ИНСТИТУТ КОСМИЧЕСКИХ ИССЛЕДОВАНИЙ РАН

Моделирование ареалов потенциального распространения древесных пород на территории России с использованием методов машинного обучения

Михайлов Н.В., Барталев С.А.
Институт Космических Исследований РАН

Работа выполнена в рамках темы «Мониторинг» (госрегистрация №1220042500031-8). Обработка данных ДЗЗ проводилась с использованием ресурсов Центра коллективного пользования «ИКИ-Мониторинг».

Карта преобладающих пород

Использовано 92 признака (данные MODIS):

- 5-ти дневные композиты из 3 спектральных (b01, b02, b07) каналов за период, полученные путем интерполяции и фильтрации, разработанные в ИКИ РАН;
- зимние композиты;

Карты априорных вероятностей для каждой породы по данным на уровне лесничеств (наличие породы в пределах лесничества)

Выборка, отфильтрованная по ежегодным изменениям за 2000-2024г.

Метод LAGMA*

*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.

Цель проведения эксперимента – опробовать методику машинного обучения и сверточных нейронных сетей для воссоздания ареалов распространения пород.

Данные

Разрешение 0,5°

BIOCLIM, осредненный за 20 лет:

BIO1 - среднегодовая температура, °C;
BIO2 - среднесуточная амплитуда, °C (среднее значение (макс. темп. - мин. темп.) по месяцам);
BIO3 - изотермальность (BIO2/BIO7) (×100), %;
BIO4 - температурная сезонность (стандартное отклонение ×100), %;
BIO5 - максимальная температура наиболее теплого месяца, °C;
BIO6 - минимальная температура наиболее холодного месяца, °C;
BIO7 - годовая амплитуда температур (BIO5-BIO6), °C;
BIO8 - средняя температура наиболее влажного квартала, °C;
BIO9 - средняя температура наиболее сухого квартала, °C;
BIO10 - средняя температура наиболее теплого квартала, °C;
BIO11 - средняя температура наиболее холодного квартала, °C;
BIO12 - годовое количество осадков, мм;
BIO13 - сумма осадков наиболее влажного месяца, мм;
BIO14 – сумма осадков наиболее сухого месяца, мм;
BIO15 - сезонность осадков (коэффициент вариации), %;
BIO16 - осадки наиболее влажного квартала, мм;
BIO17 - осадки наиболее сухого квартала, мм;
BIO18 - осадки наиболее теплого квартала, мм;
BIO19 - осадки наиболее холодного квартала, мм;

Разрешение 230м

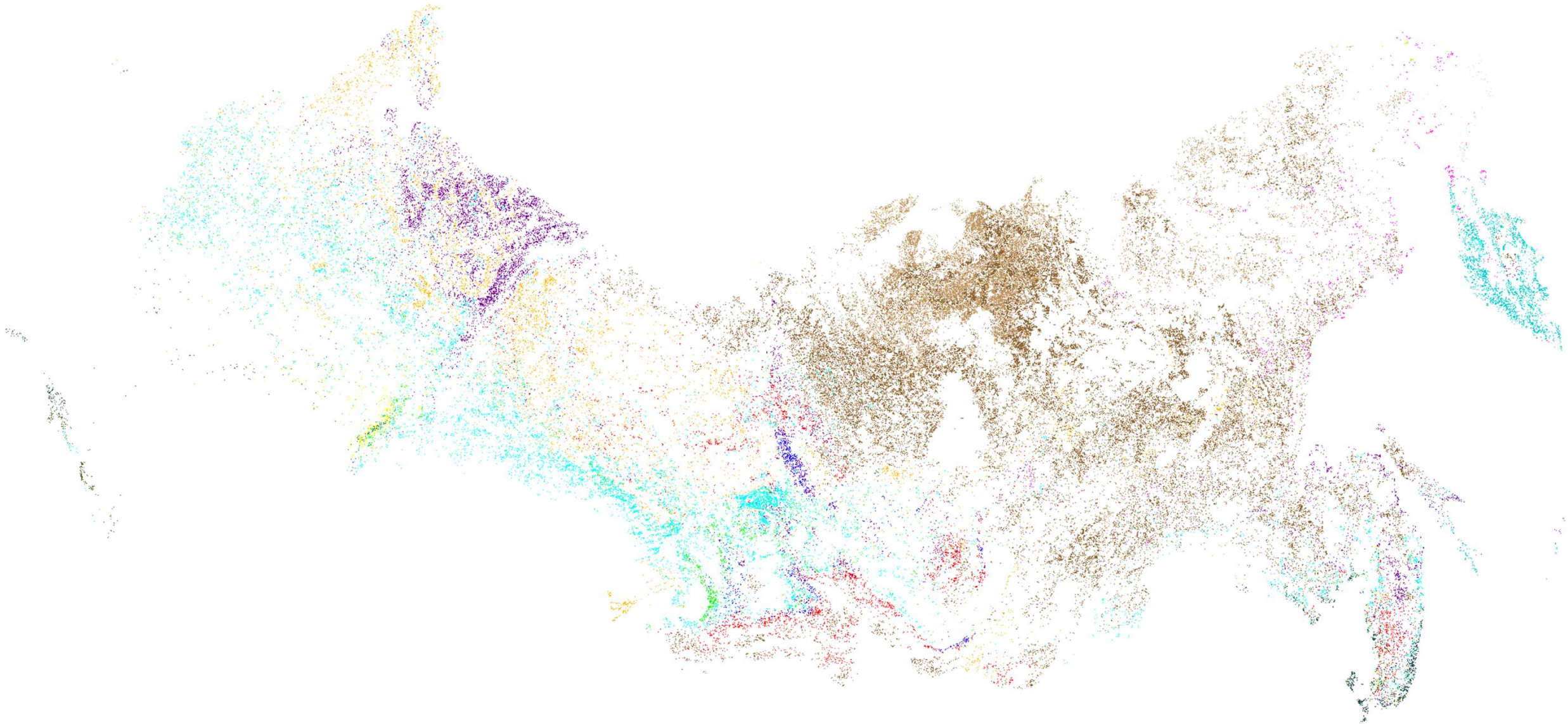
Цифровая модель рельефа (SRTM):
абс. Высота (м) + углы наклона (°)

Карта почв масштаба 2 500 000

Источник данных о преобладающей породе - Актуализированная Цифровая Основа Государственной Инвентаризации Лесов (АЦО ГИЛ).

Результат - обученная модель, позволяющая воссоздавать ареалы распространения каждой породы в виде вероятностных карт, отражающих биоклиматическую пригодность условий произрастания.

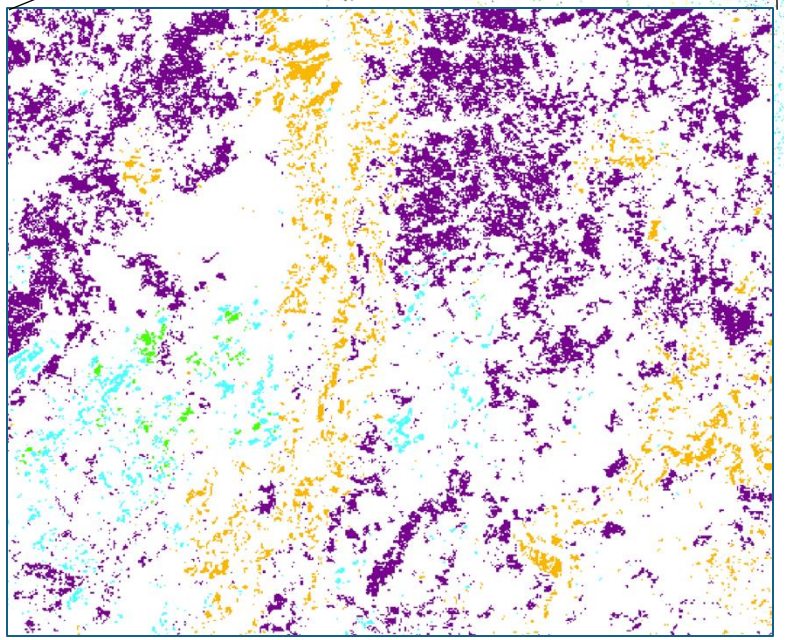
Источник данных для выборки - Актуализированная Цифровая Основа Государственной Инвентаризации Лесов (АЦО ГИЛ)

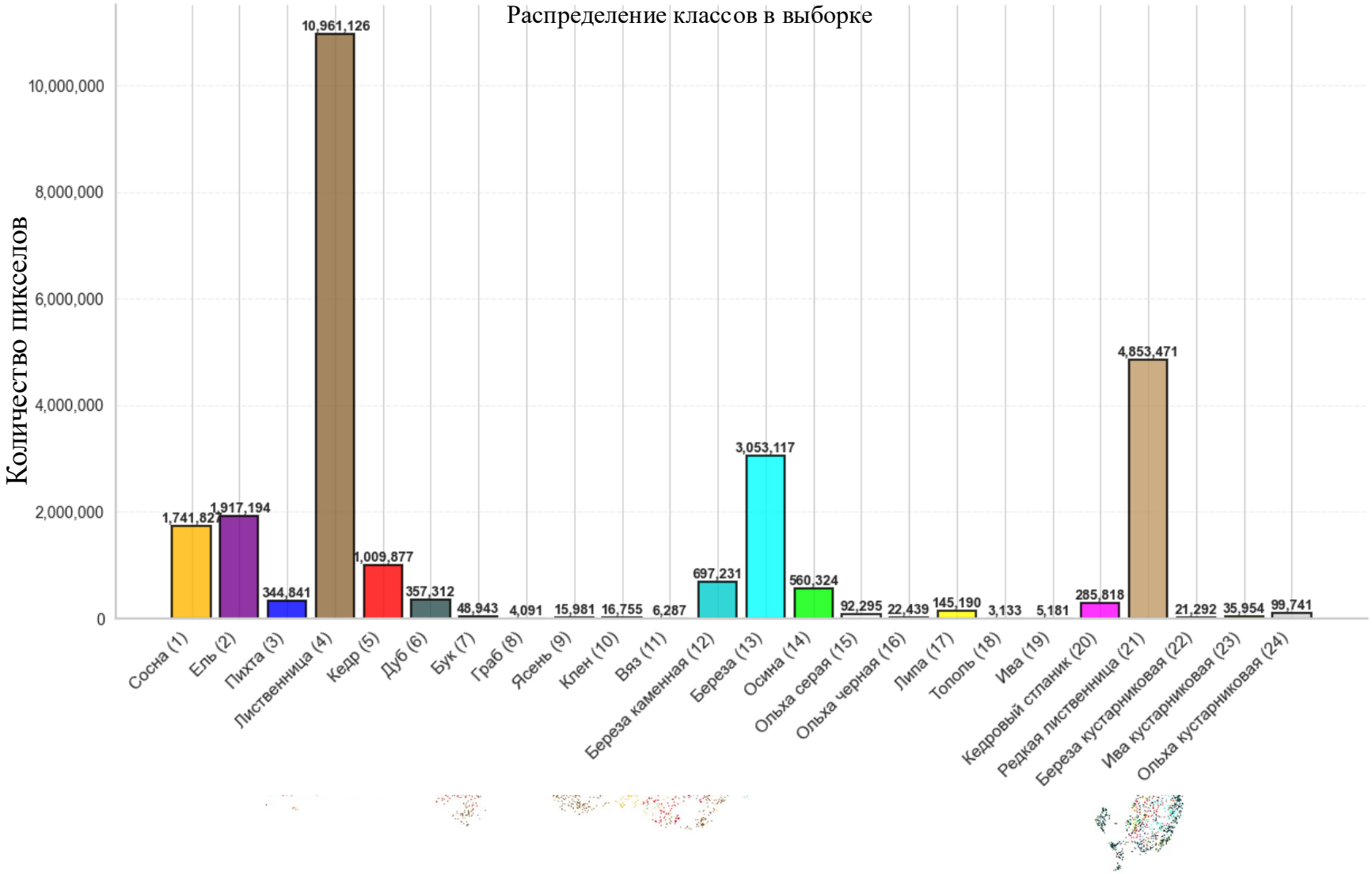


Источник данных для выборки:- Актуализированная Цифровая Основа Государственной Инвентаризации Лесов (АЦО ГИЛ)

Особенности выборки:

- сильный дисбаланс классов пород;
- разреженность разметки;
- только преобладающая порода;
- факт присутствия -1, **отсутствие/неизвестно – 0*;**



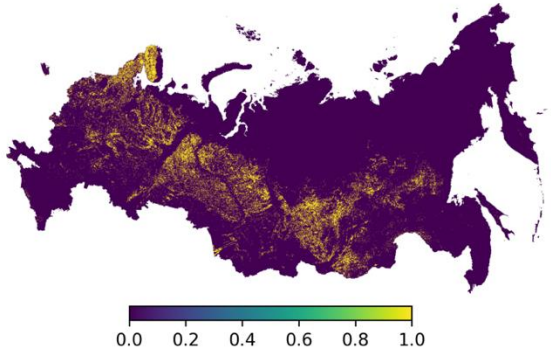


*Wang Y., Samarasekara C. L., Stone L. A machine learning method for estimating the probability of presence using presence-background data //Ecology and Evolution. – 2022. – Т. 12. – №. 6. – С. e8998.

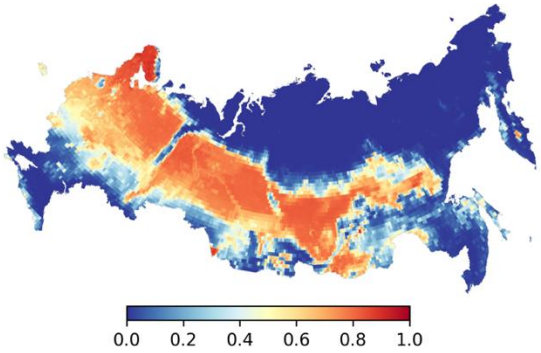
Классические ML-методы

Сосна (1)

Повыделенные данные



Random Forest

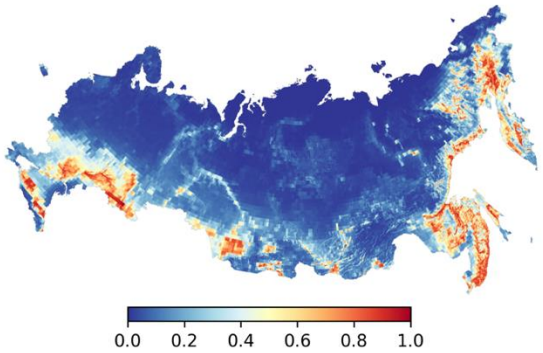


Априорные вероятности (текущие)

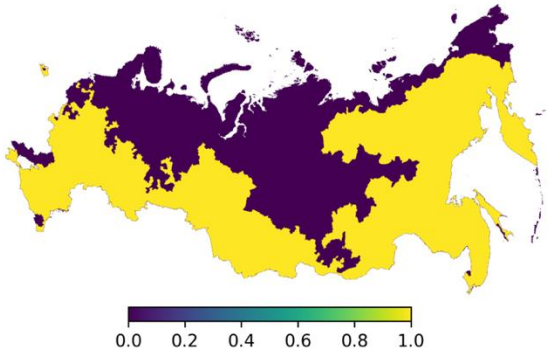
Тополь (18)

Повыделенные данные

Random Forest

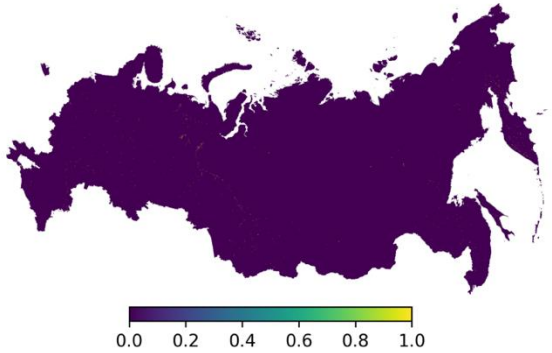


Априорные вероятности (текущие)

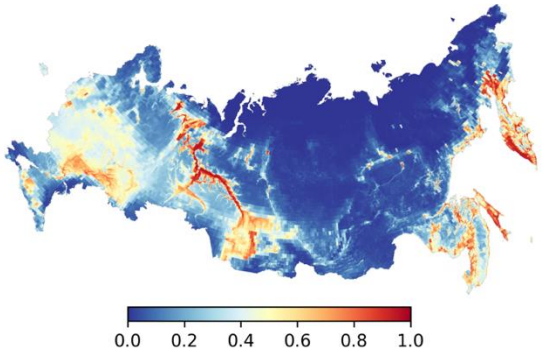


Ива (19)

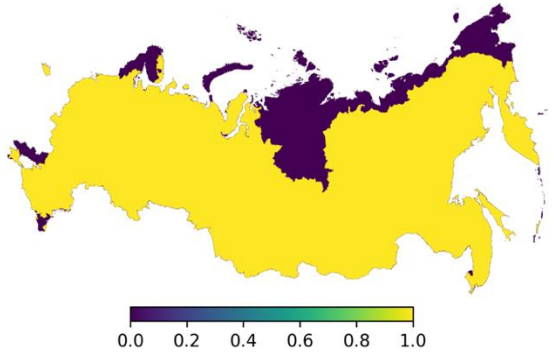
Повыделенные данные



Random Forest



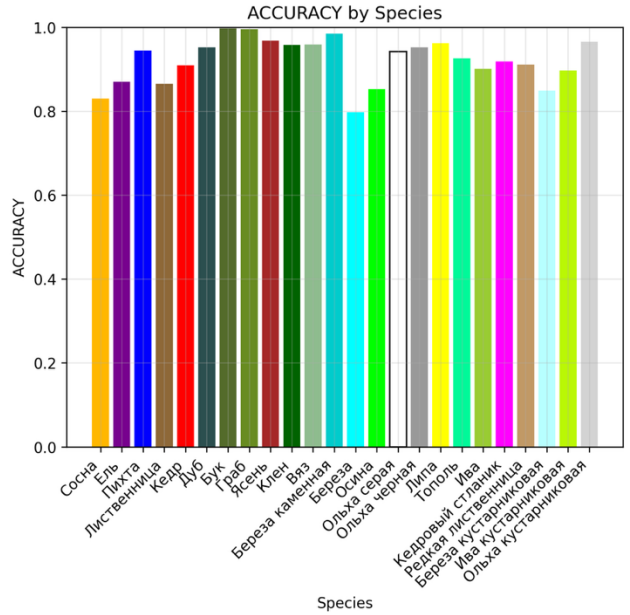
Априорные вероятности (текущие)

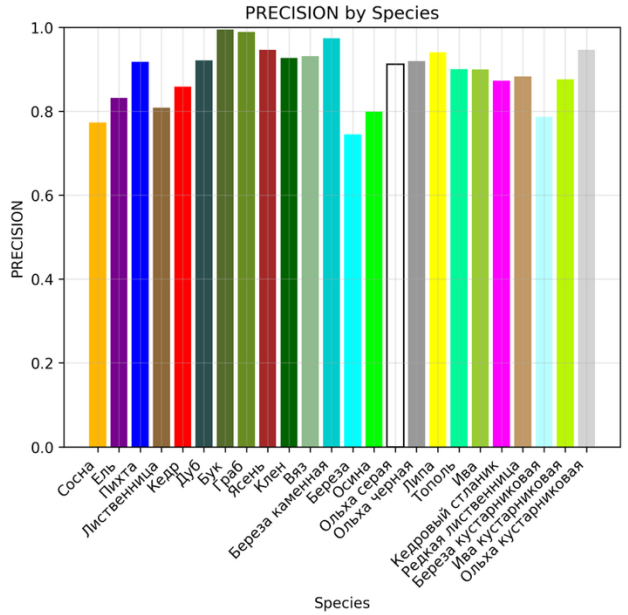


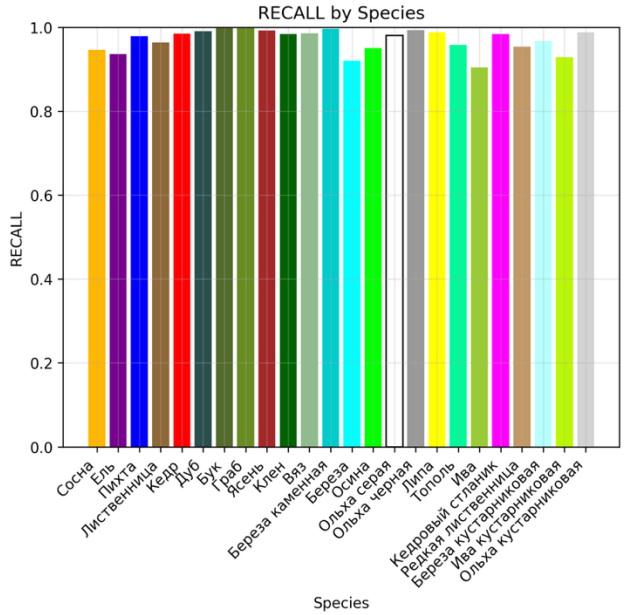
Особенности:

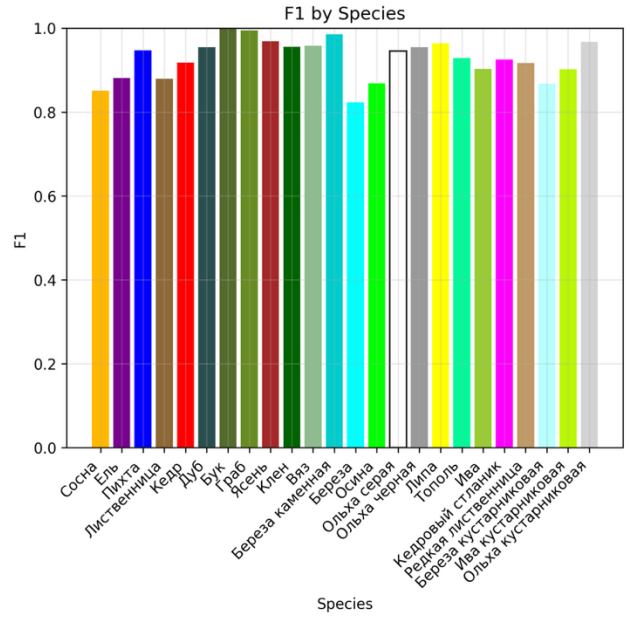
- одна независимая модель на породу (24 независимые модели) → требуется работа с признаками;
- требуется балансировка выборки (по минимальному классу);
- часть моделей опираются на один и тот же набор признаков;
- отсутствие контекста соседних пикселов → резкие границы и одиночные пикселы (шум);

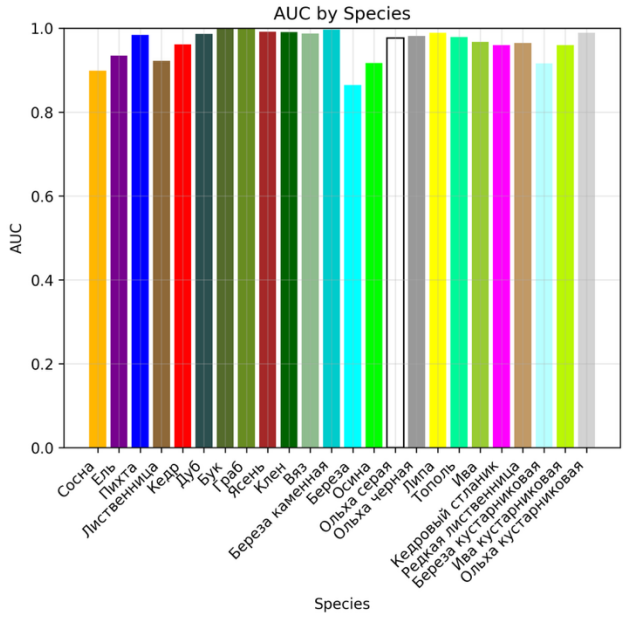
Метрики. Random Forest

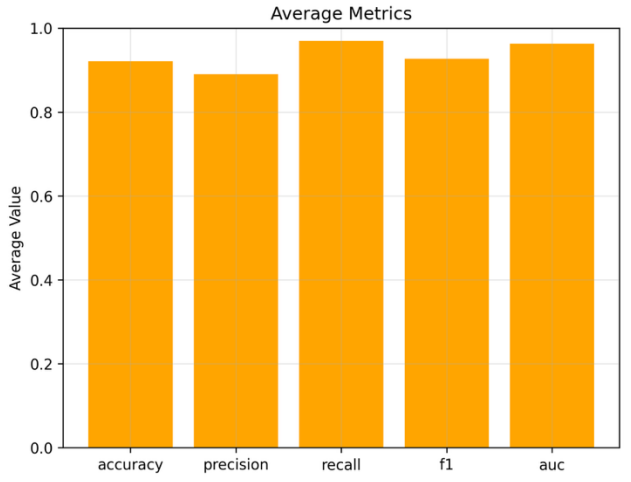












Accuracy (точность классификации): доля **всех правильно** классифицированных пикселов;

Precision (точность класса): из всех предсказанных моделью присутствий, какая доля **истинные** присутствия;

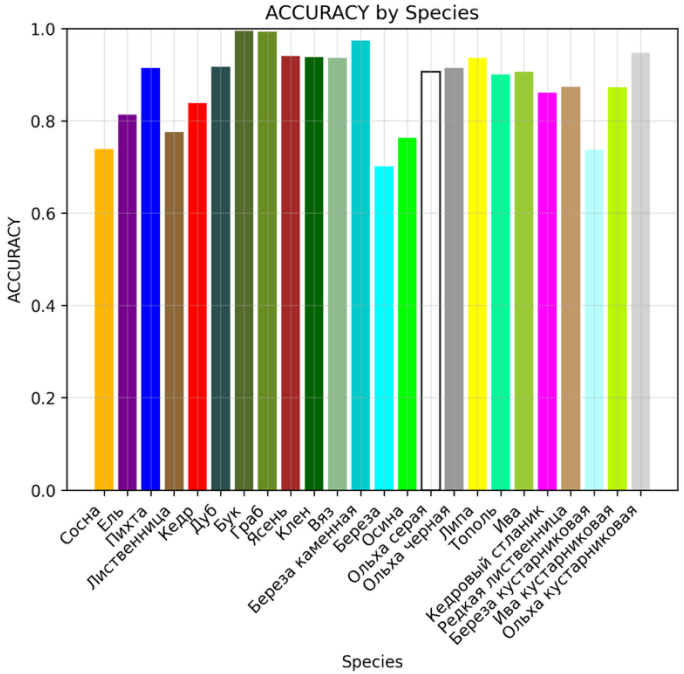
Recall (полнота): из всех реальных присутствий, какую долю модель нашла;

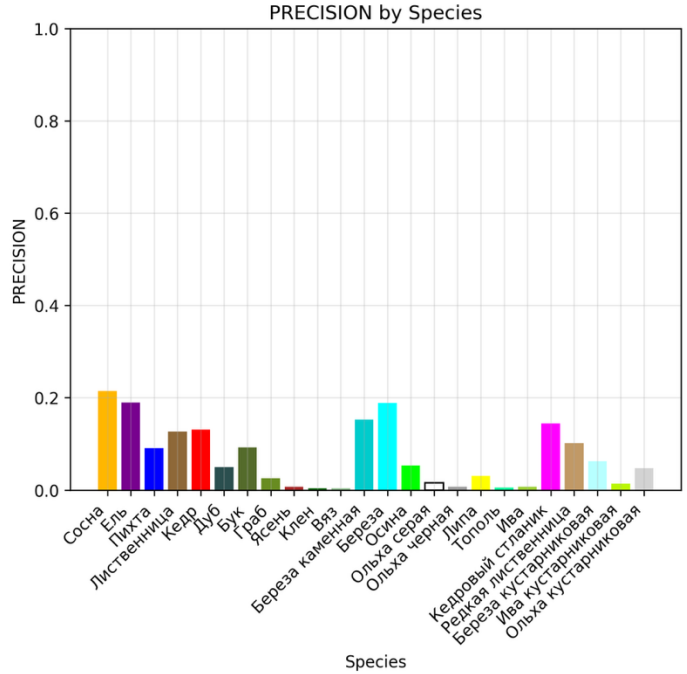
F1-score: гармоническое среднее precision и recall;

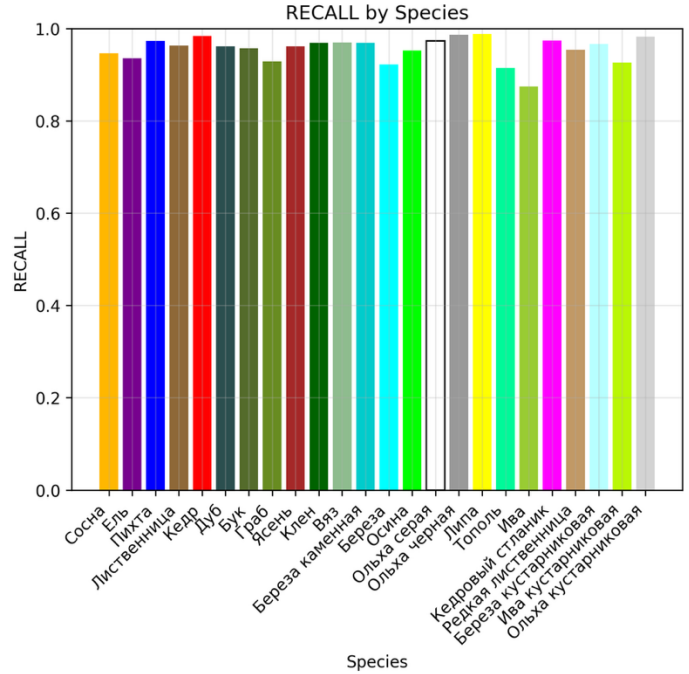
ROC-AUC (качество ранжирования): вероятность, что случайному присутствию модель даст большую вероятность, чем случайному отсутствию;

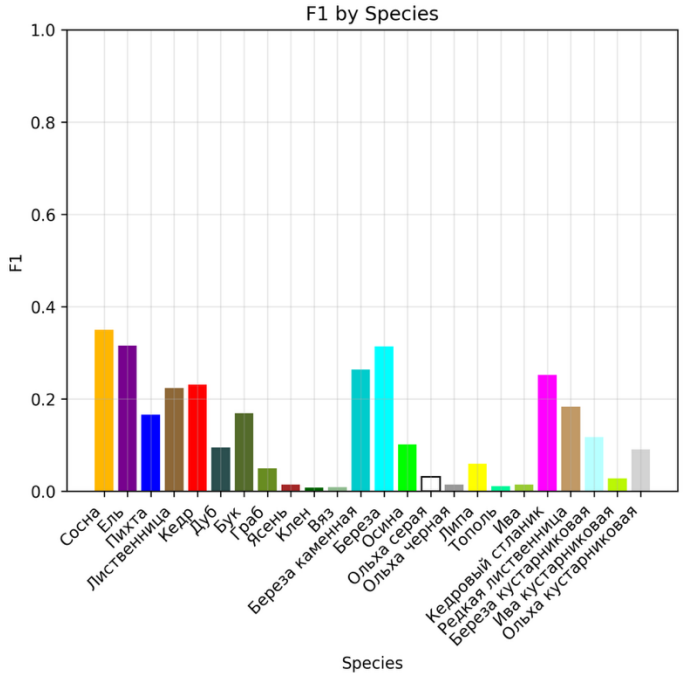
Метрики показаны по валидационной выборке

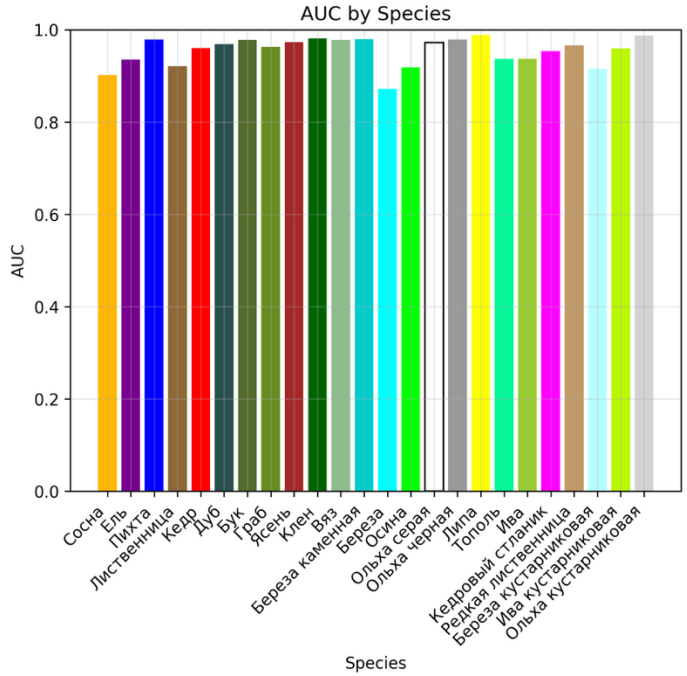
Результаты RF

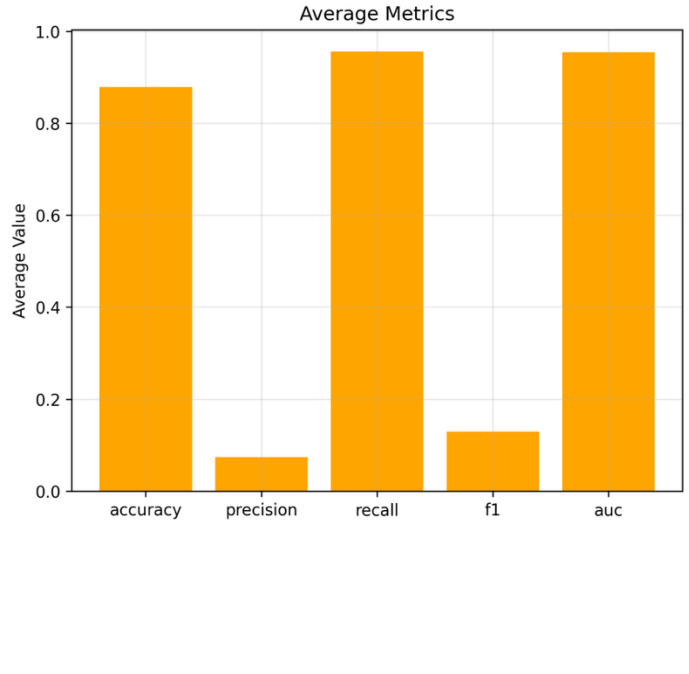


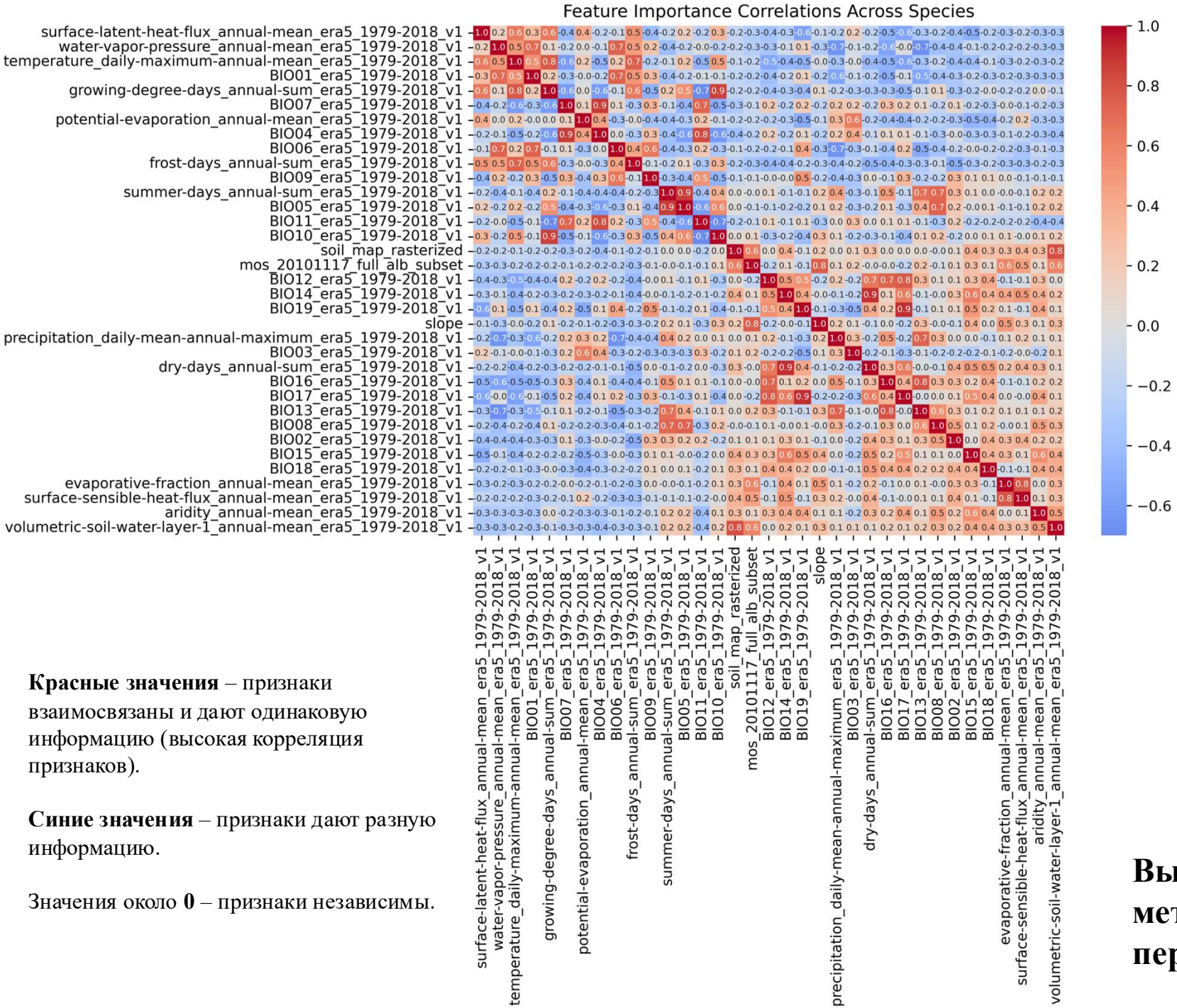












Красные значения – признаки взаимосвязаны и дают одинаковую информацию (высокая корреляция признаков).

Синие значения – признаки дают разную информацию.

Значения около **0** – признаки независимы.

Особенности ML:

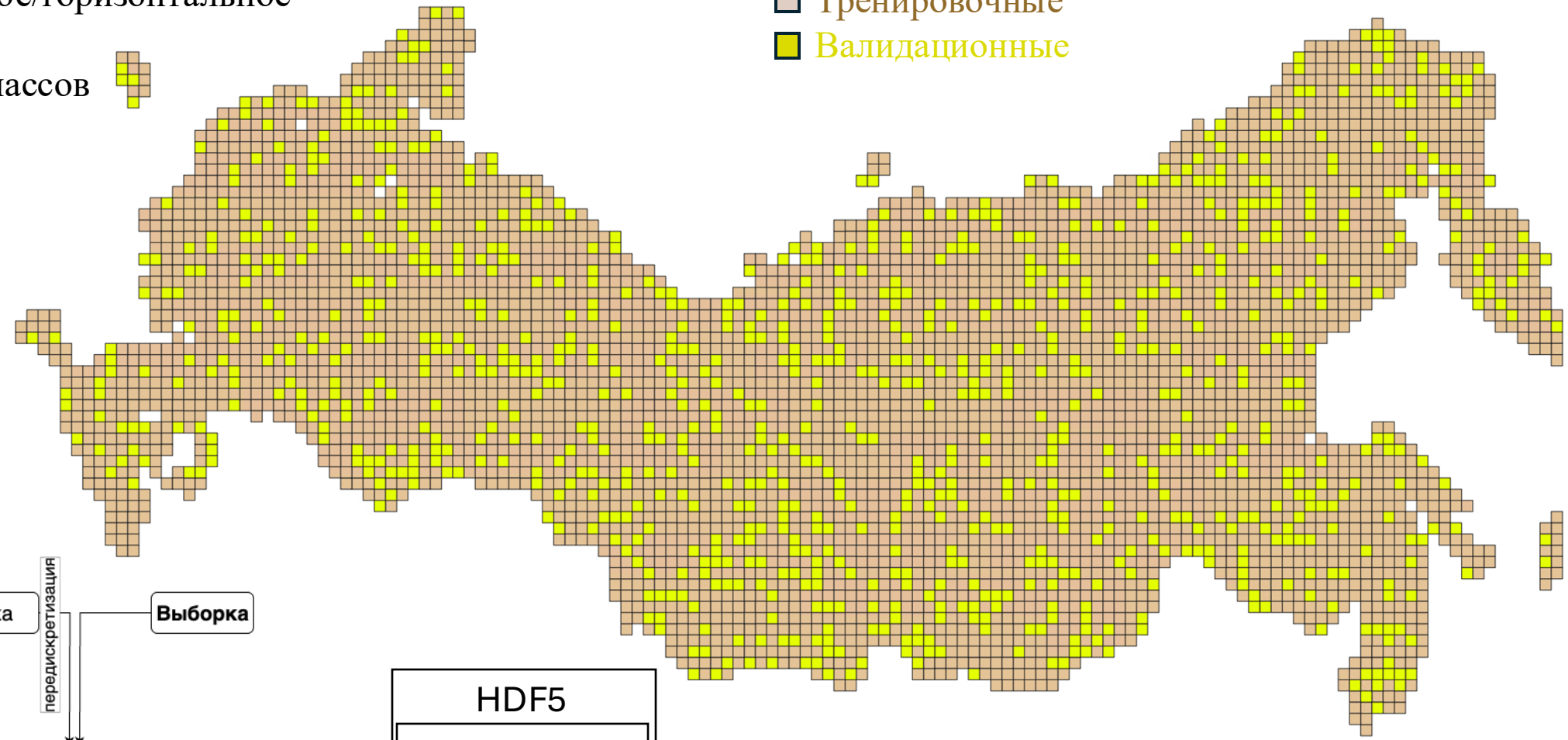
- высокая положительная корреляция важностей по породам - признак избыточности → требуется прореживание для улучшения точности и интерпретируемости, упрощения вычислений для каждой модели;
- **сильная вероятность переобучения** из-за балансировки классов;
- у каждой модели свой набор признаков;
- **определение доминирующих признаков каждой модели;**

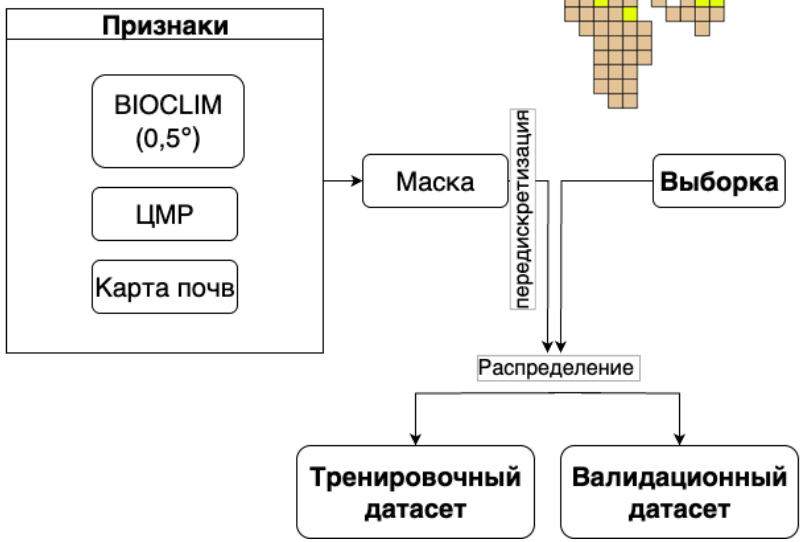
Высокие значения - косвенный признак, метрики могут быть завышены из-за переобучения

Подготовка данных для сверточной модели

- аугментация патчей (повороты, вертикальное/горизонтальное отражение);
- добавление весов классов в функцию потерь;

Патчи (256x256 пикс. MODIS (230м)):





Модель UNET

Язык: Python 3.12
Библиотеки: Tensorflow, Keras, scikit-learn

Архитектура модели:

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

Энкодер: состоит из трёх последовательных остаточных блоков с увеличением числа каналов. На каждом уровне выход энкодера сохраняется для последующих пропускных соединений, которые обеспечивают сохранение деталей.

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 (повышения разрешения) и остаточных блоков, модель восстанавливает пространственное разрешение

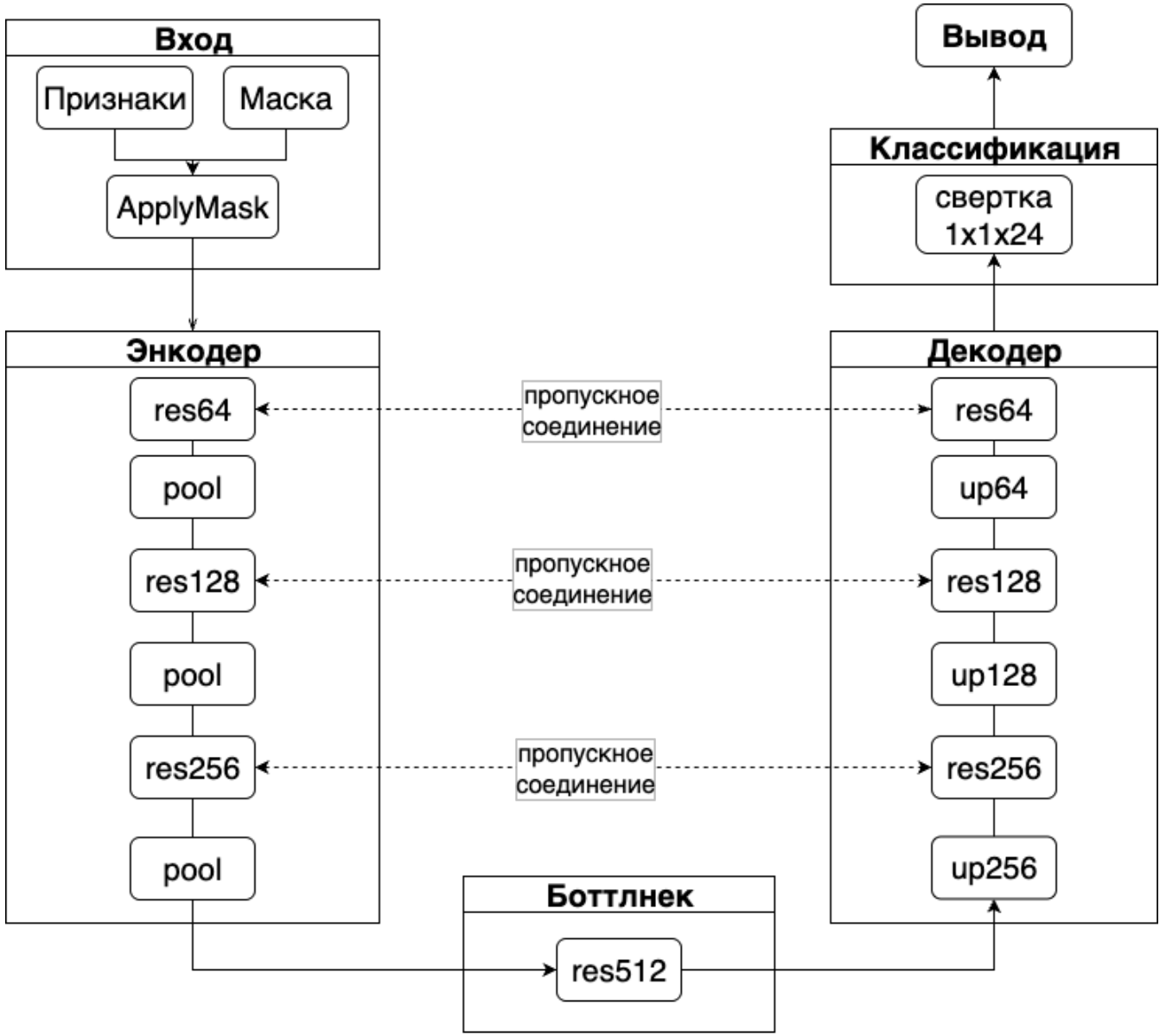
Предсказание: На выходе декодера применяется свёртка 3×3 → батч-нормализация → свёртка 1×1 (1 канал) из которой формируется итоговое изображение

Функция потерь: Binary Cross-Entropy*
Для каждого из 24 классов вычисляется отдельная потеря, затем берётся среднее. Каждый класс обрабатывается как независимая бинарная сегментация.

$$L_{class} = -\left[y \cdot pos_{weight} \cdot log\left(\sigma(x)\right) + (1-y) \cdot log\left(1-\sigma(x)\right)\right]$$

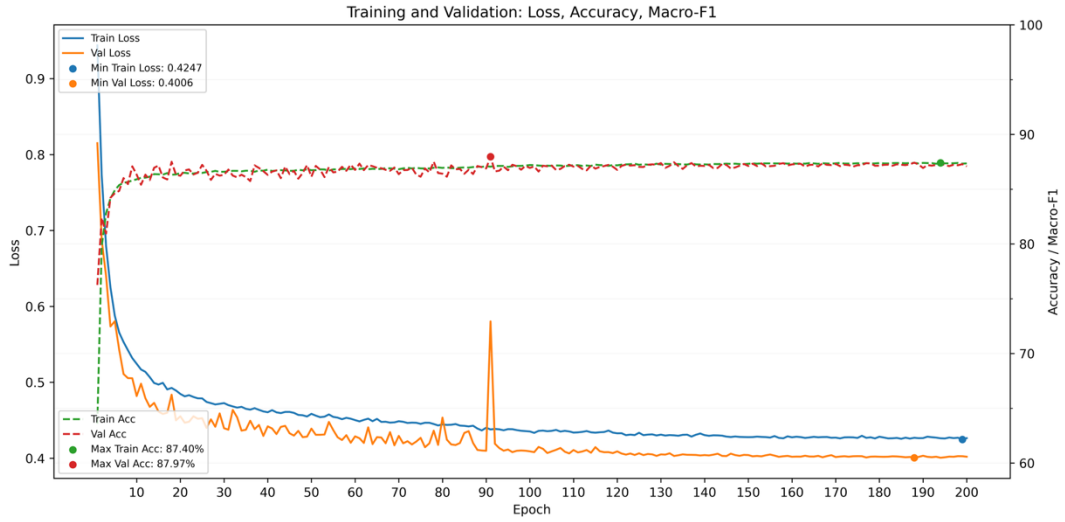
$$L_{total} = \left(\frac{1}{N}\right) \cdot \Sigma(i = 1\ to\ N)\left[L_{class_i}\right]$$

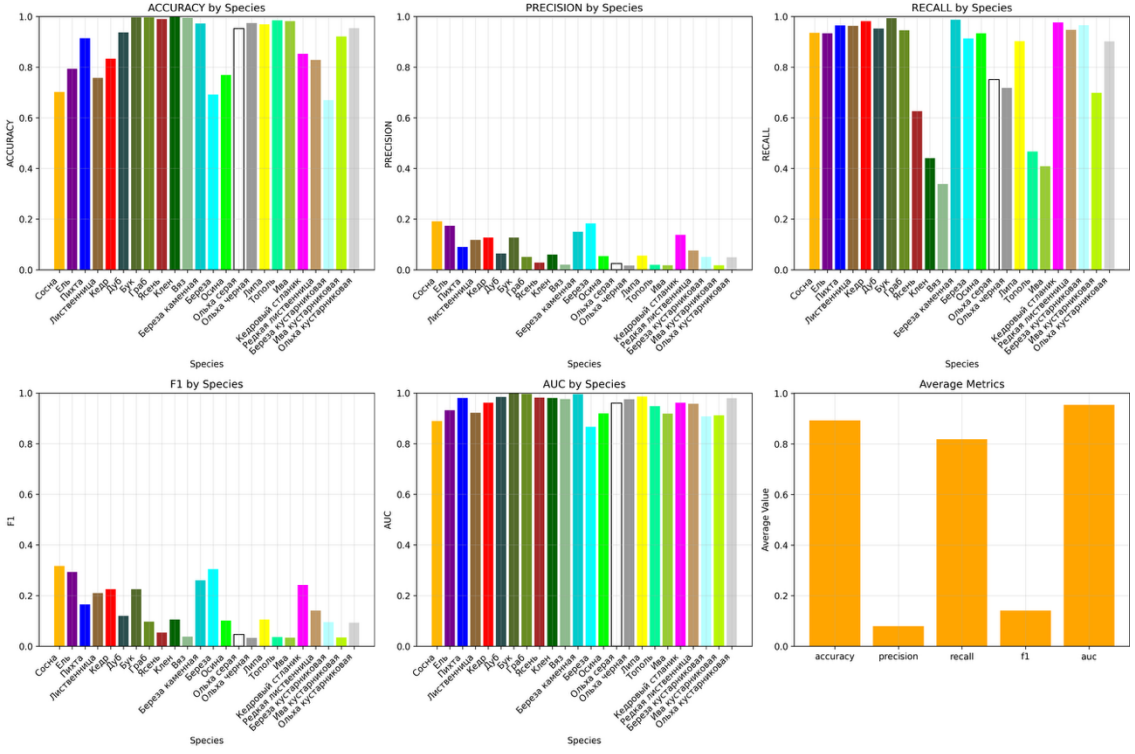
где:
$-\ \sigma(x_i) = \frac{1}{1+e^{-x_i}} -$ сигмоида;
$-\ pos_{weight_i} -$ вес положительного класса для балансировки;
$-\ y_i -$ целевое значение (0 или 1);
$-\ x\_i\ -\ logits$ для класса i (выход модели до сигмоиды);

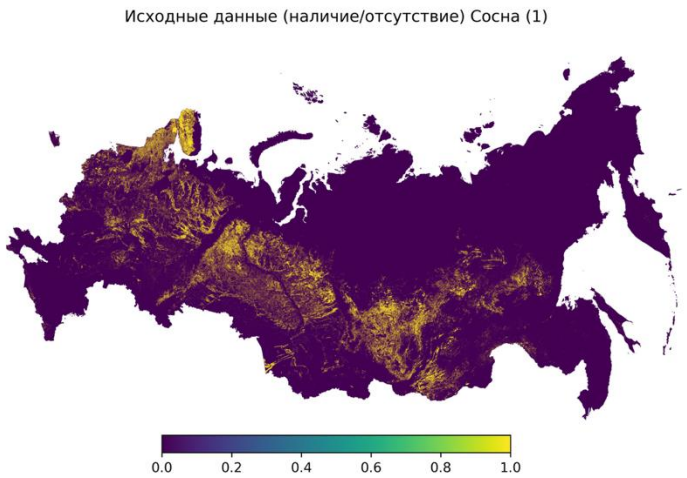


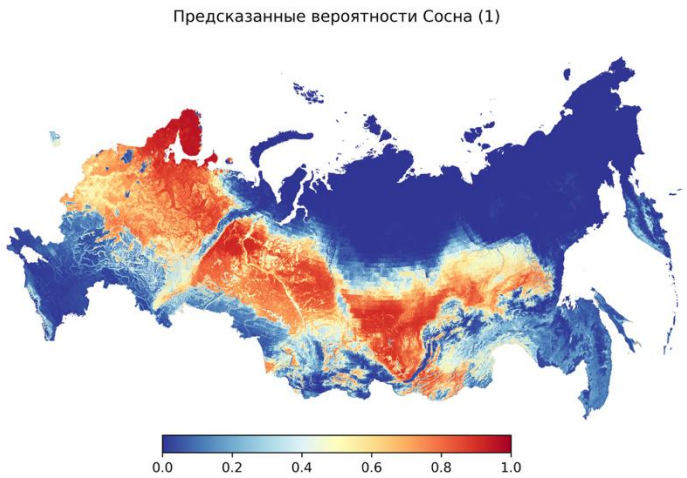
*https://docs.pytorch.org/docs/stable/generated/torch.nn.BCELoss.html

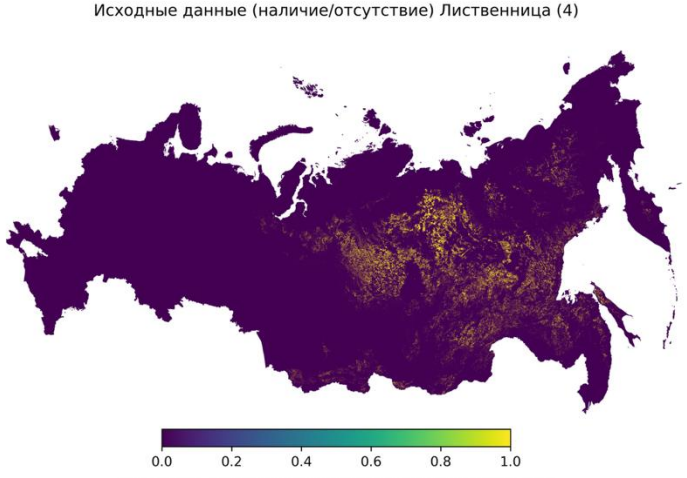
Результаты обучения UNET

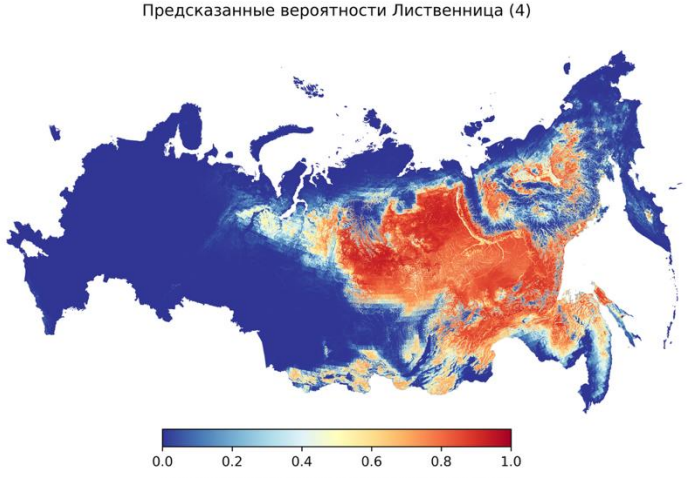


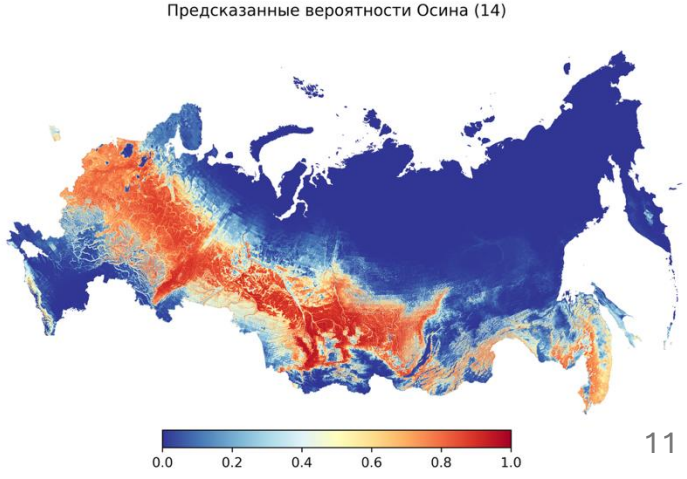






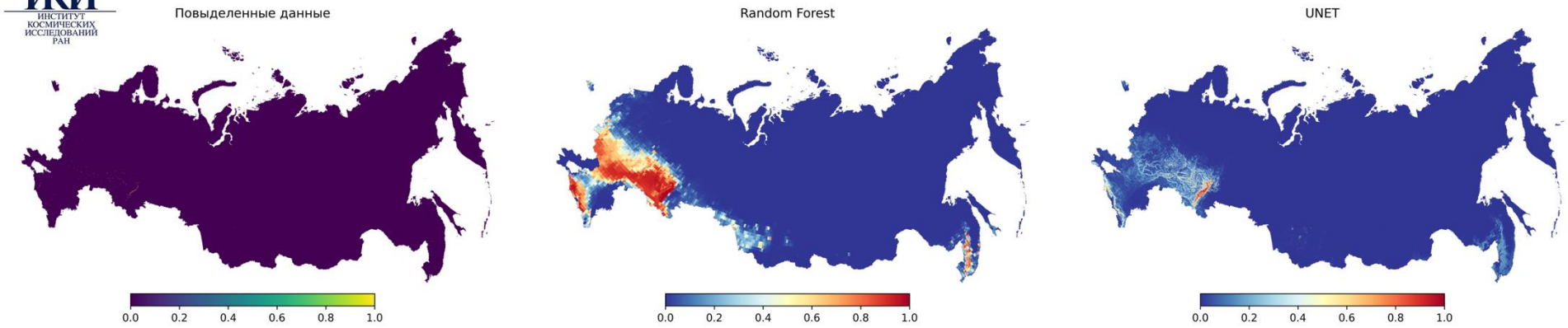


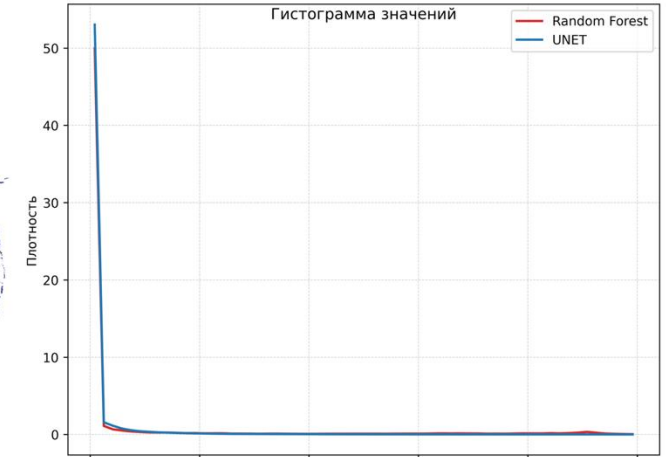




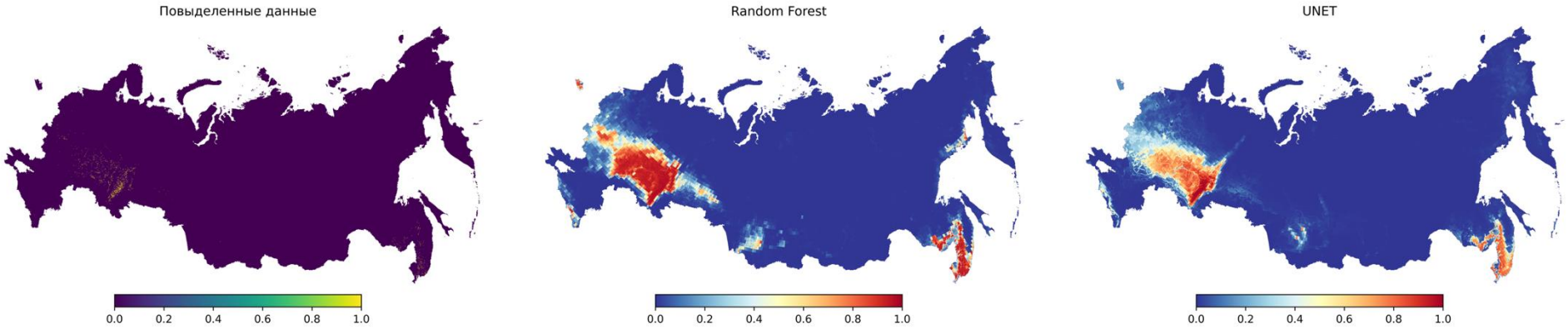
ИКИ
ИНСТИТУТ КОСМИЧЕСКИХ ИССЛЕДОВАНИЙ РАН

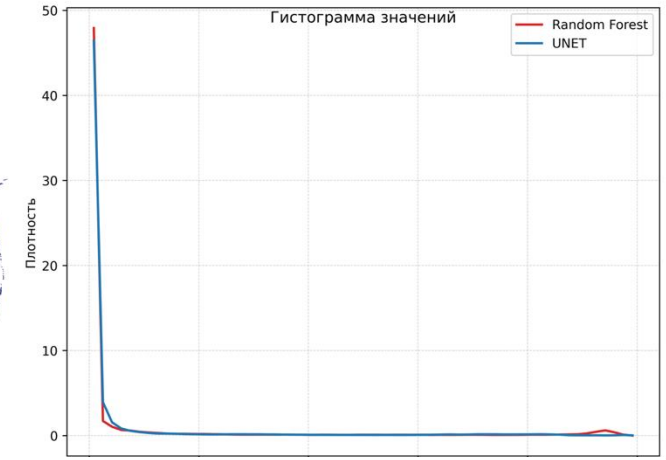
Клен



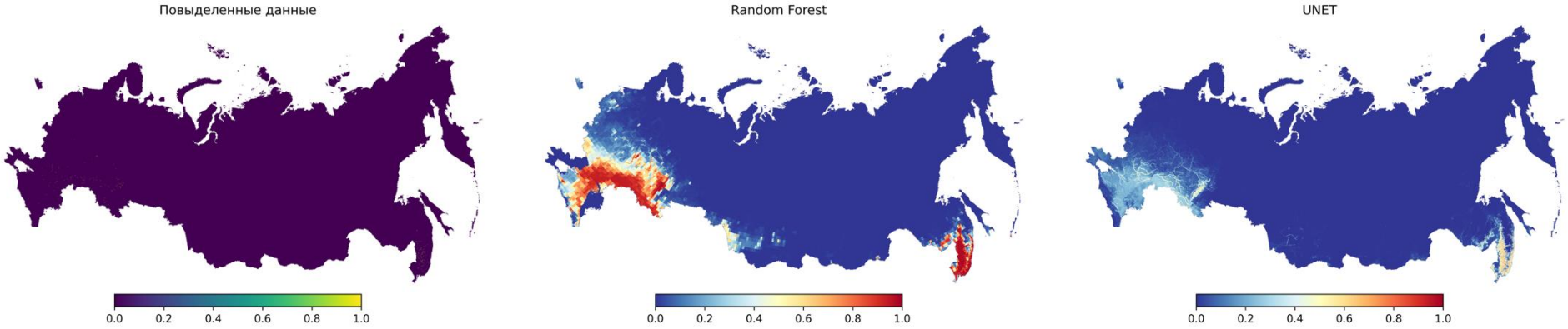


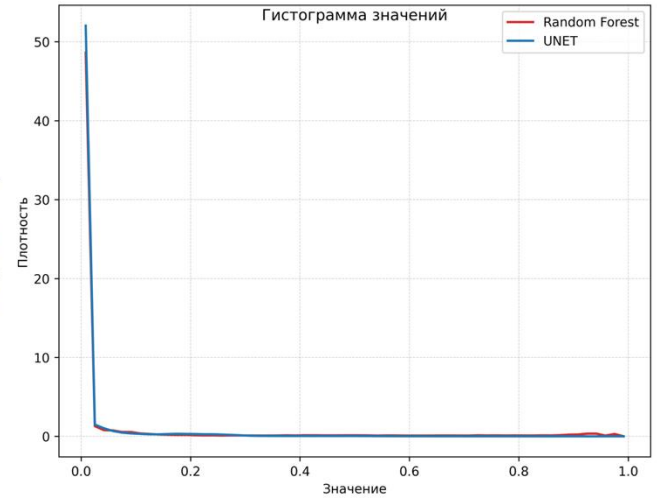
Липа





Вяз

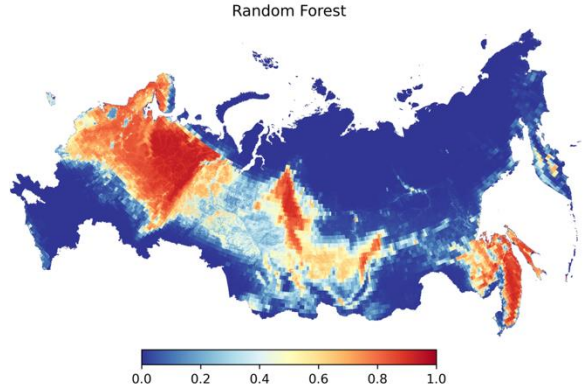


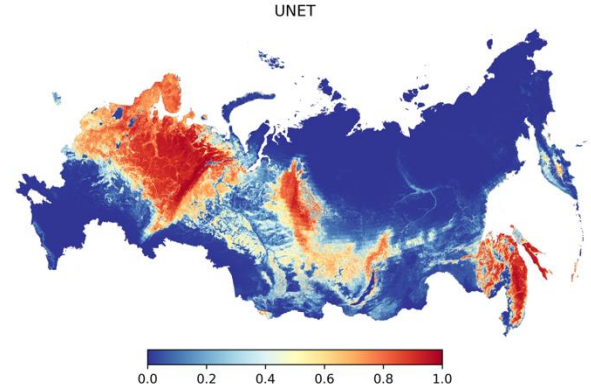


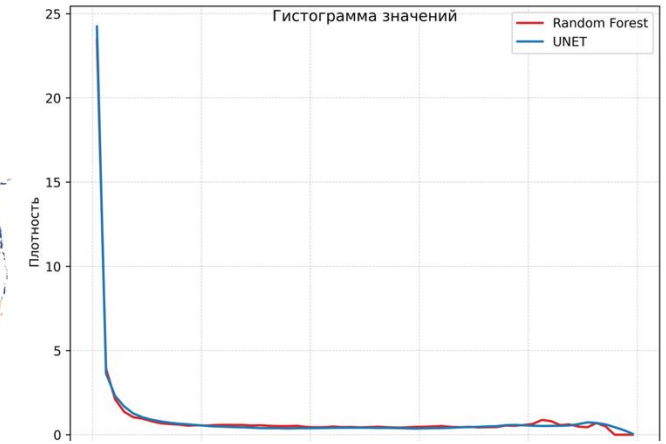
Результаты. Сложные примеры

ИКИ
ИНСТИТУТ КОСМИЧЕСКИХ ИССЛЕДОВАНИЙ РАН

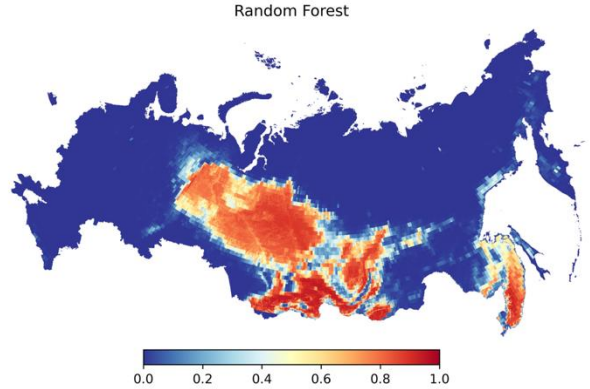
Ель

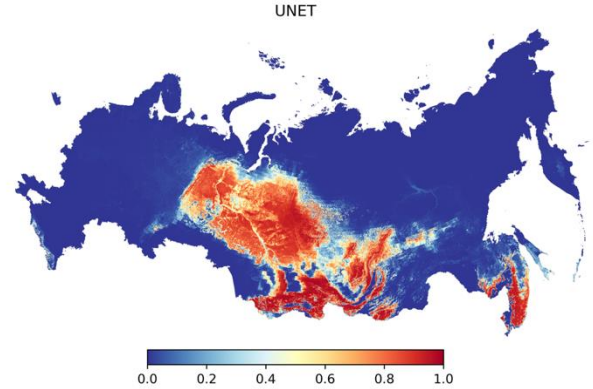


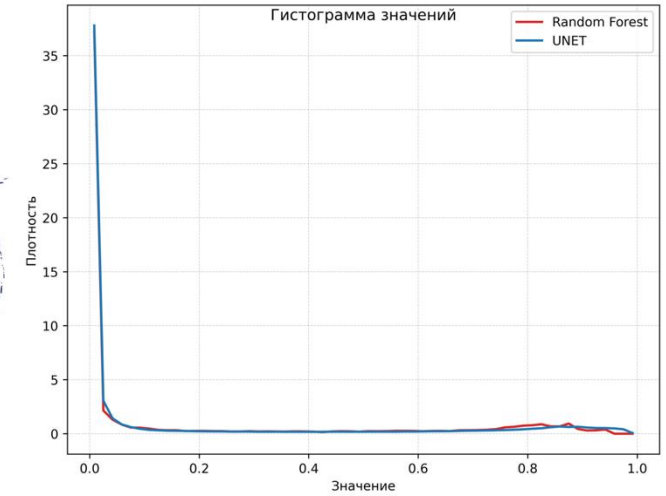




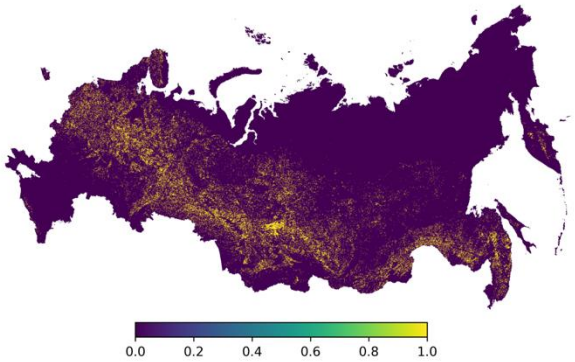
Кедр

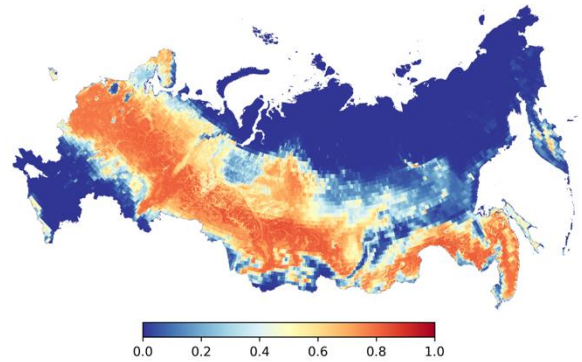


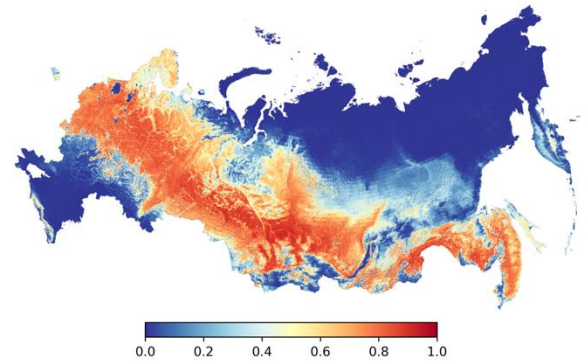


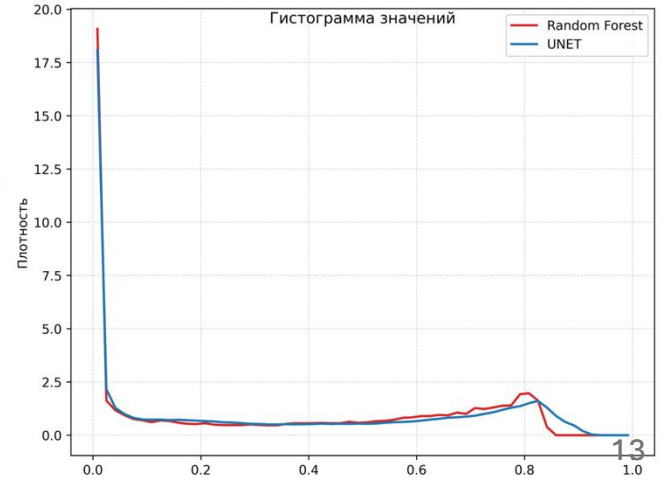


Береза



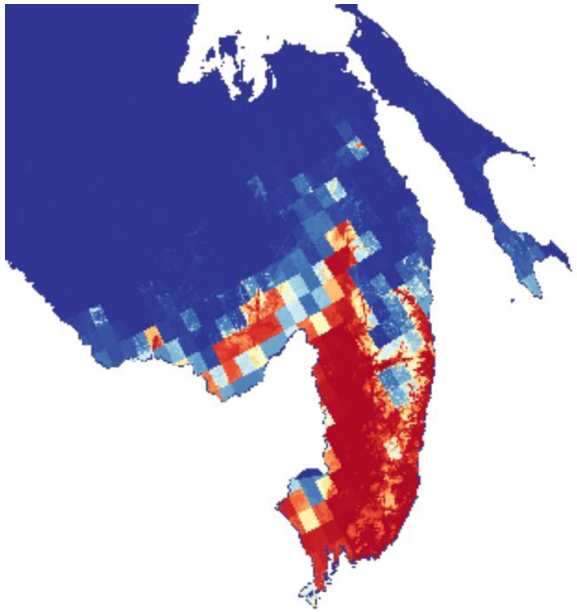
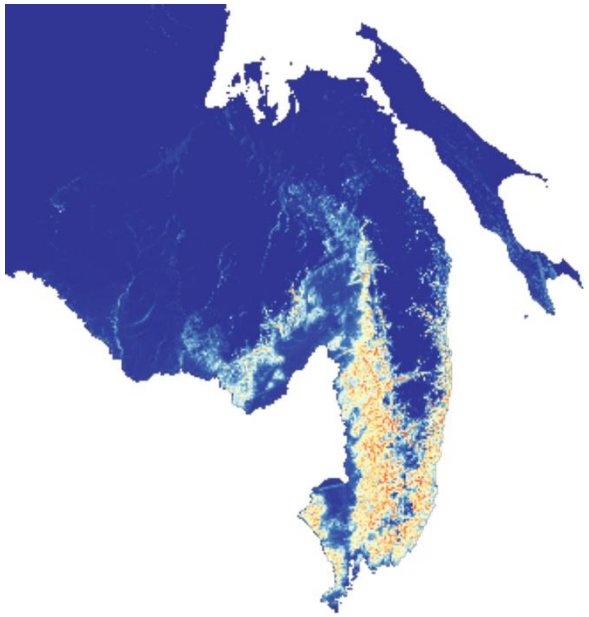
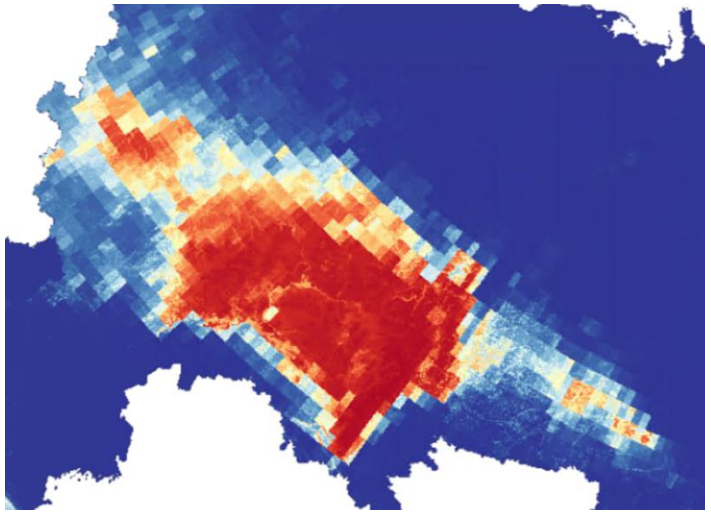
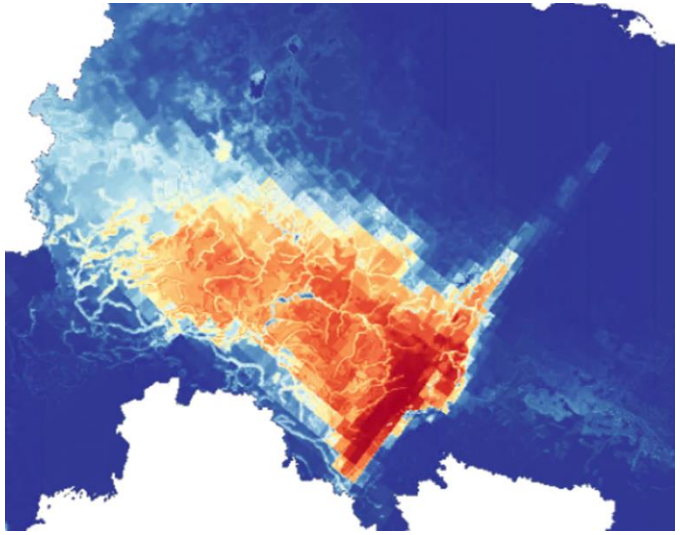




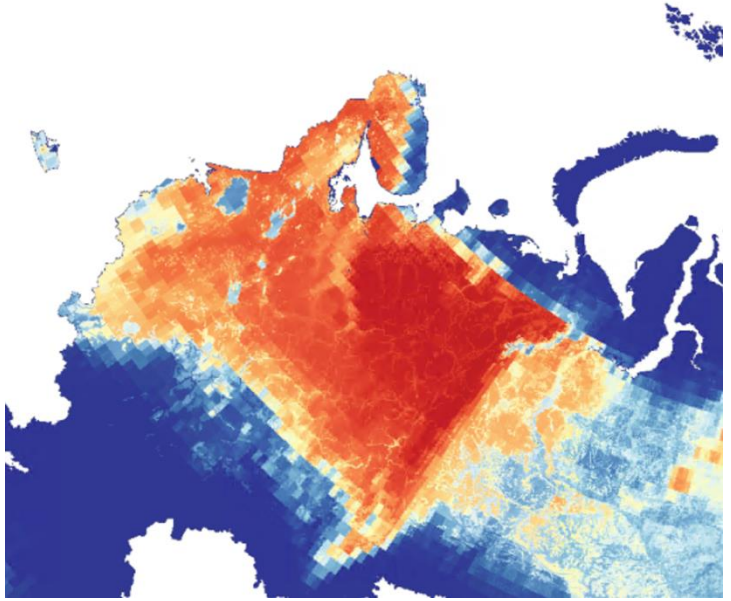
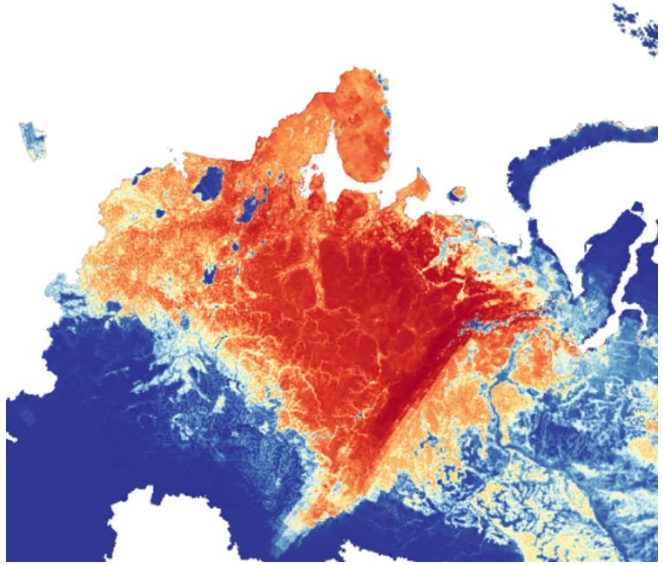
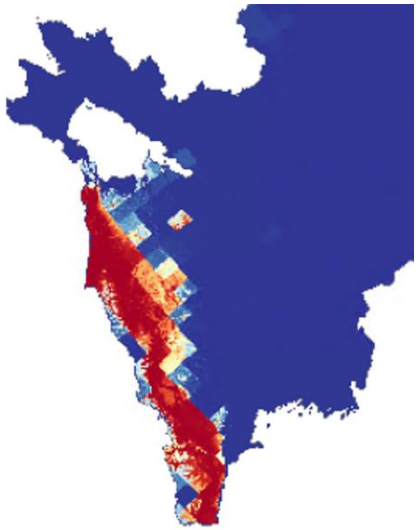
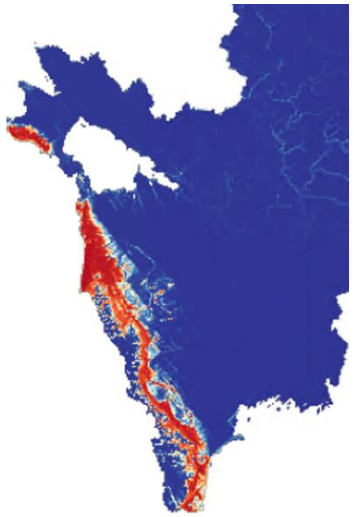


Результаты. Простые примеры

ИКИ
ИНСТИТУТ КОСМИЧЕСКИХ ИССЛЕДОВАНИЙ РАН

	Выборка	Random Forest	UNET
Клен			
Липа			

ИКИ
ИНСТИТУТ КОСМИЧЕСКИХ ИССЛЕДОВАНИЙ РАН

	Выборка	Random Forest	UNET
Ель			
Граб			

Результат применения сверточной нейросети даёт менее зашумлённые, более сглаженные и целостные карты ареалов:

+ уверенно выделяет согласованные зоны высокой и низкой вероятности;
+ лучше восстанавливает **внутреннюю структуру** ареалов и высоких вероятностей в местах с устойчивыми сочетаниями климата, рельефа и почв;
+ более **устойчива к сильной взаимосвязи** признаков: избыточные климатические каналы не ухудшают результат благодаря иерархическим сверточным признакам;
+ позволяет обучать и применять **единый классификатор** для нескольких пород и перекрывающихся ареалов;

- демонстрирует повышенную устойчивость к переобучению (особенно на малопредставленных классах), что делает её менее подходящей для задач, где требуется максимальная чувствительность к редким или очень локализованным очагам присутствия;
- для задач, где важна структура ареала – модель UNET, для задач определения наличия породы – ML.

Развитие и перспективы

Данные и признаки

- добавление новых признаков (например, удаленность от водных объектов, уровень грунтовых вод, и другие);
- добавление новых независимых источников данных, как для выборки, так и дополнительной оценки;

Перспективы:

- изучение взаимосвязи признаков и пород, поиск наиболее важных признаков;
- переход более короткие временные промежутки построения (5 лет) для изучения поведения ареалов;
- доработка текущей модели для повышения эффективности обучения и повышения метрик на разреженных данных;
- улучшение обучающей выборки;
- применение дополнительных весов входных данных;
- внедрение ареалов в качестве априорных вероятностей для карты преобладающих пород (потенциальная оптимизация, улучшение метрик карты преобладающих пород);

ИНСТИТУТ КОСМИЧЕСКИХ ИССЛЕДОВАНИЙ РАН

Моделирование ареалов потенциального распространения древесных пород на территории России с использованием методов машинного обучения

Михайлов Н.В., Барталев С.А.
Институт Космических Исследований РАН

Работа выполнена в рамках темы «Мониторинг» (госрегистрация №1220042500031-8). Обработка данных ДЗЗ проводилась с использованием ресурсов Центра коллективного пользования «ИКИ-Мониторинг».

Литература:

[1] Elith J., Leathwick J. R. Species distribution models: ecological explanation and prediction across space and time //Annual review of ecology, evolution, and systematics. – 2009. – Т. 40. – №. 1. – С. 677-697.

[2] Booth T. H. et al. BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MAXENT studies //Diversity and Distributions. – 2014. – Т. 20. – №. 1. – С. 1-9.

[3] Лупян Е.А., Прошин А.А., Бурцев М.А., Кашницкий А.В., Балашов И.В., Барталев С.А., Константинова А.М., Кобец Д.А., Мазуров А.А., Марченков В.В., Матвеев А.М., Радченко М.В., Сычугов И.Г., Толпин В.А., Уваров И.А. 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.