.3%	82.7%
	Пихта	1 050	32004	175 678	3 271	43 182	10 714	66.1%	69.5%
	Лиственница	13 771	8 620	1 010	3 320 941	10 449	99 242	96.1%	91.9%
	Кедр	10 510	30 242	27 601	16 371	494 614	23 391	82.1%	79.8%
	Другие	15 858	13 436	5 193	340 780	27 265			
	Recall	92.1%	84.2%	73.4%	88.1%	77.6%			

Где k – класс (преобладающей породы)

TP_k - число корректно классифицированных элементов класса k (True Positive)

FN_k - число элементов класса k, классифицированных как другие классы (False Negative)

FP_k - число элементов других классов, ошибочно классифицированных как класс k (False Positive)

Оценка точности карты пород

Полная матрица ошибок для карты пород выглядит следующим образом:

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Precision	F1-score
Сосна	**1**	597930	14174	3708	62360	16453	219	16	10	30	16	25	5	13181	1592	52	34	21	23	19	55	1670	12	37	1	84.0%	**87.9%**
Ель	**2**	9900	523357	26295	26466	45444	102	7	0	56	2	34	284	9631	1219	88	36	212	19	16	129	785	6	35	2	81.3%	**82.7%**
Пихта	**3**	1050	32004	175678	3271	43182	84	19	0	38	3	11	412	6785	3067	8	0	98	15	24	65	78	2	1	4	66.1%	**69.5%**
Лиственница	**4**	13771	8620	1010	3320941	10449	339	0	0	12	0	13	372	12481	1009	41	3	28	69	26	2340	81417	643	197	252	96.1%	**91.9%**
Кедр	**5**	10510	30242	27601	16371	494614	3620	0	6	1768	124	563	73	13197	954	35	0	2894	54	5	37	53	3	4	1	82.1%	**79.8%**
Дуб	**6**	238	5	3	150	1063	162388	350	442	1130	442	525	114	11114	4708	370	267	3695	236	213	0	4	0	22	0	86.6%	**86.2%**
Бук	**7**	3	0	3	0	0	293	10276	209	14	38	2	0	58	42	27	64	17	0	1	0	0	0	0	0	93.0%	**92.0%**
Граб	**8**	3	2	0	0	5	497	334	711	47	4	3	0	35	28	52	14	13	4	4	0	0	0	0	0	40.5%	**44.1%**
Ясень	**9**	62	93	18	7	2699	1781	5	3	6508	117	598	3	1245	575	58	35	841	38	106	0	0	0	7	0	44.0%	**49.6%**
Клен	**10**	2	0	0	0	7	189	7	1	20	2612	60	0	232	117	8	11	2232	9	4	0	0	0	0	0	47.4%	**52.6%**
Вяз	**11**	38	70	9	5	677	645	0	0	540	135	2171	8	608	154	23	15	187	75	76	0	1	0	10	0	39.9%	**42.6%**
Береза кам.	**12**	3	753	209	858	192	552	0	0	7	71	17	108459	3544	94	500	0	32	46	96	403	171	1	0	135	93.4%	**92.7%**
Береза	**13**	9982	8432	3177	70148	17198	10254	222	65	402	244	311	4085	994123	60068	2870	655	3522	370	502	116	2388	37	28	9	83.6%	**83.4%**
Осина	**14**	1659	552	1306	4555	1603	4044	5	5	345	78	55	60	100460	158324	562	81	2402	18	93	4	124	2	4	0	57.3%	**61.7%**
Ольха серая	**15**	133	82	2	89	2	284	8	5	12	12	32	1683	11526	1974	10669	324	254	11	105	46	23	5	0	83	39.0%	**49.0%**
Ольха черная	**16**	75	42	0	14	2	157	28	8	7	25	9	0	3057	385	546	2476	64	10	51	7	2	0	0	0	35.5%	**44.6%**
Липа	**17**	2	154	34	16	3115	3326	4	2	454	446	268	16	5795	2006	118	54	25246	28	11	0	1	0	1	0	61.4%	**60.9%**
Тополь	**18**	39	88	23	661	34	113	1	1	13	3	27	44	1036	172	15	7	49	881	35	74	106	10	21	3	25.5%	**32.3%**
Ива	**19**	404	237	75	1738	64	199	0	2	34	35	24	127	4466	436	58	68	19	70	1186	75	166	14	69	9	12.4%	**19.5%**
Кедровый стланик	**20**	78	299	107	17007	278	6	0	0	0	4	0	744	61	1	18	0	0	5	8	73892	5488	393	39	239	74.9%	**82.0%**
Редк. Лиственница	**21**	2971	2463	200	225759	305	33	0	0	2	2	2	135	2350	148	4	1	2	9	6	2962	877234	1525	216	262	78.6%	**83.4%**
Карл. Береза	**22**	62	112	5	12690	4	0	0	0	0	0	0	3	93	3	0	0	0	1	2	1008	9878	12292	301	180	33.6%	**47.2%**
Куст. Ива	**23**	99	44	21	3986	17	12	0	0	5	1	2	1	193	10	1	1	1	10	18	98	4626	315	4585	82	32.5%	**46.4%**
Куст. Ольха	**24**	5	8	1	3097	0	0	0	0	0	0	0	1288	18	0	18	0	0	0	2	319	3735	145	55	10238	54.1%	**67.3%**
Recall		92.1%	84.2%	73.4%	88.1%	77.6%	85.9%	91.1%	48.4%	56.9%	59.2%	45.7%	92.0%	83.2%	66.8%	66.1%	59.7%	60.4%	44.0%	45.5%	90.5%	88.8%	79.8%	81.4%	89.0%		**85.5%**

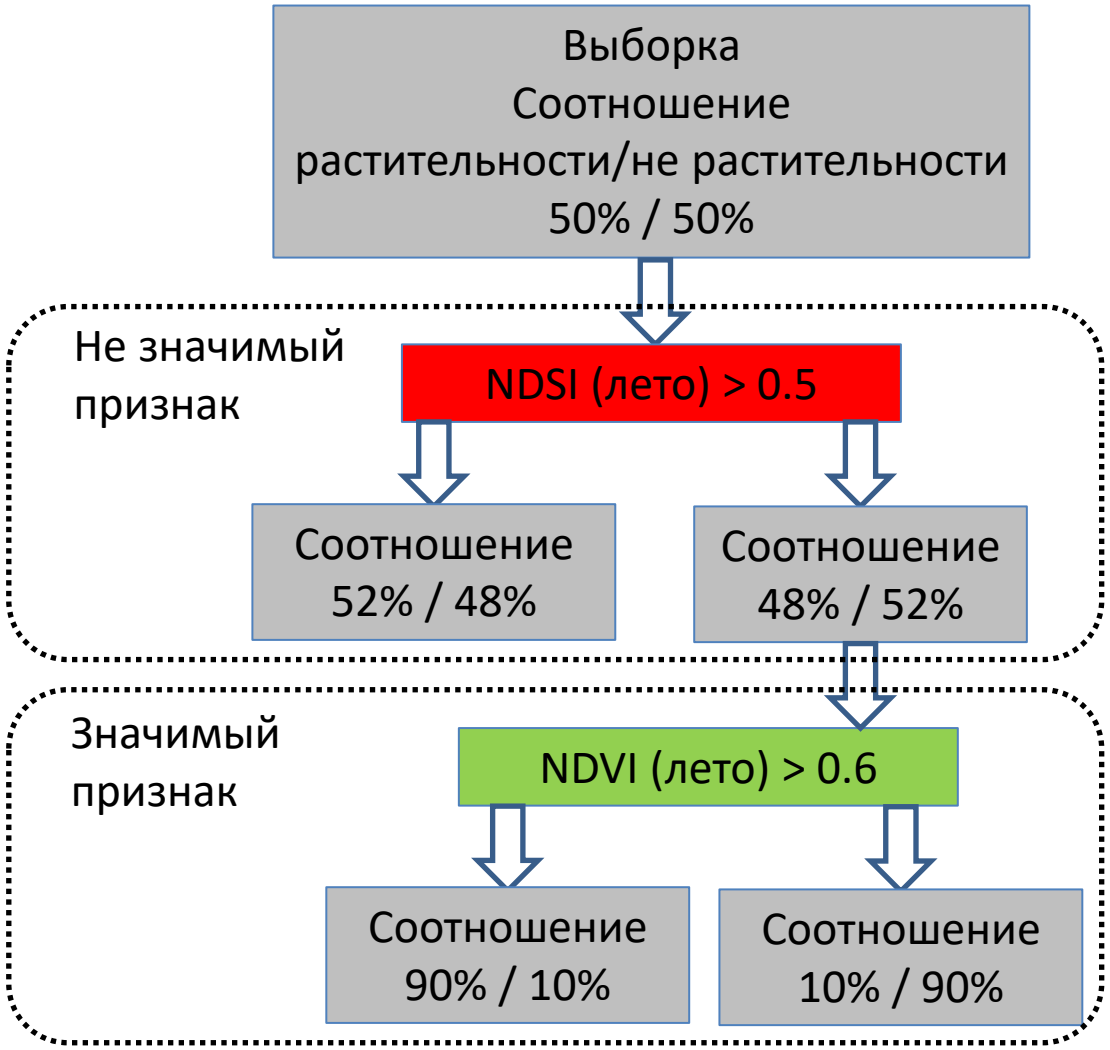
Продуктивность лесов России

Методы оценки значимости, деревья решений

Случайные леса и другие популярные методы на основе деревьев решений используют только часть признаков в каждом узле;

Значимость признака можно определить по тому, насколько хорошо признак способен разбить исходную выборку;

Агрегация такой статистики по большому количеству узлов и деревьев может дать адекватную итоговую оценку значимости.

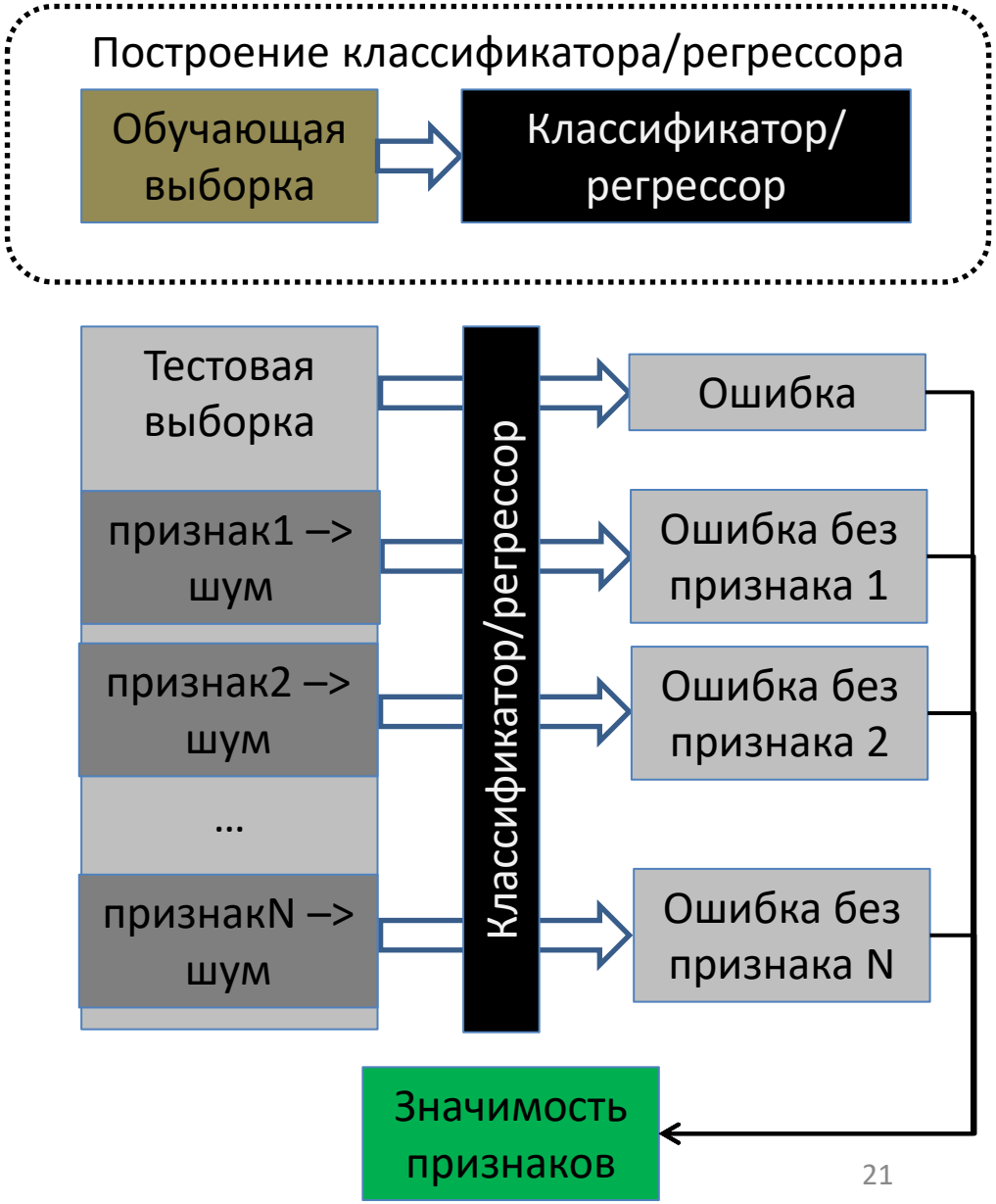


Методы оценки значимости, изменение признаков

Другой подход оценки значимости основан на использовании уже готового классификатора;

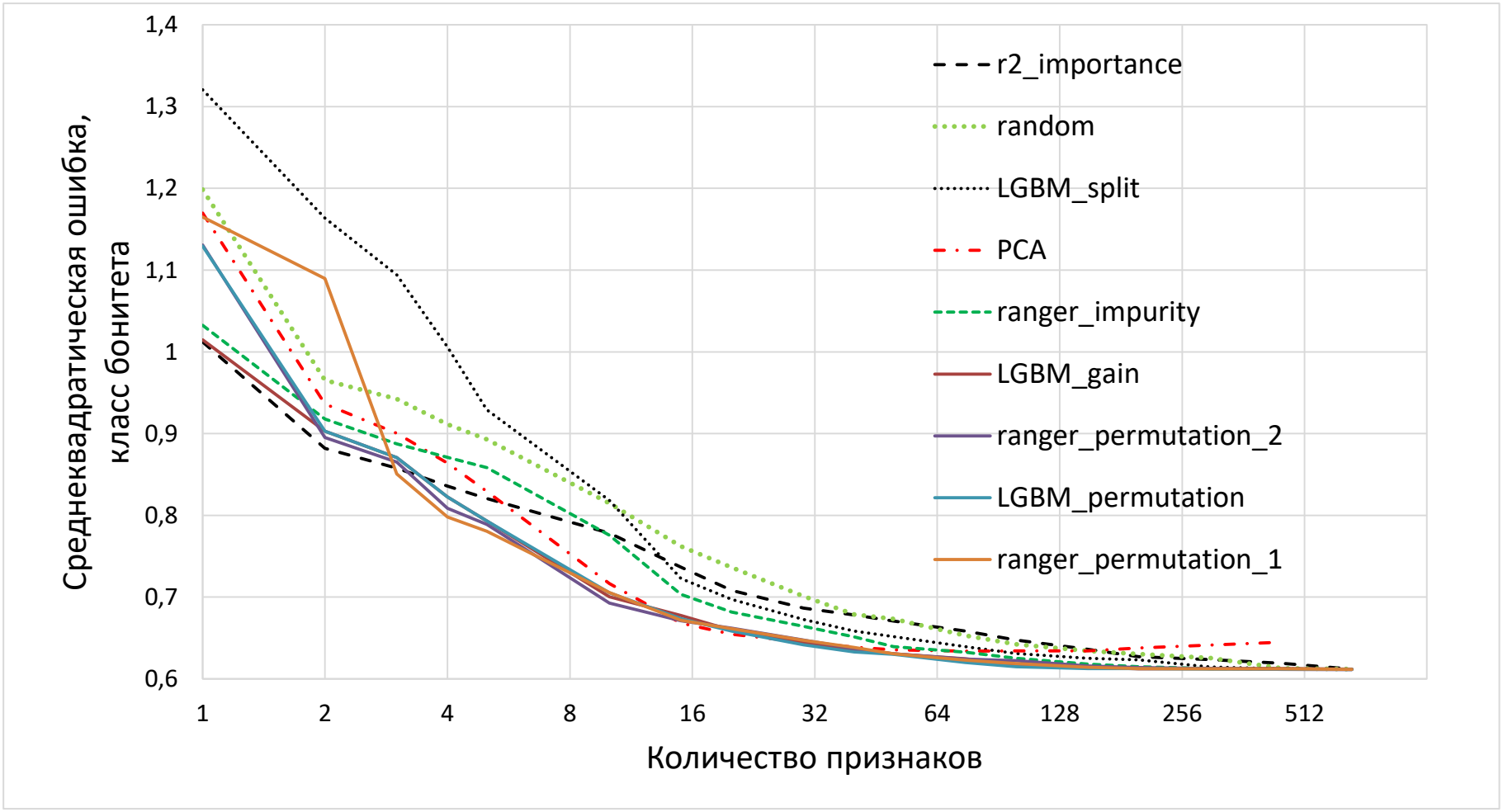
На его вход можно подать тестовую выборку, в которой один признак заменен на шум, и оценить падение точности;

Агрегация падений точности при подмене признака позволяет оценить значимость признаков.



Выбор оптимального метода оценки значимости

Анализ показал преимущество методов на основе перестановки значений признака.



Выбор оптимального метода машинного обучения

По оптимальному набору признаков было проверено несколько методов машинного обучения. Был сделан выбор в пользу градиентного бустинга на основе LightGBM

Метод		R2	RMSE	Время обучения, минут
Линейая регрессия	Базовая	0,65	0,81	0,05
	Ridge	0,65	0,81	0,05
	Lasso	0,65	0,81	2
	Elastic Net	0,65	0,81	2
К ближайших соседей		*0,7*	*0,76*	-
Метод опорных векторов		*0,73*	*0,72*	-
Случайный лес		0,776	0,65	75
XGBoost		0,783	0,64	5
LightGBM		0,816	0,59	8

Понимание общей ситуации по машинному обучение в ДЗЗ:

- Случайный лес – стандартный метод классификации;
- Градиентные бустинг - часто лучше, но требует настройки;
- Нейронные сети – область активного научного интереса, требуют некоторой экспертизы для получения результата;
- Другие методы – нишевые.

$$RMSE = \sqrt{\frac{1}{N}\sum(Y_{TRUE} - Y_{ML})^2}; R2 \approx 1 - \frac{\sum(Y_{TRUE} - Y_{ML})^2}{\sum(Y_{TRUE} - Y_{MEAN})^2}; MAE = \frac{1}{N}\sum ABS(Y_{TRUE} - Y_{ML});$$

Оценка точности полученной карты бонитета

Использование оптимального набора из 100 признаков, регрессии на основе LGBM и выборки из 25 000 000 элементов с валидацией на 3 000 000 элементах позволило получить финальный вариант оценки бонитета и возраста.

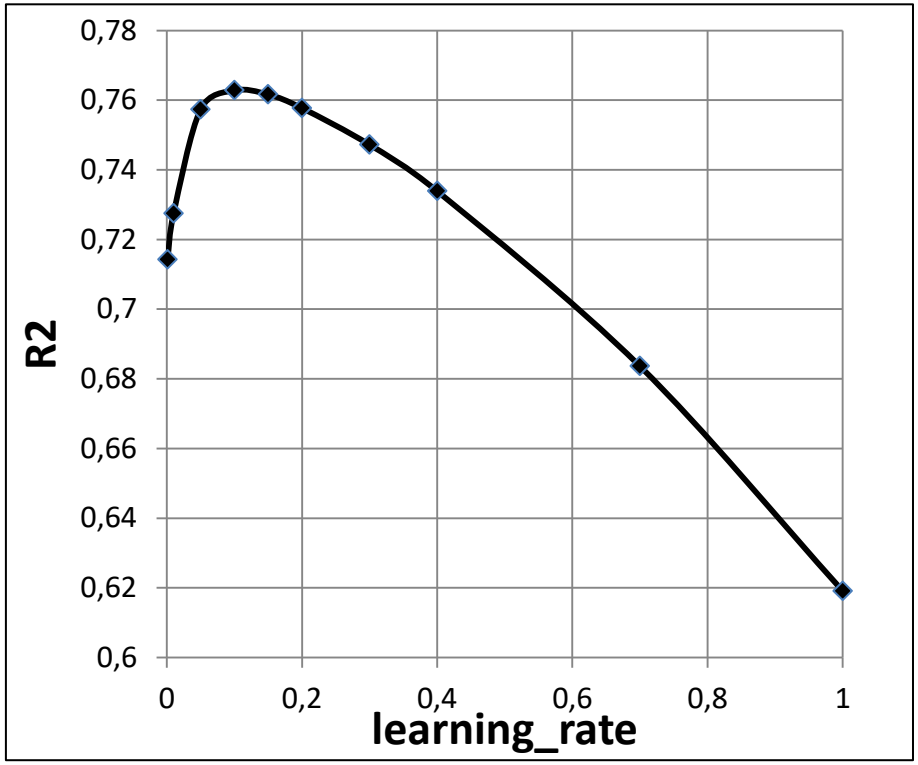
	R2	RMSE	MAE
Бонитет	0,867	0,503	0,37

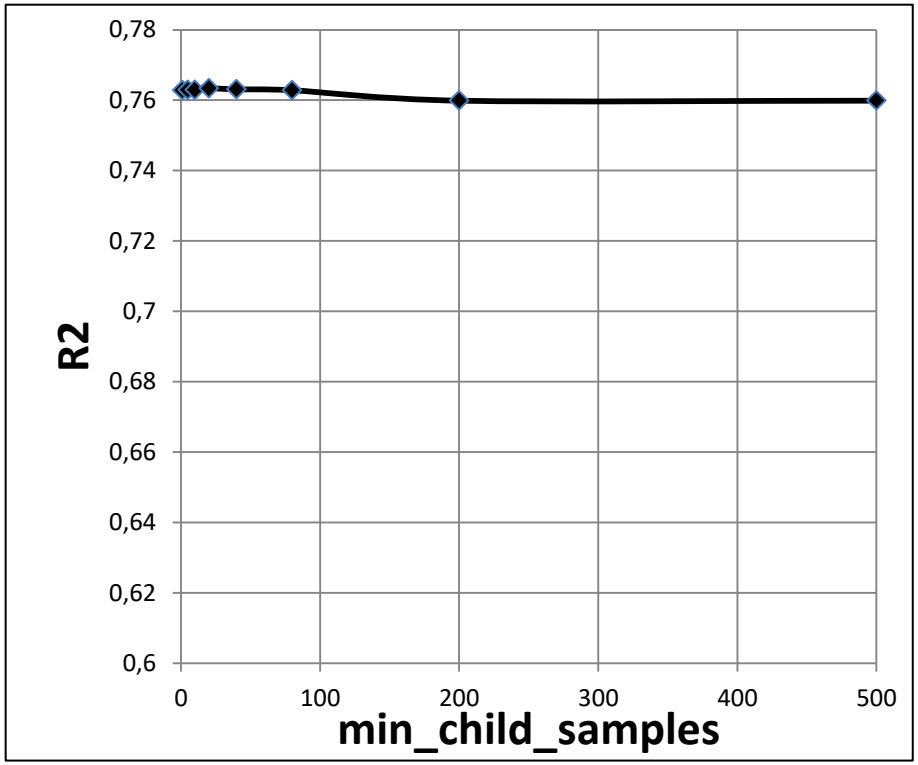
Сравнение с данными ТДУ ГИЛ показывает, что в 83% случаев предложенный подход позволяет корректно определять бонитет лесов с погрешностью, не превышающей 1 класс бонитета.

Настройка LGBM

Очень важно настроить параметр learning_rate.

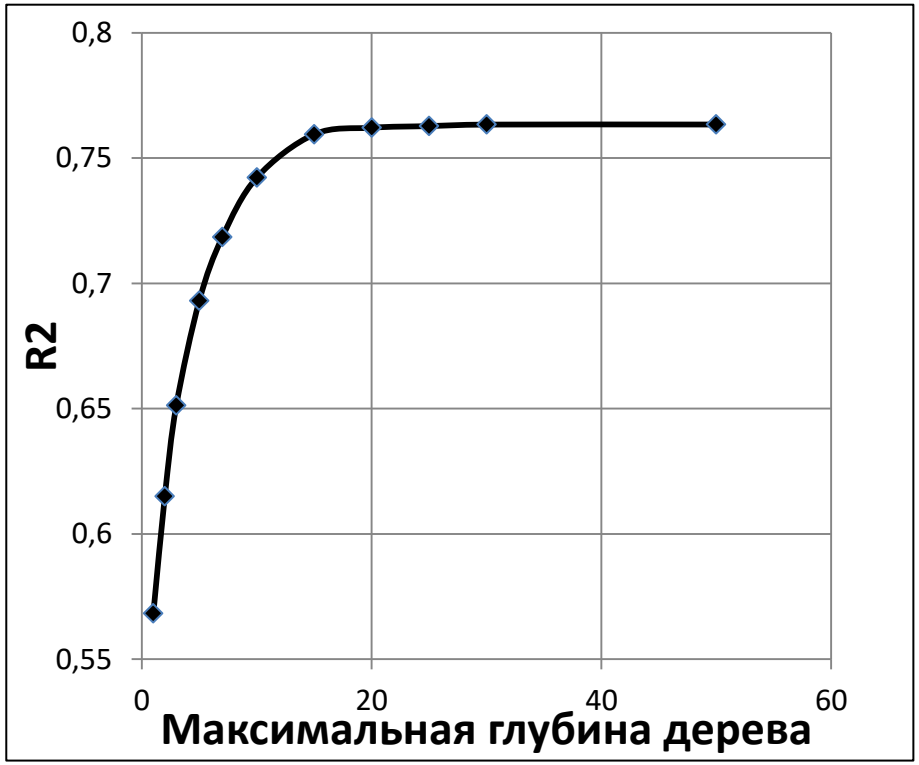
Остальные параметры играли очень небольшую роль.

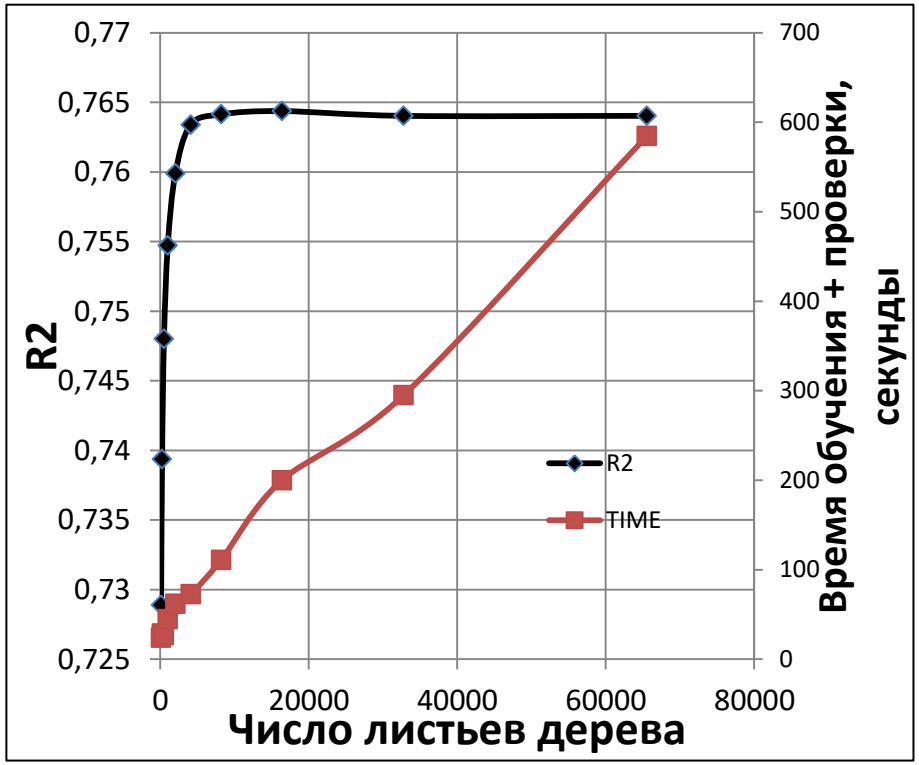




Настройка LGBM

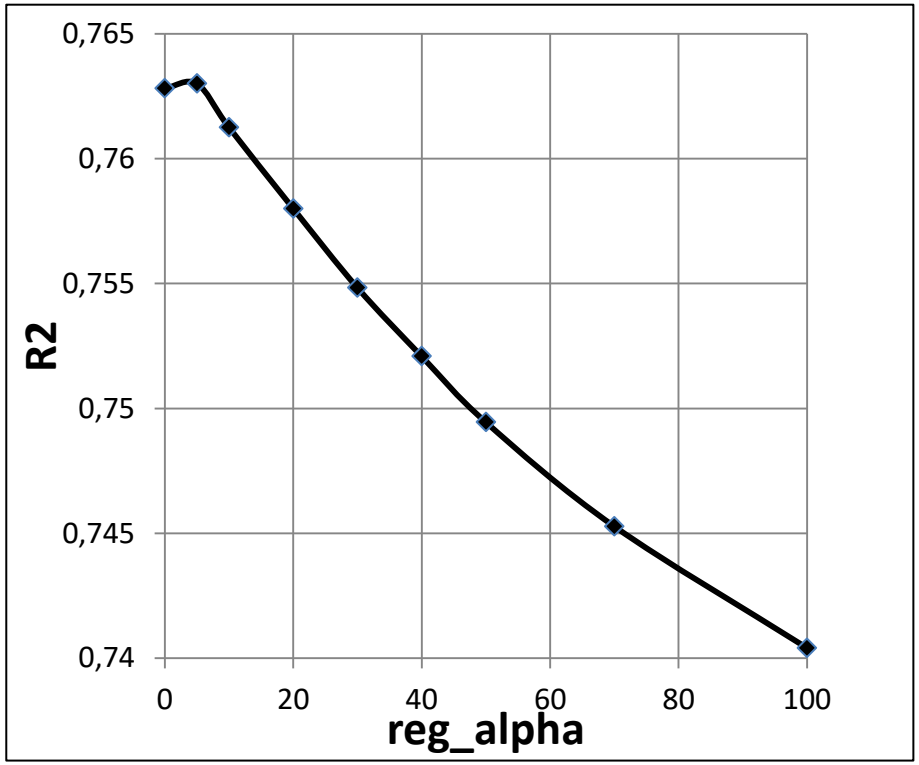
Нужно установить число листьев (узлов) и глубину дерева, который в этом случае имеют некоторый оптимальный порог.
В теории большие значения могут весть к переобучению, но этот эффект слабо проявился при оценке бонитета.

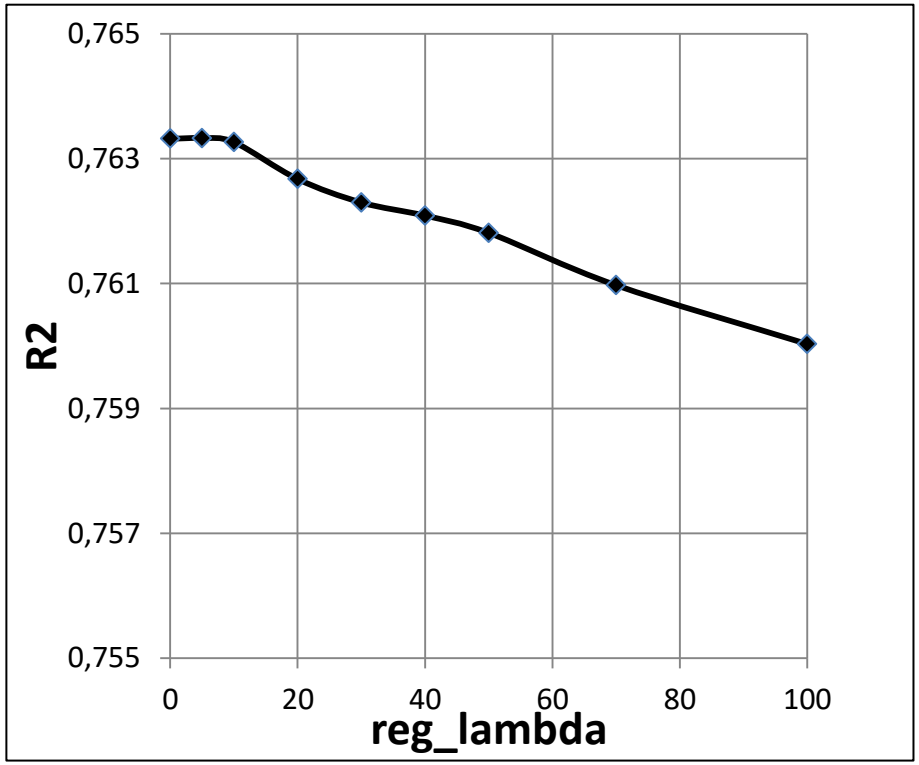




Настройка LGBM

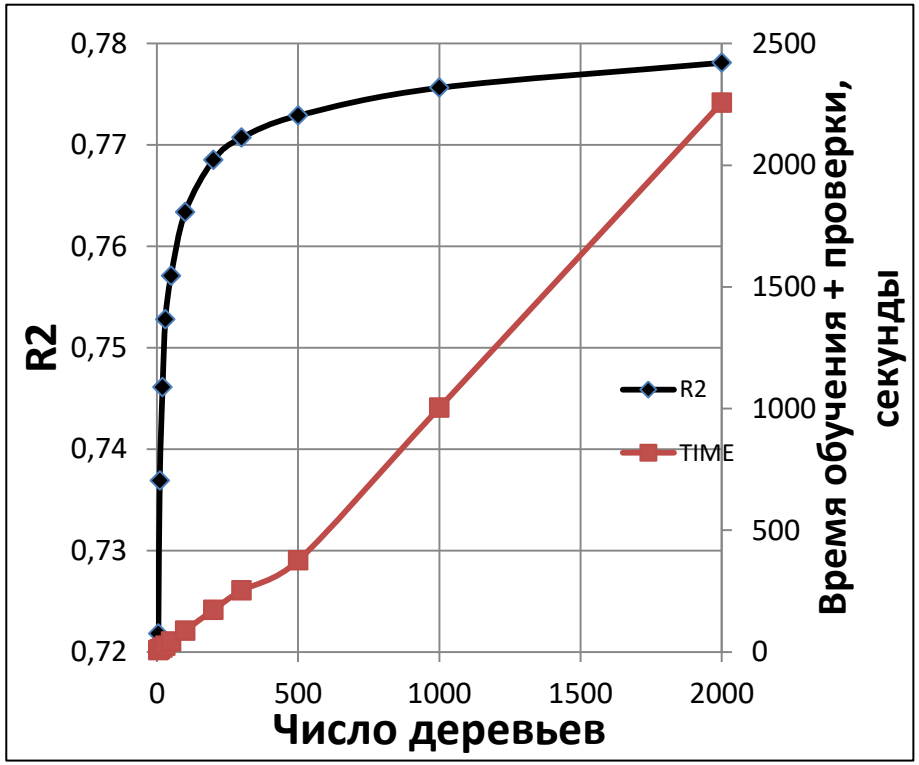
Можно отметить параметры регуляризации (reg_alpha, reg_lambda), которые способны уменьшить переобучение LightGBM

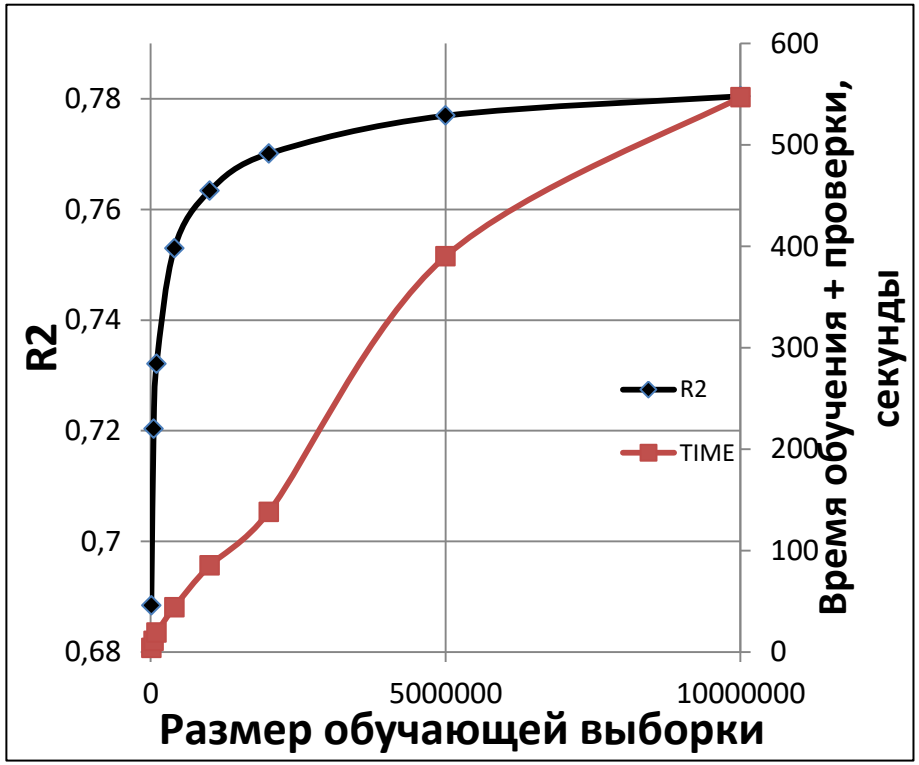




Настройка LGBM

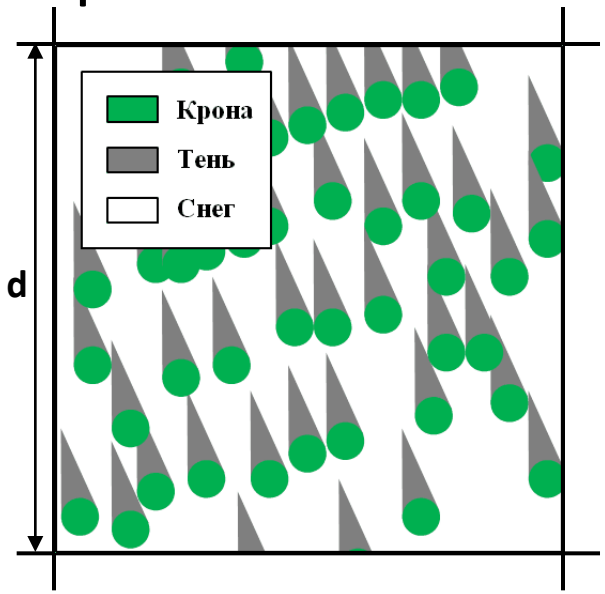
Точность LGBM сильно зависит от объема выборки, а также числа деревьев. Больше – лучше, но наблюдается насыщение.





Оценка запасов древесины лесов России

Оценка запаса древесины на основе измерений отражательной способности покрытого снегом леса



Sc – площадь снега

Sk – площадь крон

St – площадь теней

h – высота деревьев; n – число деревьев

Отражательная способность в красном канале:

$$R = f\left(S_c, S_k, S_t\right);$$

$$S_c = d^2 - S_k - S_t,$$

$$S_k = f_1\left(n\right), S_t = f_2\left(n, h\right),$$

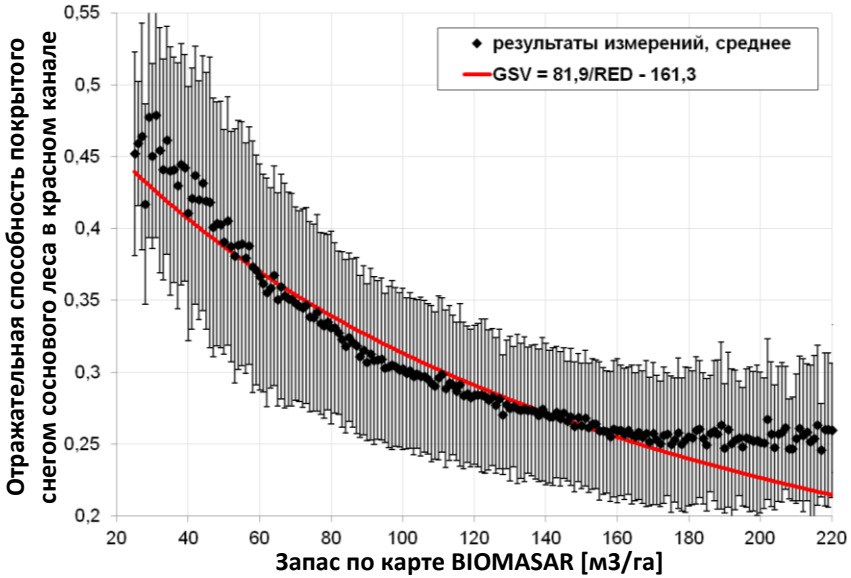
$$R = f_3(n, h);$$

Запас в пикселе:

$$GSV\left[м^3 / га\right] = f_4(n, h)$$

Модель:

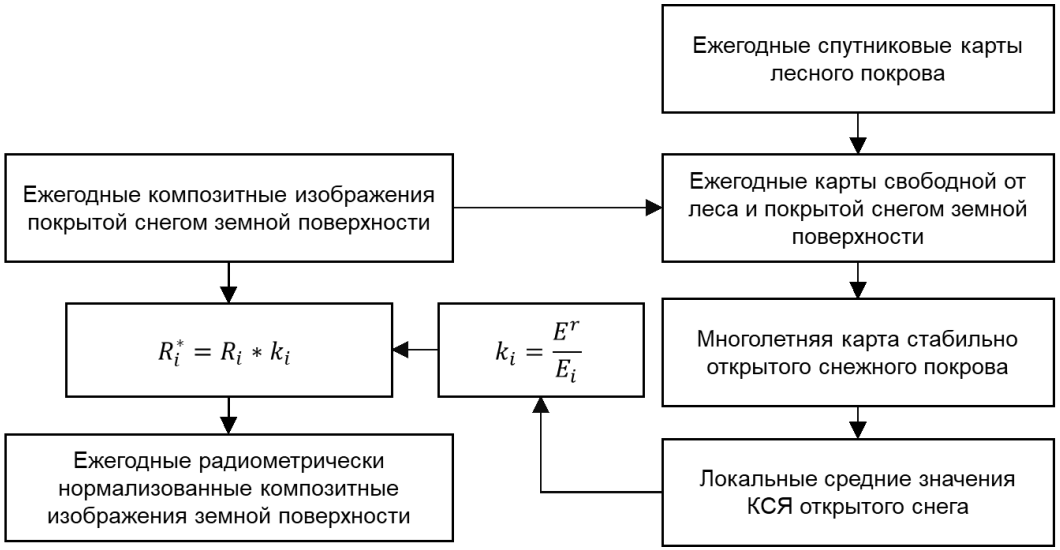
$$GSV\left[м^3 / га\right] \sim 1/R$$



Оценка запасов древесины – исходные данные

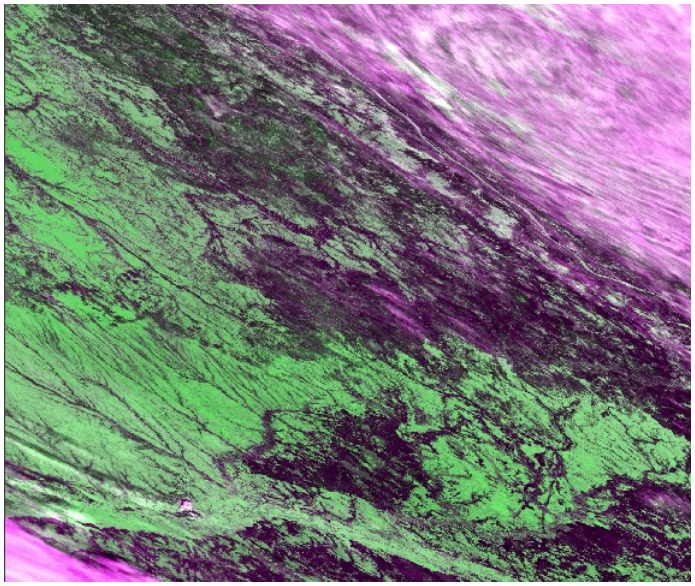
Текущий алгоритм использует гибридную обучающую выборку: наземные данные и отдельные наблюдения продукта Globbiomass;

В качестве признаков используются зимние композиты, приведенные к одному углу отражения (45 градусов), КСЯ скорректированы к 2010 году.



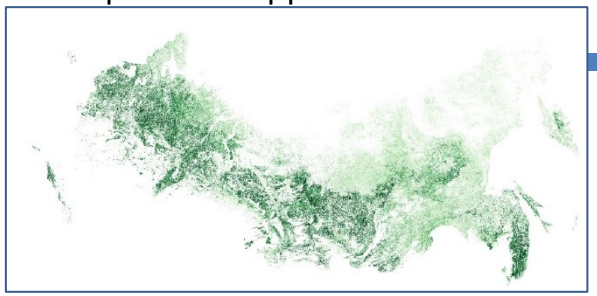
E_i и E^r - локальные средние значения КСЯ открытого снежного покрова, полученные на регулярной сети с использованием композитных изображений произвольного i – го и опорного годов; R_i и R_i^* - значения КСЯ в пикселах исходного и радиометрически нормализованного композитных изображений; k_i - заданные на регулярной сети значения коэффициентов радиометрической нормализации композитного изображения.

Terra-MODIS
Red-NIR-Red

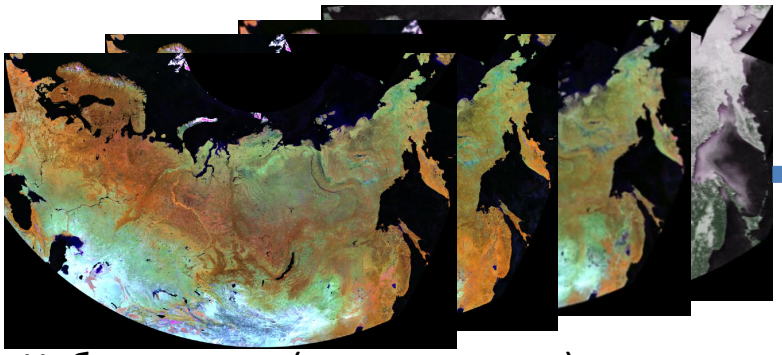


Оценка запасов древесины – схема построения

Выборка по наземным и дистанционным данным за 2010 год

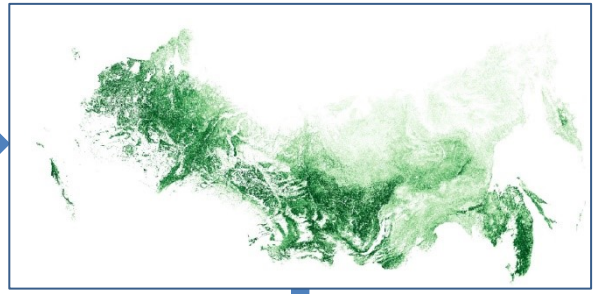


Набор сезонных композитов MODIS за 2010 год



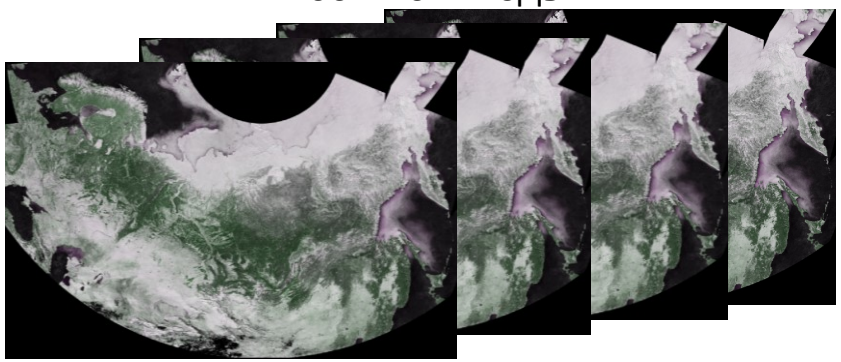
LightGBM

Расширенная выборка 2010 года

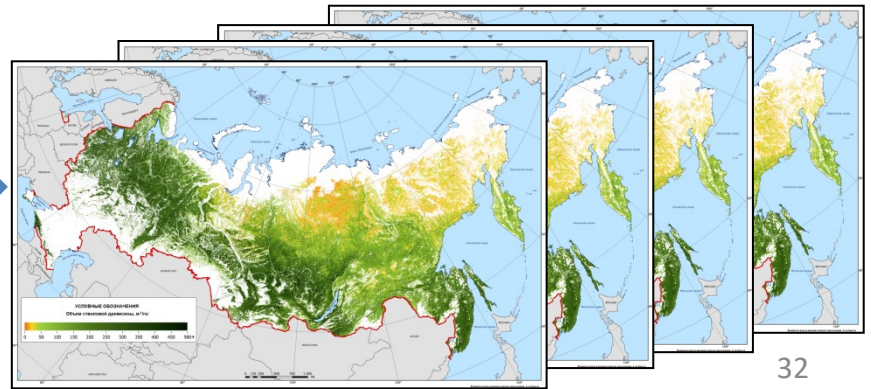


Обучение случайного леса на данных 2010 год

Набор зимних (инвариантных) композитов за 2001-2024 годы



Набор карт запаса за 2001-2024 годы



Оценка запасов древесины, сглаживание временного ряда

Погрешность оценки запаса (10%) значительно выше его ожидаемого ежегодного изменения, что потребовало добавления этапа сглаживания на основе алгоритма STL (Season-Trend decomposition using LOESS)

Больше про текущее развитие продукта можно узнать 12 ноября (среда) на 2 заседании секции F в 12:20 на докладе:

Развитие оценки запаса стволовой древесины на основе данных Sentinel-2
Ворушилов И.И., Барталев С.А., Егоров В.А.

Оценка площади лесов России

При анализе данного продукта было обнаружено, что построение двух отдельных карт – для древесной растительности и для кустарников – приводит к значительному увеличению точности оценки лесистости кустарников

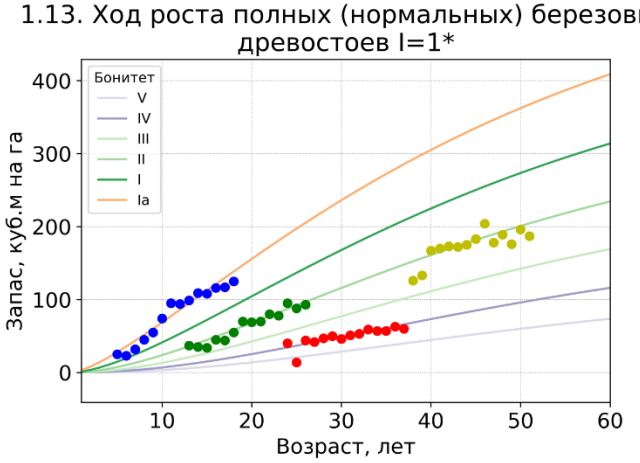
Больше про анализ данного и некоторых других продуктов можно узнать 12 ноября (среда) на 2 заседании секции F в 10:20 на докладе:

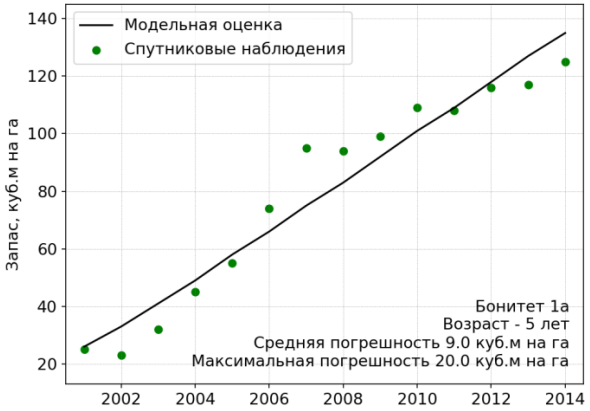
Уточнение данных о площади погибших и поврежденных лесов от пожаров и оценка скорости лесовосстановления на гарях на основе анализа спутниковых данных за период с 2001 по 2024 год
Ховратович Т.С., Барталев С.А., Стыценко Ф.В., Сайгин И.А., Стыценко Е.А.

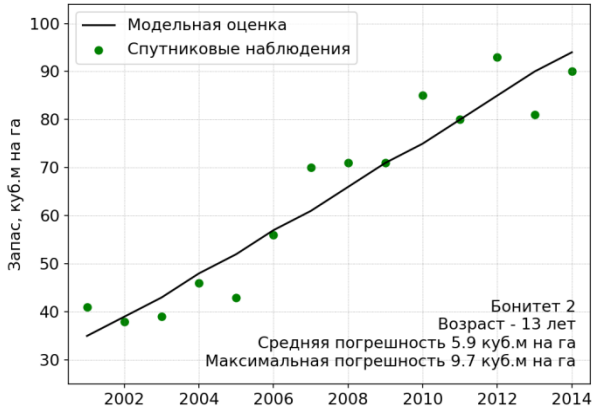
Возраст лесов России

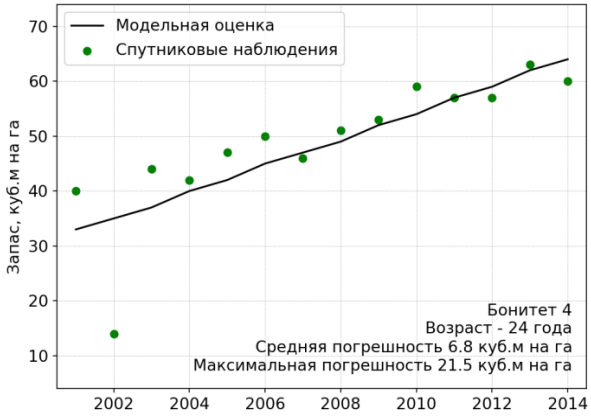
Оценка возраста лесов на основе моделей и данных ДЗЗ

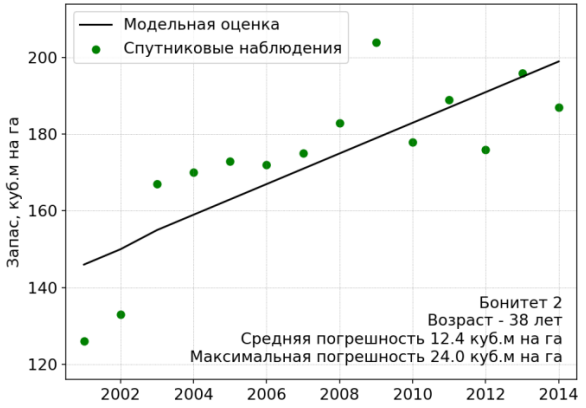
Для построения обучающей выборки используются модельные оценки возраста на основе моделей хода роста, значений запаса и бонитета лесов.









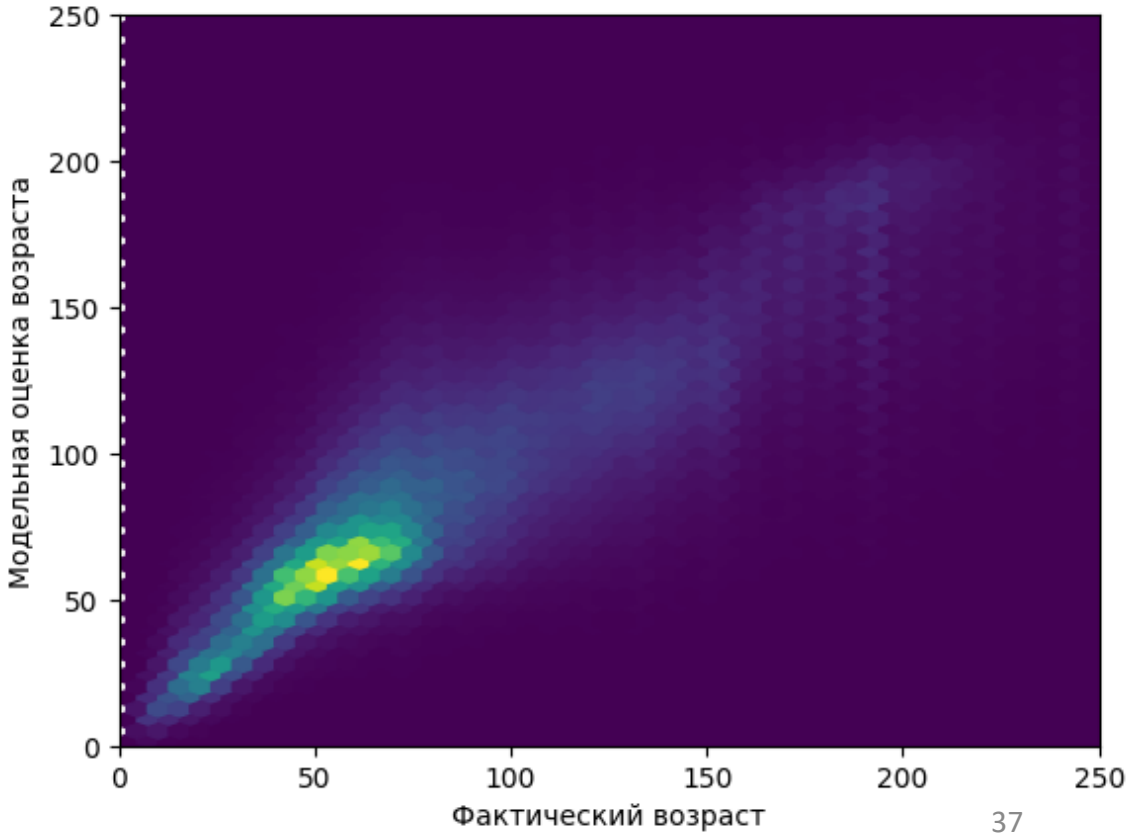


Оценка точности полученной карты возраста

Построение карты возраста лесов выполнялось по аналогии с методом построения карты бонитета. Был получен адекватный уровень точности.

	R2	RMSE	MAE
Возраст	0,74	32,1	22,7

Сравнение с данными ТДУ ГИЛ показывает, что в 79% случаев предложенный подход позволяет корректно определять класс возраста лесов с погрешностью, не превышающей 1 класс.



Эксперимент по применению UNET для оценки запаса

Модель UNET была применена к задаче оценки запаса, описанной ранее.

Язык: Python 3.11
Библиотеки: Tensorflow, Keras, scikit-learn

Архитектура модели:

Вход: К растрам применяется генератор масок (ApplyMask) и создается тензор признаков (Высота × Ширина × Канал)

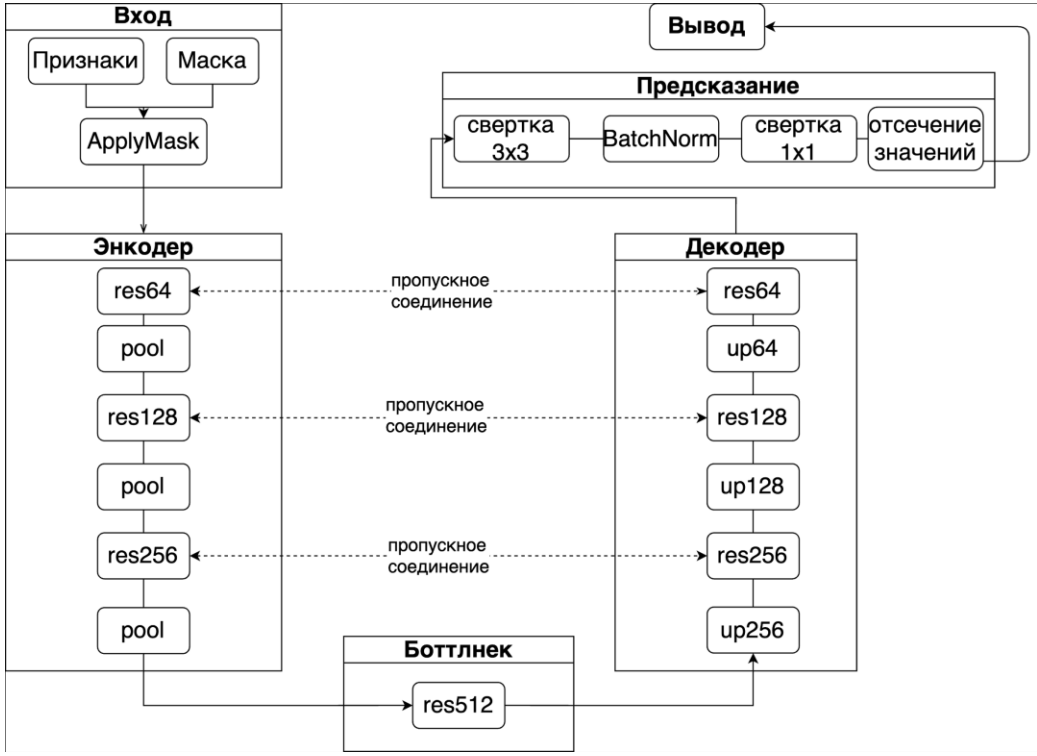
Энкодер: остоит из трёх последовательных остаточных блоков с увеличением числа каналов. На каждом уровне выход энкодера сохраняется для последующих пропускных соединений.

Боттлнек: состоит из остаточного блока res512, в котором достигается наибольшее число каналов и минимальное пространственное разрешение. Этот блок обрабатывает обобщённые признаки перед началом восстановления.

Декодер: Данные проходят через три этапа апсемплинга и остаточных блоков, модель восстанавливает пространственное разрешение

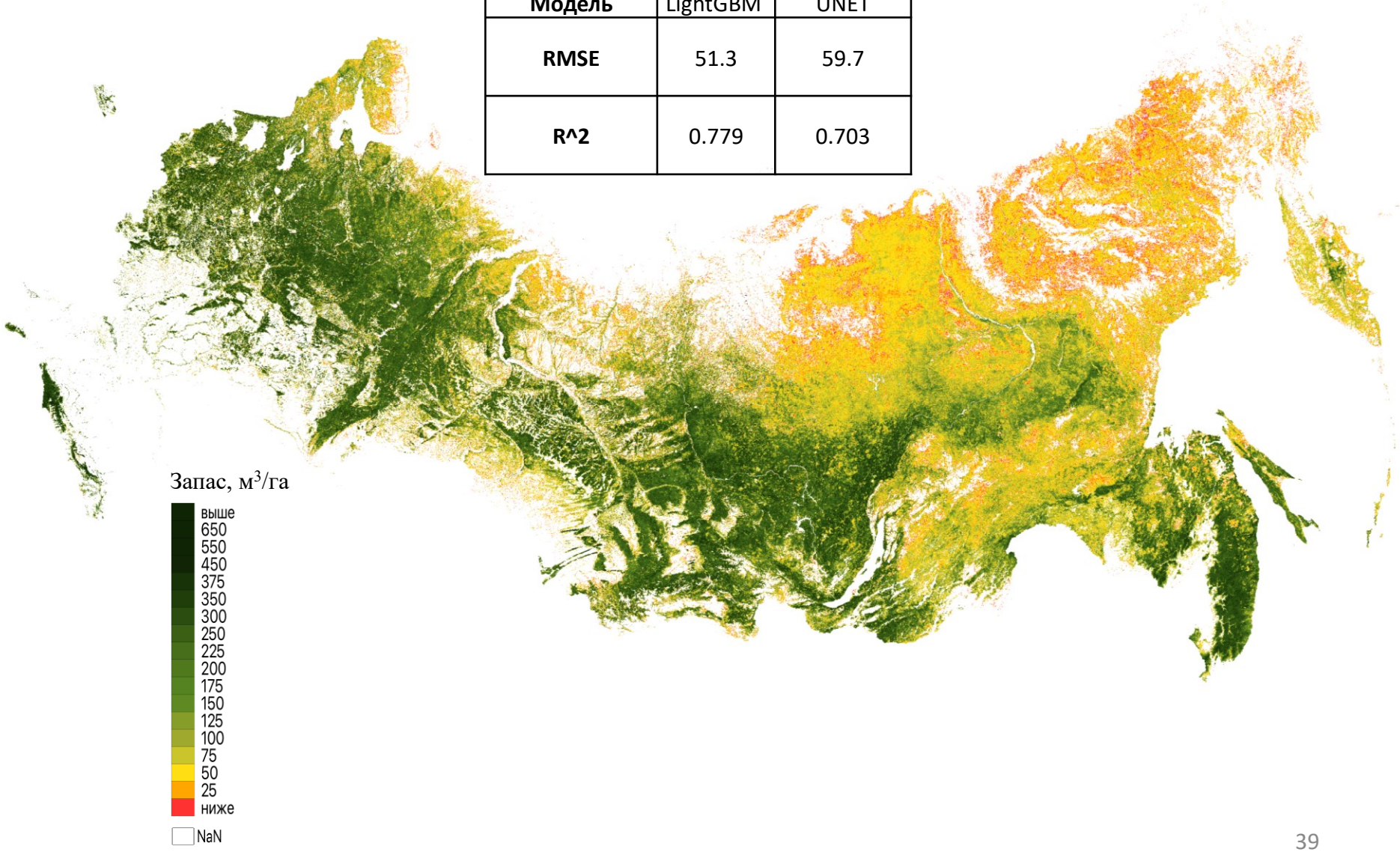
Предсказание: На выходе декодера применяется свёртка 3×3 → батч-нормализация → свёртка 1×1 (1 канал) → операция **отсечения значений**, приводящая значения пикселей в допустимый диапазон (от 0 до 1200), и формируется итоговое изображение

Компиляция: оптимизатор Adam



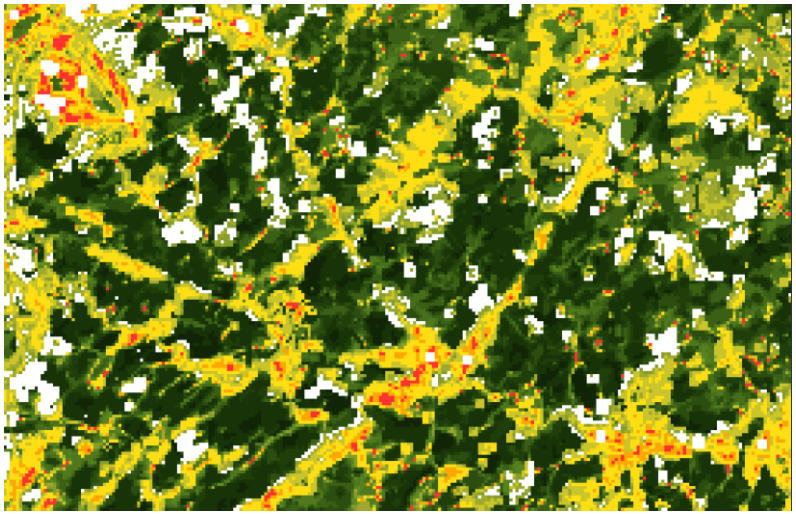
Результаты применения модели UNET для запаса

Модель	LightGBM	UNET
RMSE	51.3	59.7
R^2	0.779	0.703



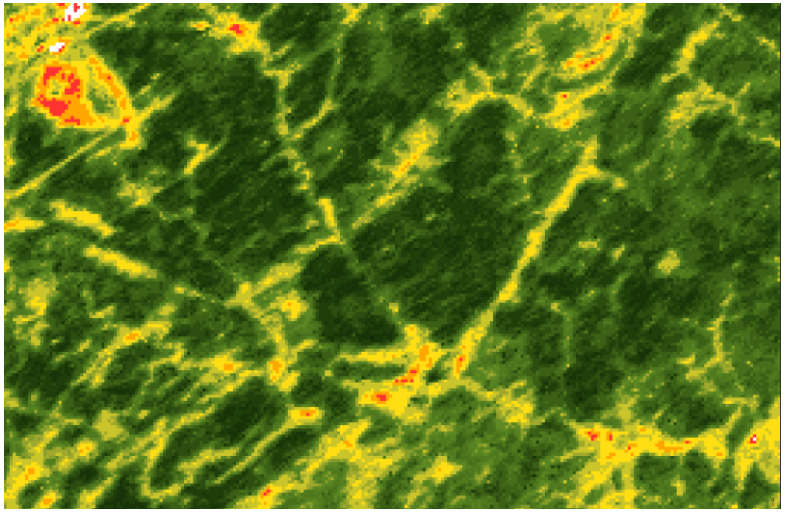
Результаты применения модели UNET для запаса

Исходная выборка

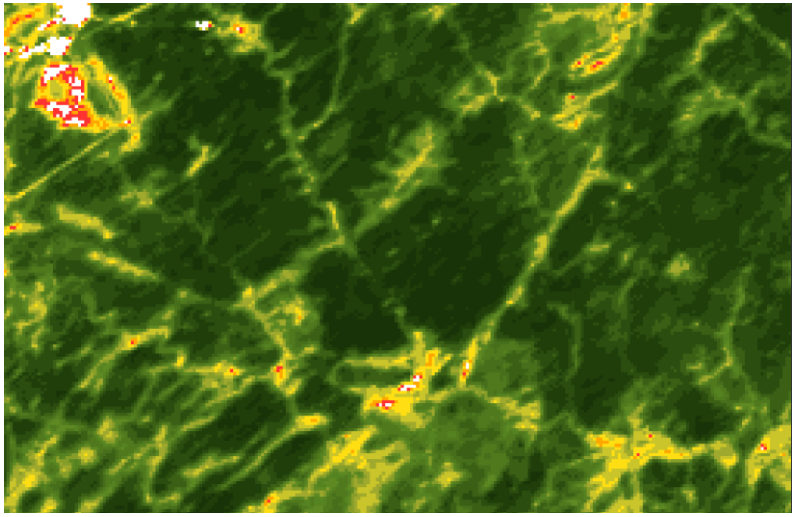


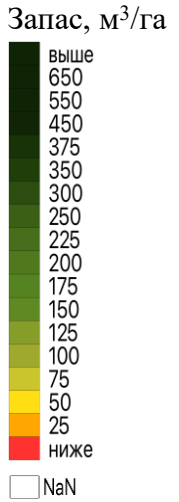
Видимый диапазон

LGBM



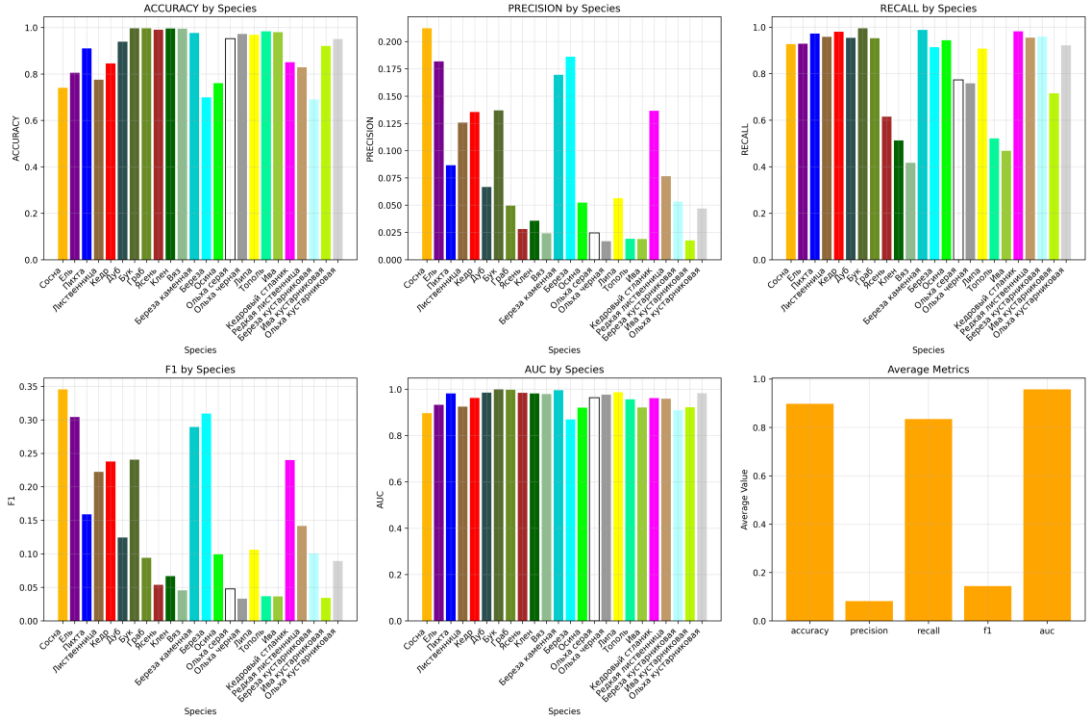
UNET

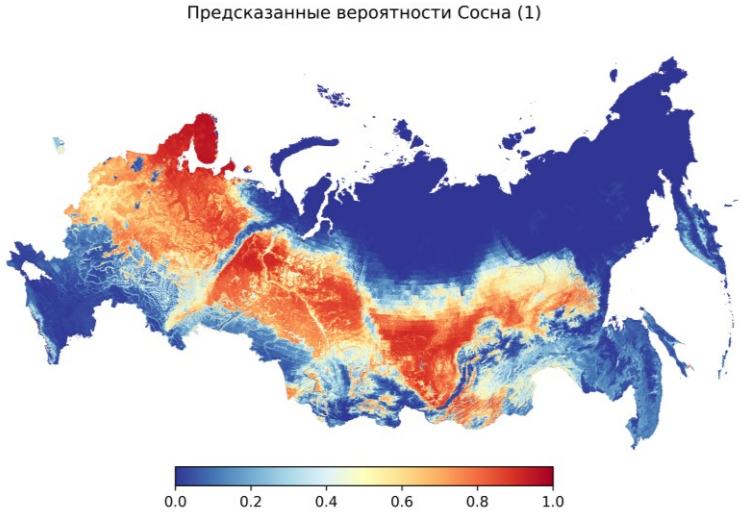


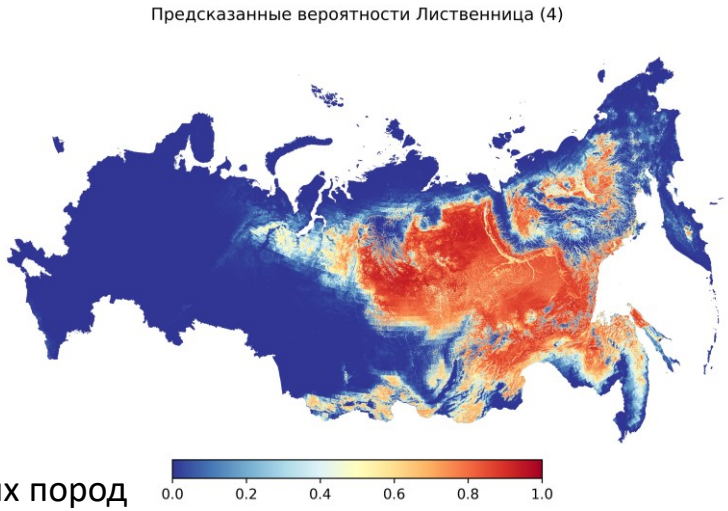


Результаты применения модели UNET для оценки ареалов распространения древесных пород

Первые результаты оценки ареалов пород согласуются их реальным распространением и показывают адекватную точность.







Больше про этот продукт можно узнать 12 ноября (среда) в 15:00 на 3 заседании секции F на докладе:

Моделирование ареалов потенциального распространения древесных пород на территории России с использованием методов машинного обучения
Михайлов Н.В., Барталев С.А.

Спасибо за внимание!

Основные этапы:

1. Постановка задачи – определить область, классы
2. Выбрать индикаторы-признаки
3. Собрать выборку для обучения/проверки точности
4. Обучить метод машинного обучения и построить карту
5. Оценить точность