# ВЕРОЯТНОСТНОЕ КАРТОГРАФИРОВАНИЕ ЛЕСОВ СЕВЕРО-ЗАПАДНОГО КАВКАЗА НА ОСНОВЕ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СПУТНИКОВЫХ ДАННЫХ ГЛУБОКОГО ОБУЧЕНИЯ

***Гаврилюк Е.А.***, *Браславская Т.Ю., Шевченко Н.Е.*

**Центр по проблемам экологии и продуктивности лесов им. А.С. Исаева РАН**
г. Москва, Россия

# ЦЕЛЬ РАБОТЫ

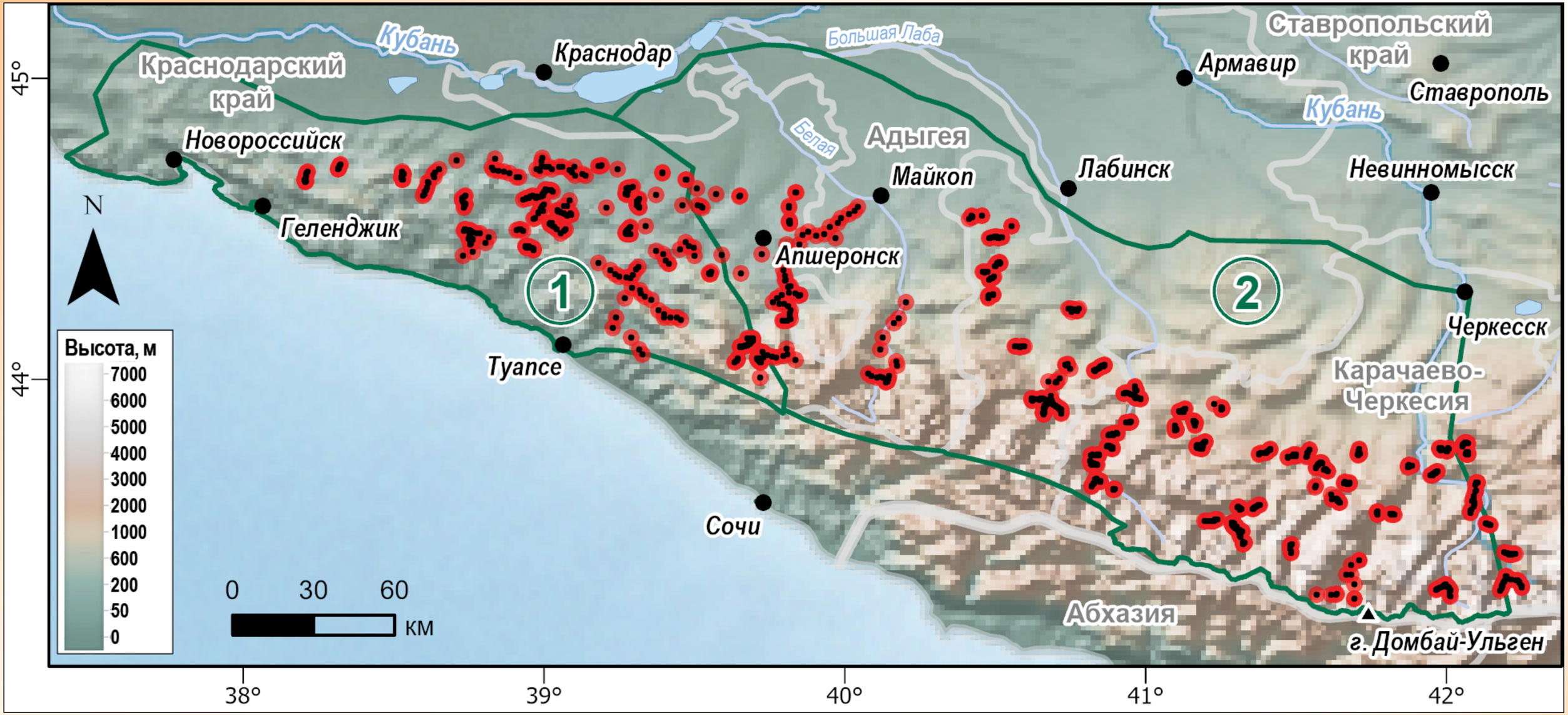


- **Северо-Западный Кавказ** представляет собой один из **ключевых мировых центров биоразнообразия**, где горные лесные экосистемы выполняют важнейшие биосферные и климаторегулирующие функции.
- В условиях выраженных **глобальных изменений** разработка надежных **методов пространственной инвентаризации лесного покрова** становится критически важной задачей.
- **Тематическое картографирование** растительности методами **машинного обучения** и **дистанционного зондирования Земли** сегодня является основным инструментом в решении подобных задач.
- Однако при работе в **горных регионах** исследователи сталкиваются со **значительными трудностями**: высокой облачностью, экстремальными эффектами рельефа и необходимостью ресурсоемкой гармонизации временных серий спутниковых изображений.
- Перспективной **альтернативой** ручному конструированию спектральных индексов и временных метрик становится использование предобученных геопространственных фундаментальных моделей и формируемых ими векторных представлений – **спутниковых эмбеддингов**.

**Целью** данной работы являлось **вероятностное картографирование синтаксонов лесной растительности Северо-Западного Кавказа** методами пространственного моделирования с **оценкой эффективности** современных **спутниковых эмбеддингов** по сравнению с **традиционными наборами геоданных**.

# ТЕРРИТОРИЯ ИССЛЕДОВАНИЯ



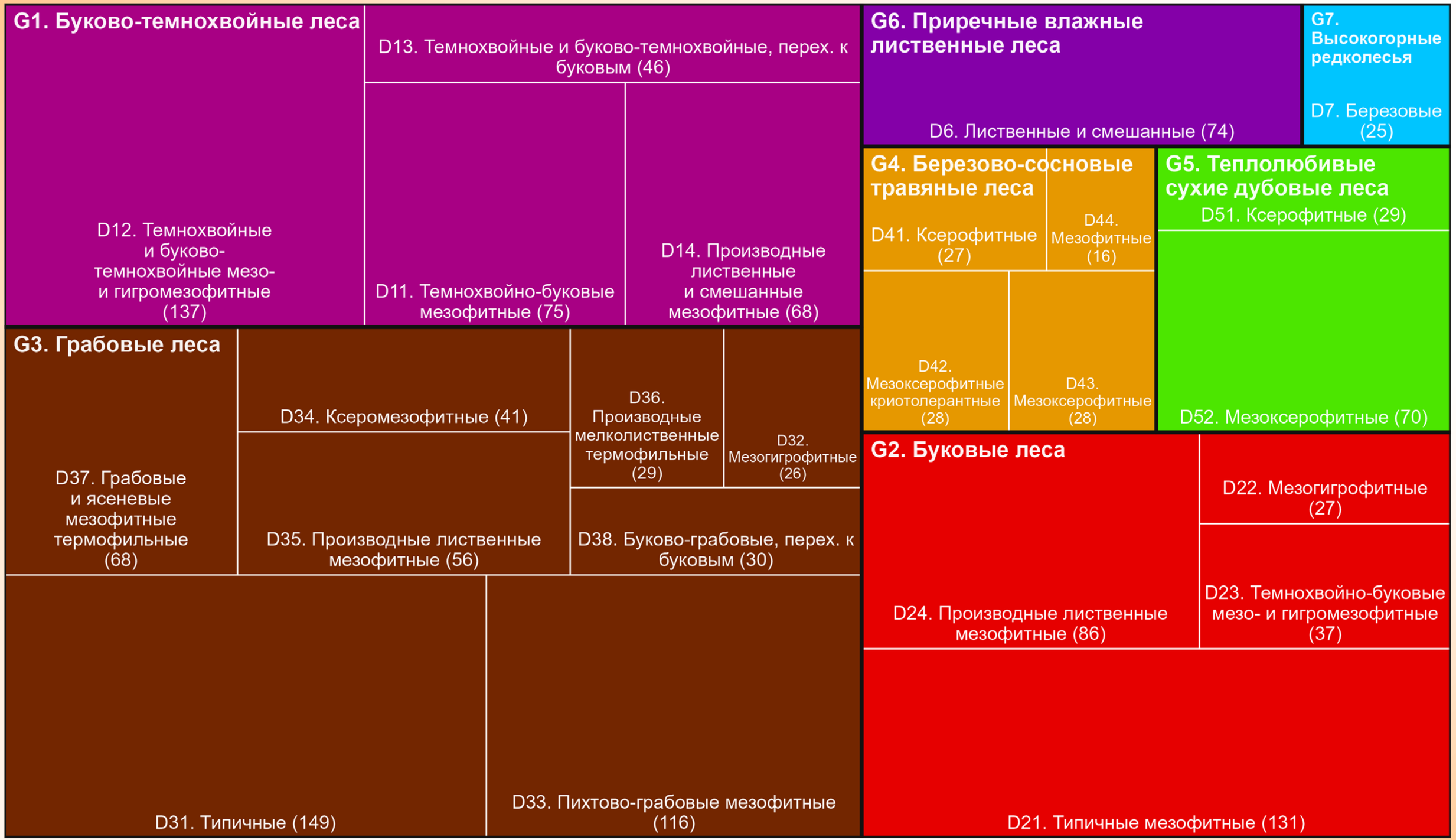



**Исходные полевые данные:**

**1419 геоботанических площадок** размером **10x10 м**, заложенные маршрутным методом в **2014-2024 гг.** в относительно крупных и малофрагментированных участках лесов на территории Крымско-Новороссийского (участок Утриш-Туапсе, 1) и Северо-Западнокавказского (2) оробиомов.

# КЛАССИФИКАЦИЯ ПОЛЕВЫХ ДАННЫХ

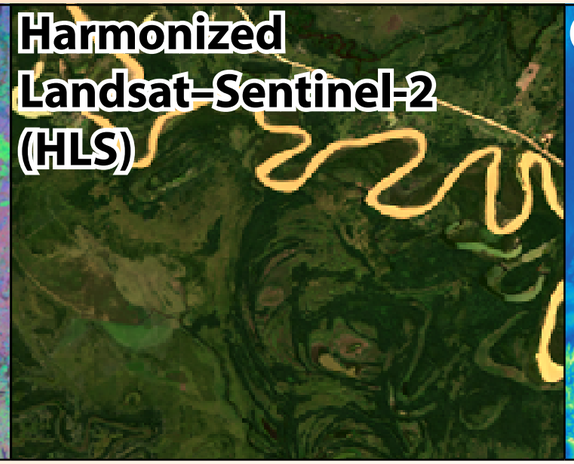
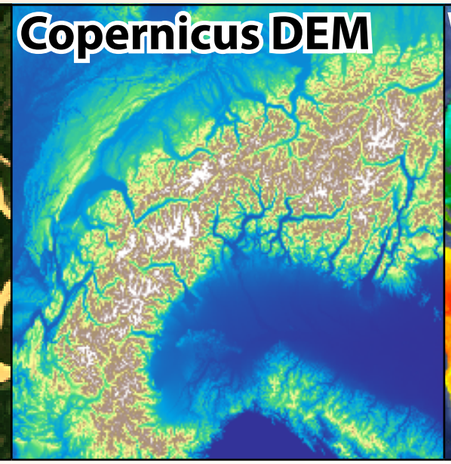
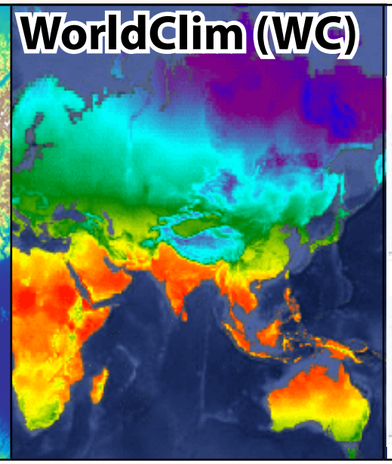
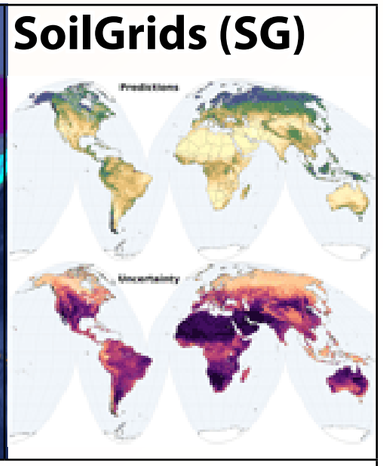





***Подписи синтаксонов:*** Обозначение (объем исходной обучающей выборки)

# ГЕОПРОСТРАНСТВЕННЫЕ ДАННЫЕ



| Данные | Спутниковые эмбеддинги AlphaEarth Foundation (AEF) | Harmonized Landsat–Sentinel-2 (HLS) | Copernicus DEM | WorldClim (WC) | SoilGrids (SG) |
|---|---|---|---|---|---|
| |  |  |  |  |  |
| **Свойства на уровне пикселей** | | | | | |
| **Разрешение** | 10 м | 30 м | 30 м | 1 км | 250 м |
| **Ресемплинг** | – | – | cubic (к 10 м) | cubic (к 10/30 м) | |
| **Временной ряд** | 1 годовой композит (среднее за 2021-2025 гг) | 25 внутригодовых композитов с шагом в 15 дней (медоиды за 2020-2024 гг) | – | – | – |
| **Набор каналов** | 64 латентных вектора | 11 (RGB+3RE+2NIR+2SWIR+TIR) | 1 (высота над уровнем моря) | 19 био-климатических переменных | 61 (10 параметров по 6 горизонтам + запас C) |
| **Производные признаки** | – | 45 попарных нормализованных отношений каналов | 19 морфометрических и 8 гидрологических признаков и индексов | – | – |
| **Признаки на уровне площадок – медоиды в буфере от центра** | | | | | |
| **Радиус поиска** | 30 м | 40 м | – | – | – |
| **Сжатие** | Базовая модель AlphaEarth | FPCA (по временному профилю) + PCA | – | – | – |
| **Всего переменных** | **64** | **64** | **28** | **19** | **61** |

***Маска лесов:*** *данные проекта* ***EarthDaily*** *(https://doi.org/10.5281/zenodo.15741437), 10 м, 2023 г.*

# СХЕМА КРОСС-ВАЛИДАЦИИ И НАСТРОЙКИ МОДЕЛЕЙ



## ЭТАП 1
## Извлечение признаков

**База геоботанических площадок**

**Наборы переменных:**

- **AEF**
- **AEF + WC**
- **AEF + WC + SG**
- **HLS**
- **HLS+WC+SG+DEM (HLS+)**

## ЭТАП 2
## Группировка фолдов

**Пространственный антикластеринг**

Мера: обратное евклидово расстояние
Минимизация расстояния между точками одного фолда и максимизация расстояний между фолдами с учетом стратификации

**5 равных фолдов**

| Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|

- Идентичный объем выборок
- Сохранение пропорций классов
- Пространственная автономия
- Защита от автокорреляции

## ЭТАП 3
## Вложенная кросс-валидация

**ВНЕШНИЙ ЦИКЛ (оценка модели)**

**1 фолд (20%) - Тест**

***ВНУТРЕННИЙ ЦИКЛ (4 фолда)***

**3 фолда (60%) - Обучение (Train)**

**1 фолд (20%) - MBO тюнинг (Val)**

4-кратная ротация Train/Val → подбор параметров
Оптимальная модель оценивается на Тесте

**Extra-Trees + MBO vs TabPFN 3.0**

- Extra-Trees:
Экстремально рандомизированные деревья (ntree = 1000)
Байесовская оптимизация (MBO):
mtry, subsample,
min.node.size, num.random.split
- TabPFN: предобученная нейросетевая модель-трансформер (настройка не требуется)

## ЭТАП 4
## Оценки эффективности

**20 повторов**

**5 фолдов × 20 реплик = 100 независимых оценок на тестовых фолдах**

**Выходные метрики**

- OAcc (общая точность)
- MCC (баланс точности)
- AUCμ (качество ранжирования)
- mBrier (СКО прогноза)
- ECE (ошибка калибровки)

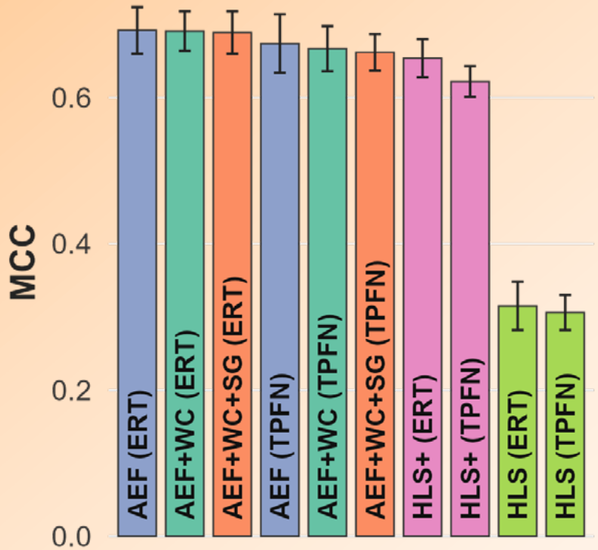
Расчет Mean ± SD

# СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МОДЕЛЕЙ

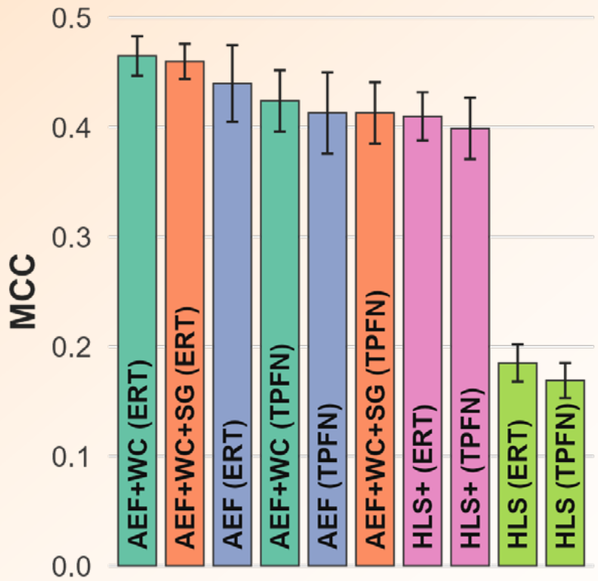


## MCC (баланс точности)

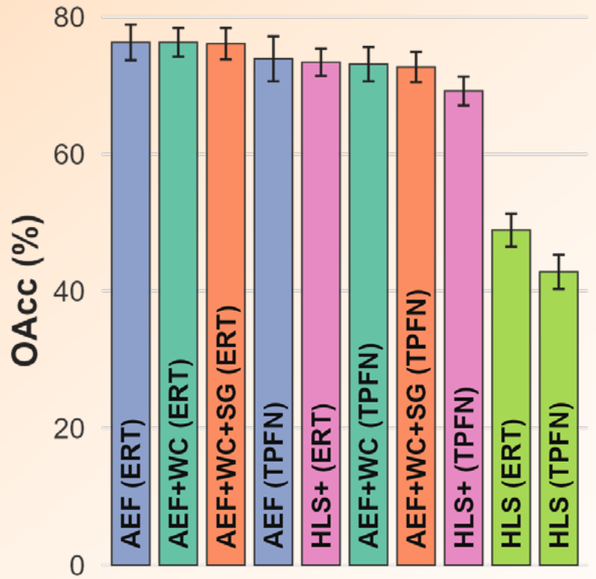
### Генерализованный уровень

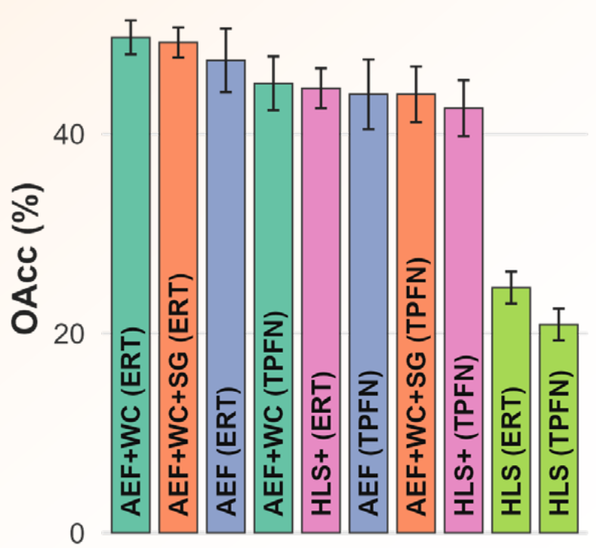



### Детальный уровень

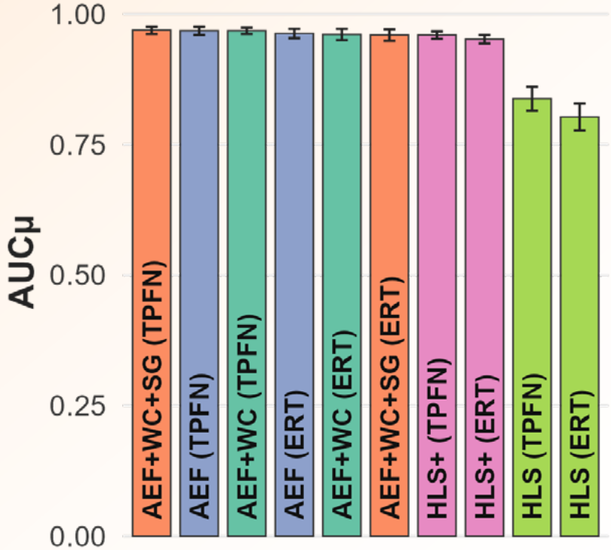


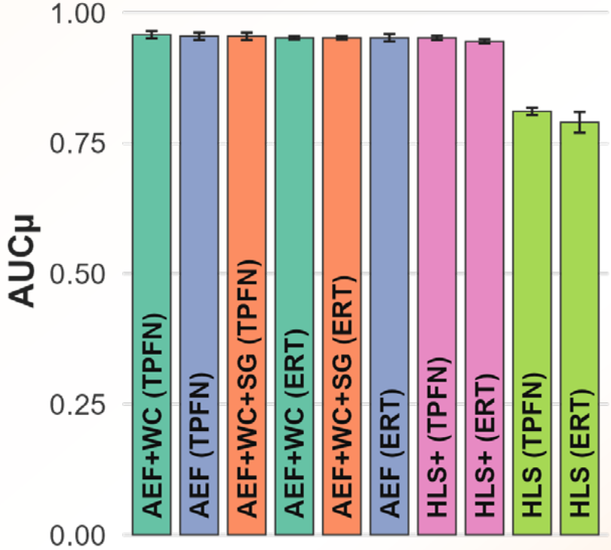

## OAcc, % (общая точность)



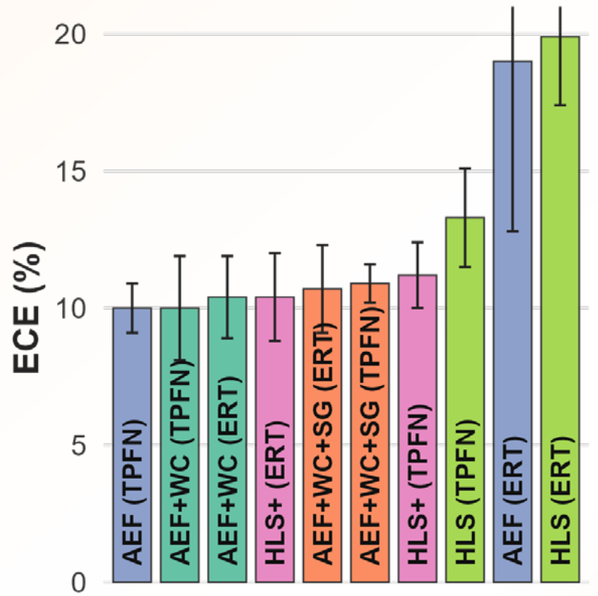



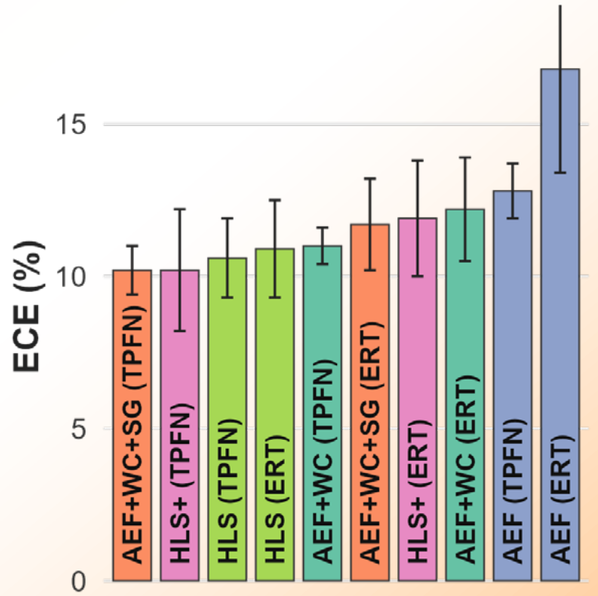

## AUCμ (ранги вероятностей)




## ECE, % (ошибка калибровки)

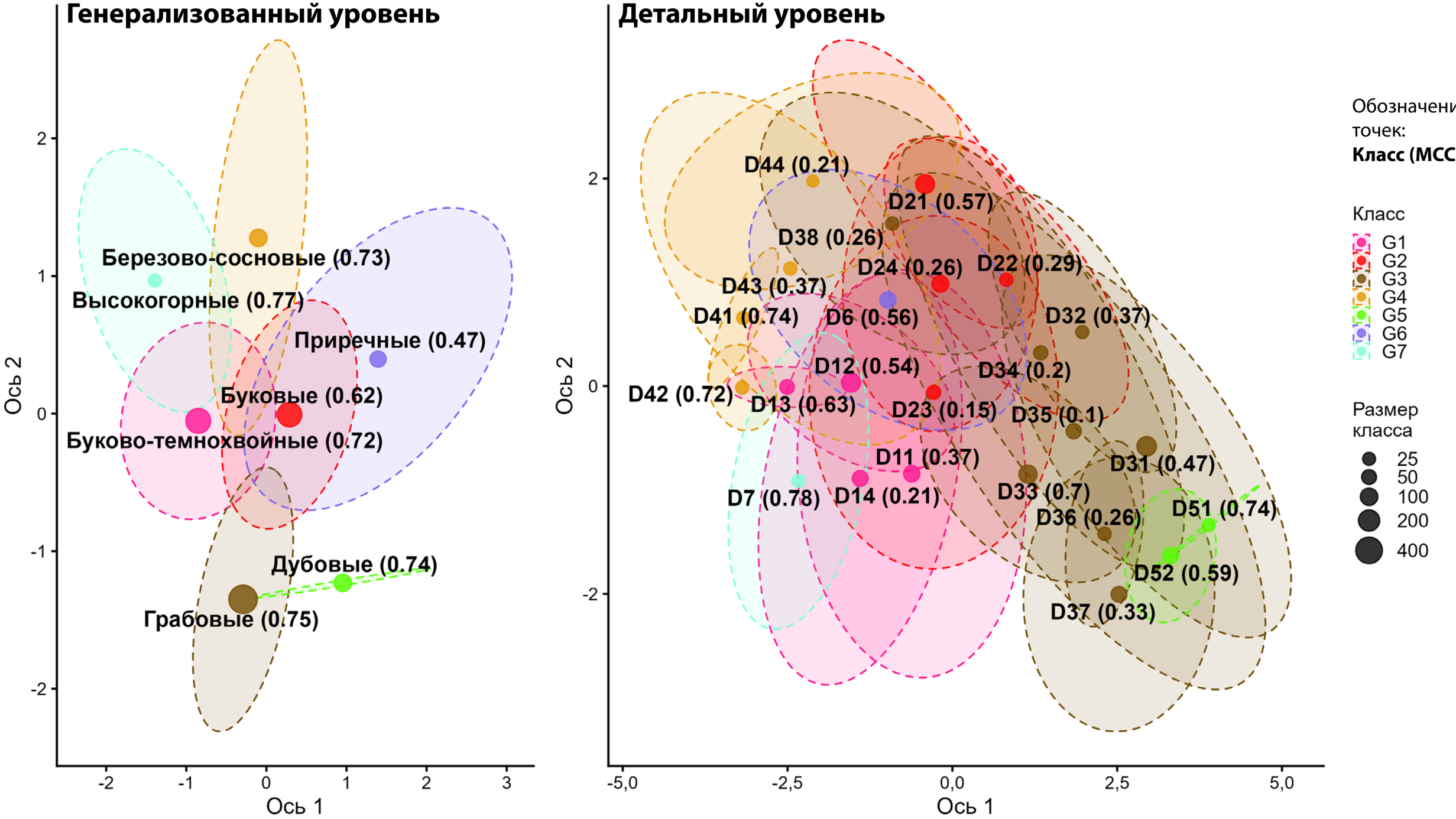



***Оптимальный вариант:*** **AEF + WC (ERT)**

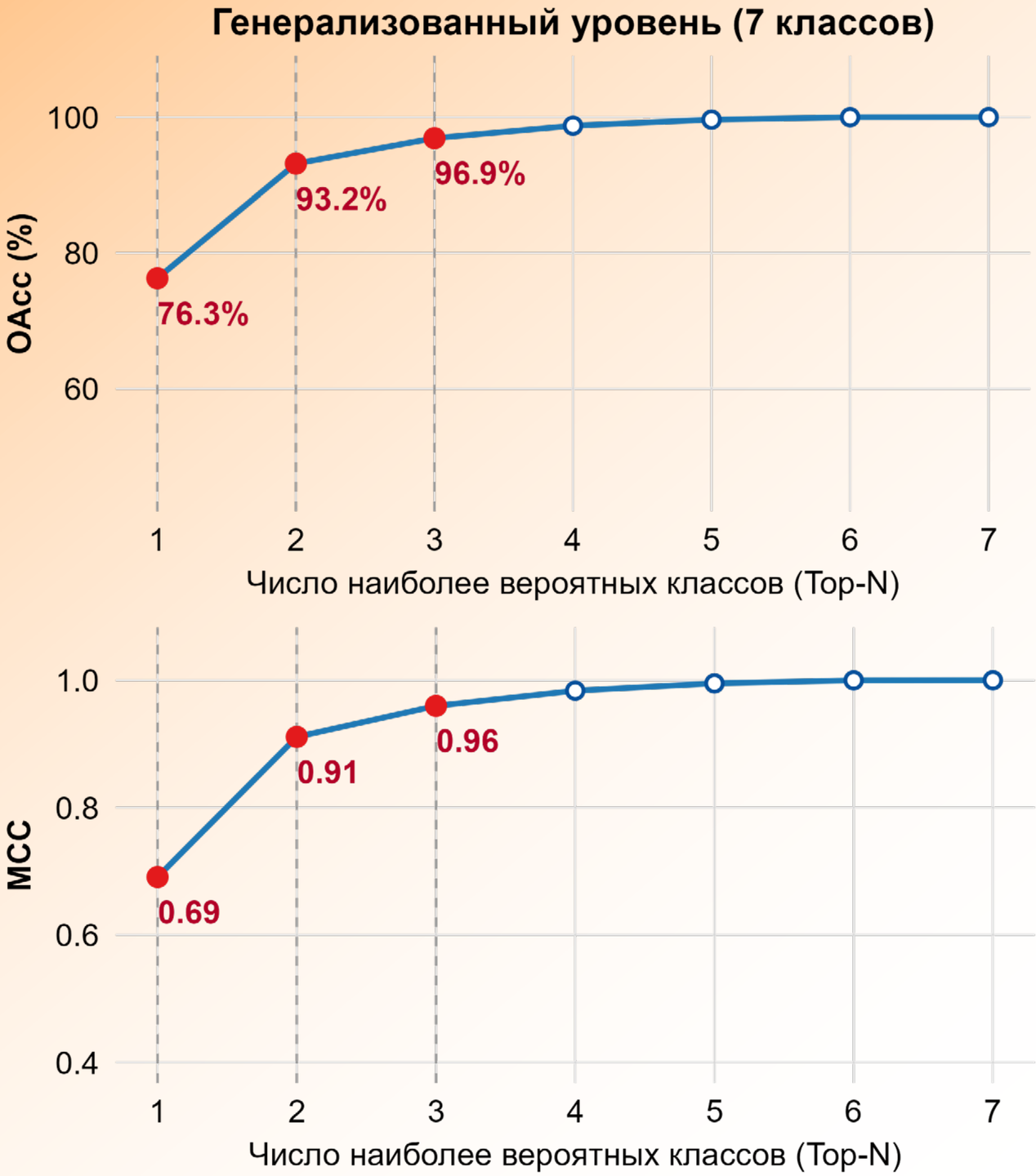
# РАЗДЕЛИМОСТЬ КЛАССОВ (UMAP-ординация по матрице ошибок)

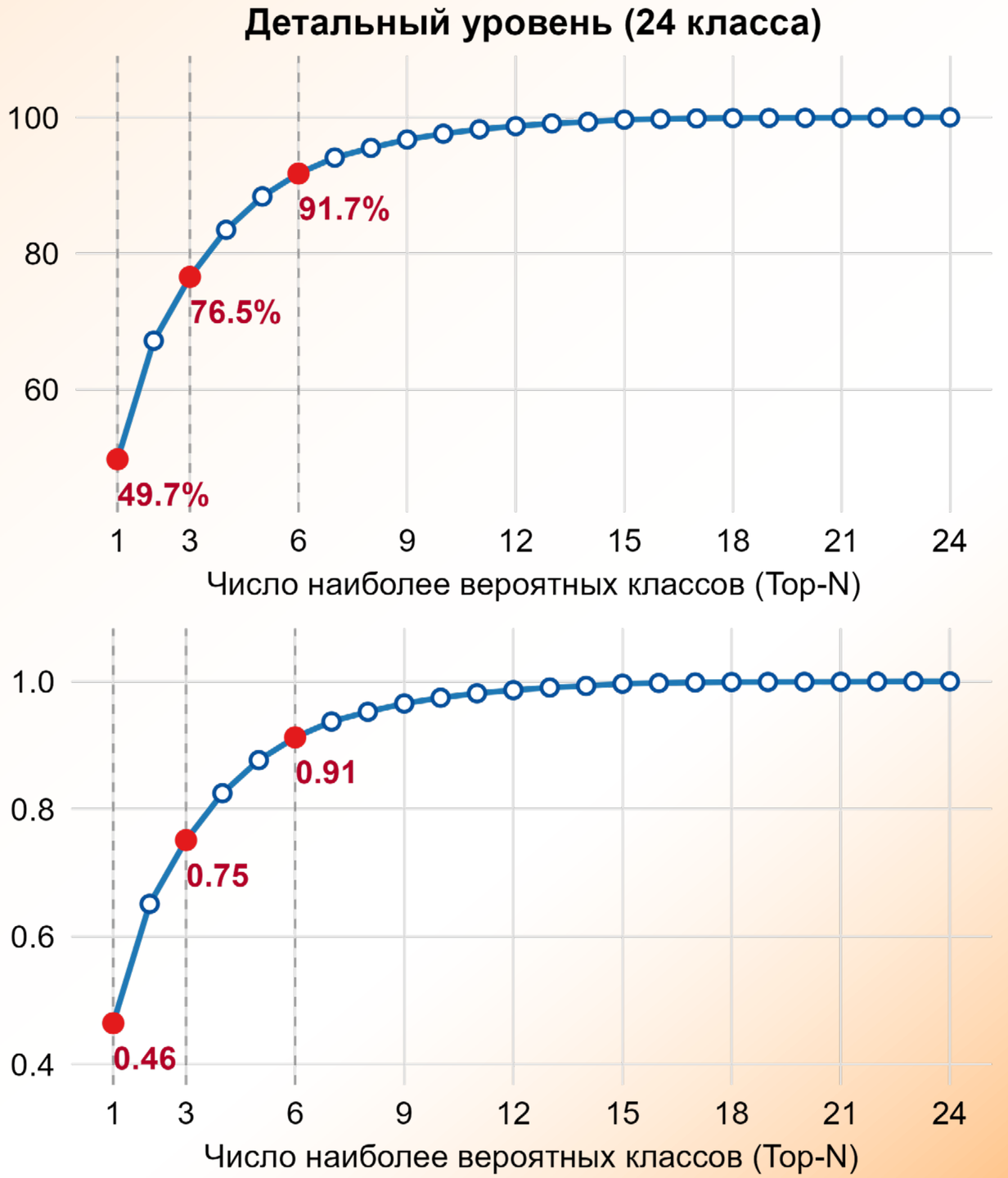




# АНАЛИЗ ТОП-N ПРЕДСКАЗАНИЙ





# ПРЕДПОСЫЛКИ К МЕТА-МОДЕЛИРОВАНИЮ ОШИБОК



» **Ограниченность жесткой (дискретной) классификации:**
Высокая дробность синтаксонов и континуальность горных лесов ограничивают точность прямого предсказания (OAcc = 50–76%). Однако высокий AUCμ > 0.95 и скачок точности при Топ-N анализе доказывают, что модель надежно улавливает структуру вероятностей.

» **Переход к вероятностной оценке площадей:**
Попиксельный подсчет доминирующих классов неизбежно ведет к потере редких и экотонных сообществ. Сохранить их баланс позволяет переход к агрегированию непрерывных вероятностей (т.н. Soft Counting).

» **Проблема локальных искажений («самоуверенность» модели):**
Сырые вероятности базового классификатора нельзя суммировать напрямую: в зонах экологических оптимумов они надежны, но на границах сообществ и в сложных условиях рельефа – систематически смещены. При этом, максимальная вероятность класса – слабый индикатор надежности (модель может «уверенно ошибаться»). Требуется независимый «гид», оценивающий риск ошибки в каждом пикселе.

» **Неприменимость глобальной пост-калибровки:**
Полевая выборка собрана маршрутным методом и не является строго пропорциональной площадям классов. Это исключает прямое использование классических глобальных эстиматоров по матрице ошибок (Olofsson et al., 2014).

**Решение:**

Разработка **пространственной мета-модели рисков ошибки**, которая оценивает достоверность каждого пикселя и позволяет локально (гибко) интерполировать предсказания между базовой моделью и априорной (кросс-валидационной) матрицей ошибок. **Мета-модель –** вероятностный бинарный классификатор, который прогнозирует риск ошибки основной модели, то есть вероятность того, что предсказанный класс (имеющий максимальное значение вероятности) определен неверно.

# МЕТА-МОДЕЛЬ РИСКОВ ОШИБКИ



В качестве **алгоритма для построения мета-модели** использовалась стандартная **логистическая регрессия с L2-регуляризацией** (т.н. Ridge-регрессия), обученная на результатах кросс-валидационных предсказаний основной классификационной модели. Для получения **доверительных интервалов оценок площадей** синтаксонов был сформирован ансамбль из **100 мета-моделей**, обученных на **случайных 50%-ных подвыборках** из общего пула кросс-валидационных предсказаний.

| **Переменные мета-модели** | **Коэффициенты** | | **Отношения шансов** | |
|---|---|---|---|---|
| | **G** | **D** | **G** | **D** |
| Максимальная предсказанная вероятность | -1,56 | -1,56 | 0,21 | 0,21 |
| Маржинальность предсказания | -0,94 | -1,12 | 0,39 | 0,33 |
| Энтропия вероятностей | 0,65 | 0,44 | 1,91 | 1,56 |
| Предсказанный класс (с минимальным риском – G7/D51) | -0,51 | -0,65 | 0,60 | 0,52 |
| Предсказанный класс (с максимальным риском – G2/D35) | 0,19 | 0,47 | 1,21 | 1,60 |
| Свободный член регрессии | -0,21 | 0,22 | 0,81 | 1,25 |
| Расстояние до выборки (географическое) | 0,01 | -0,01 | 1,01 | 0,99 |
| Расстояние до выборки (в пространстве признаков) | 0,00 | 0,00 | 1,00 | 1,00 |

**Площадь** каждого **синтаксона** рассчитывалась как **сумма скорректированных значений вероятности** соответствующего класса по всем пикселям в пределах маски лесов, умноженная на площадь одного пикселя. При этом исходные (предсказанные основной моделью) вероятности $P_{raw}$ каждого из классов **корректировались с учетом риска ошибки** по формуле:

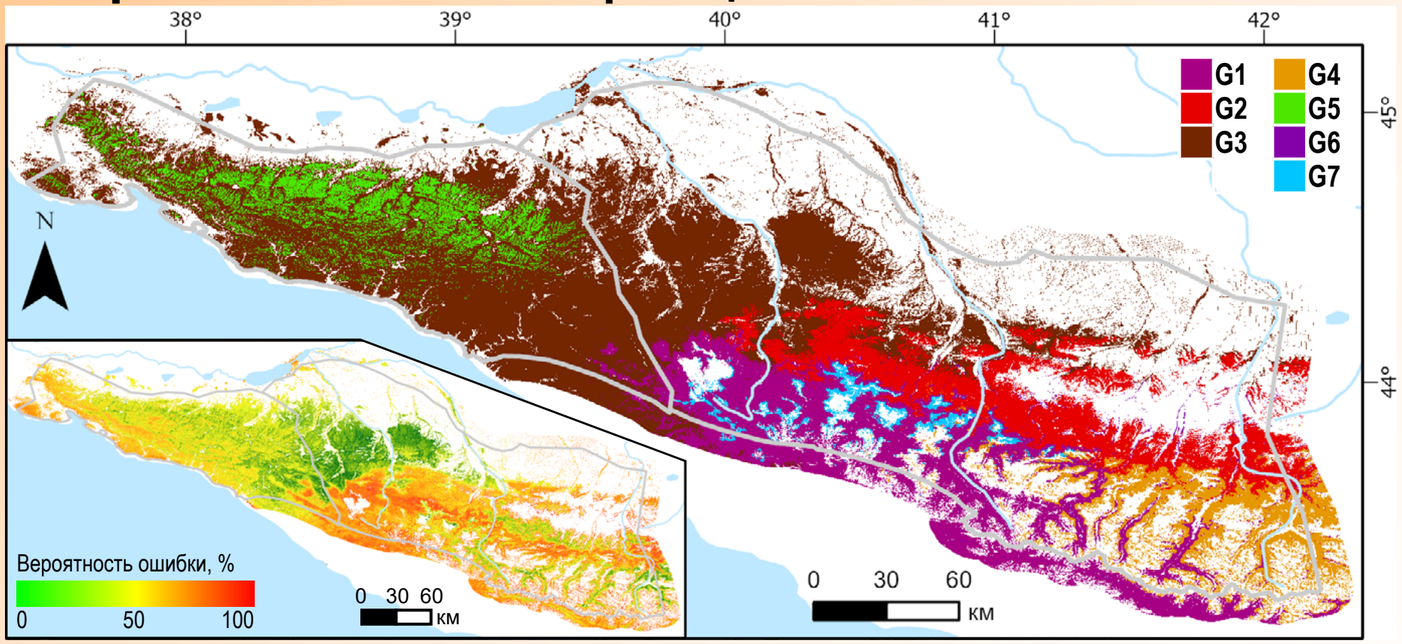
$$P_{adj} = (1 - P_{err}) \times P_{raw} + P_{err} \times P_{cor}$$

где $P_{adj}$ – скорректированная вероятность для расчета площадей, $P_{err}$ – вероятность ошибки по мета-модели, $P_{cor}$ – априорная вероятность отнесения пикселя к данному классу в случае ошибки классификатора (доля от всего объема ошибочных случаев в кросс-валидационной матрице перепутывания, приходящуюся на конкретный класс).

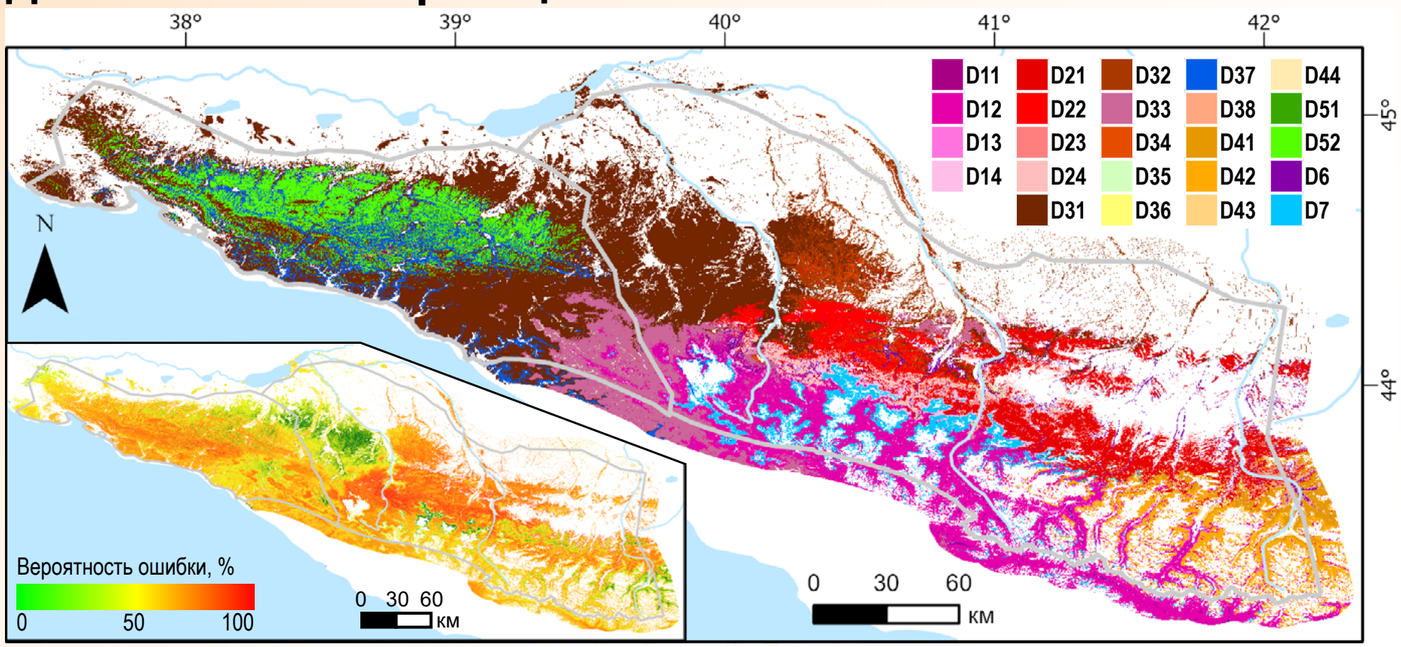
# РЕЗУЛЬТАТЫ ГЕОПРОСТРАНСТВЕННОГО МОДЕЛИРОВАНИЯ



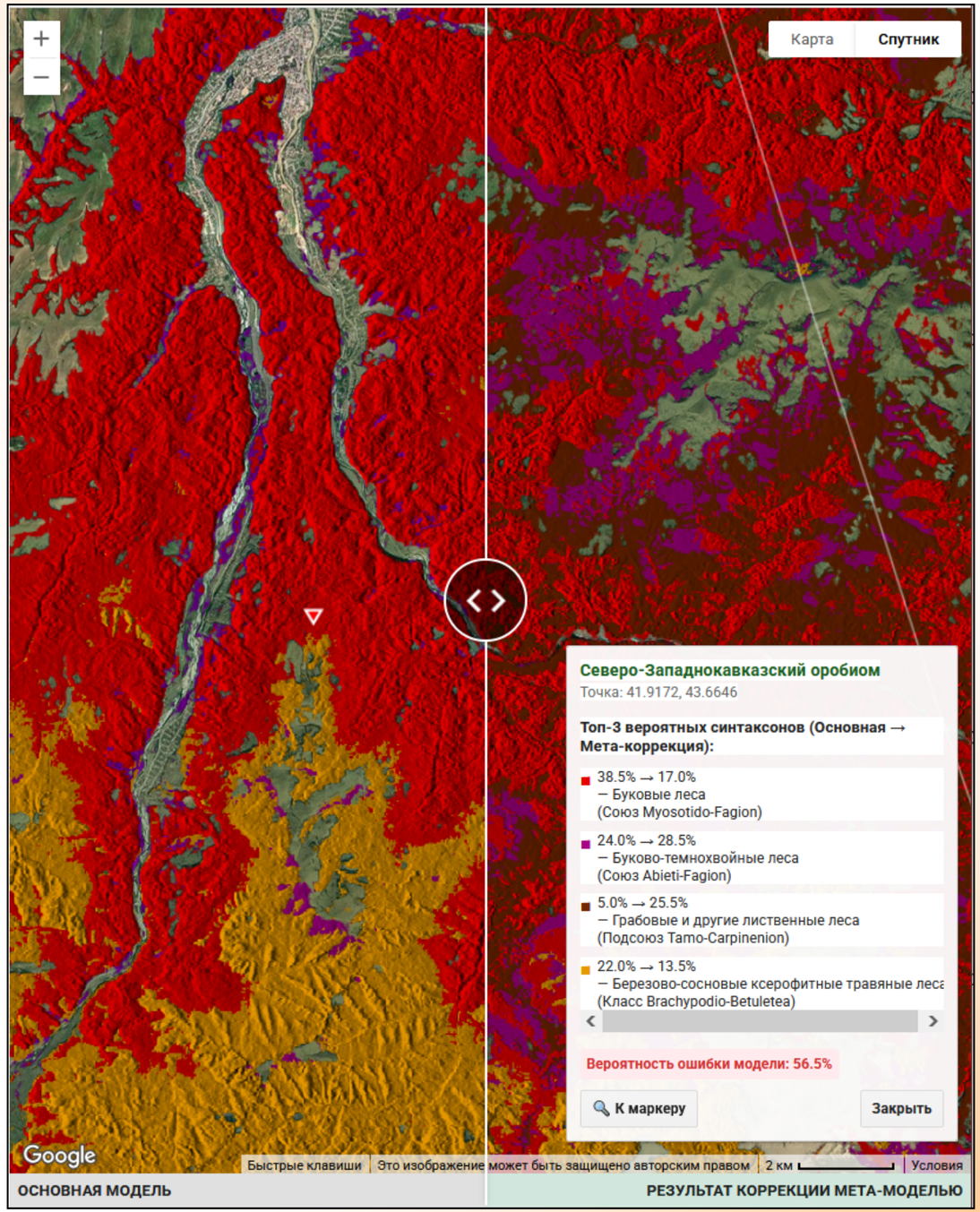
## Генерализованная классификация




## Детальная классификация



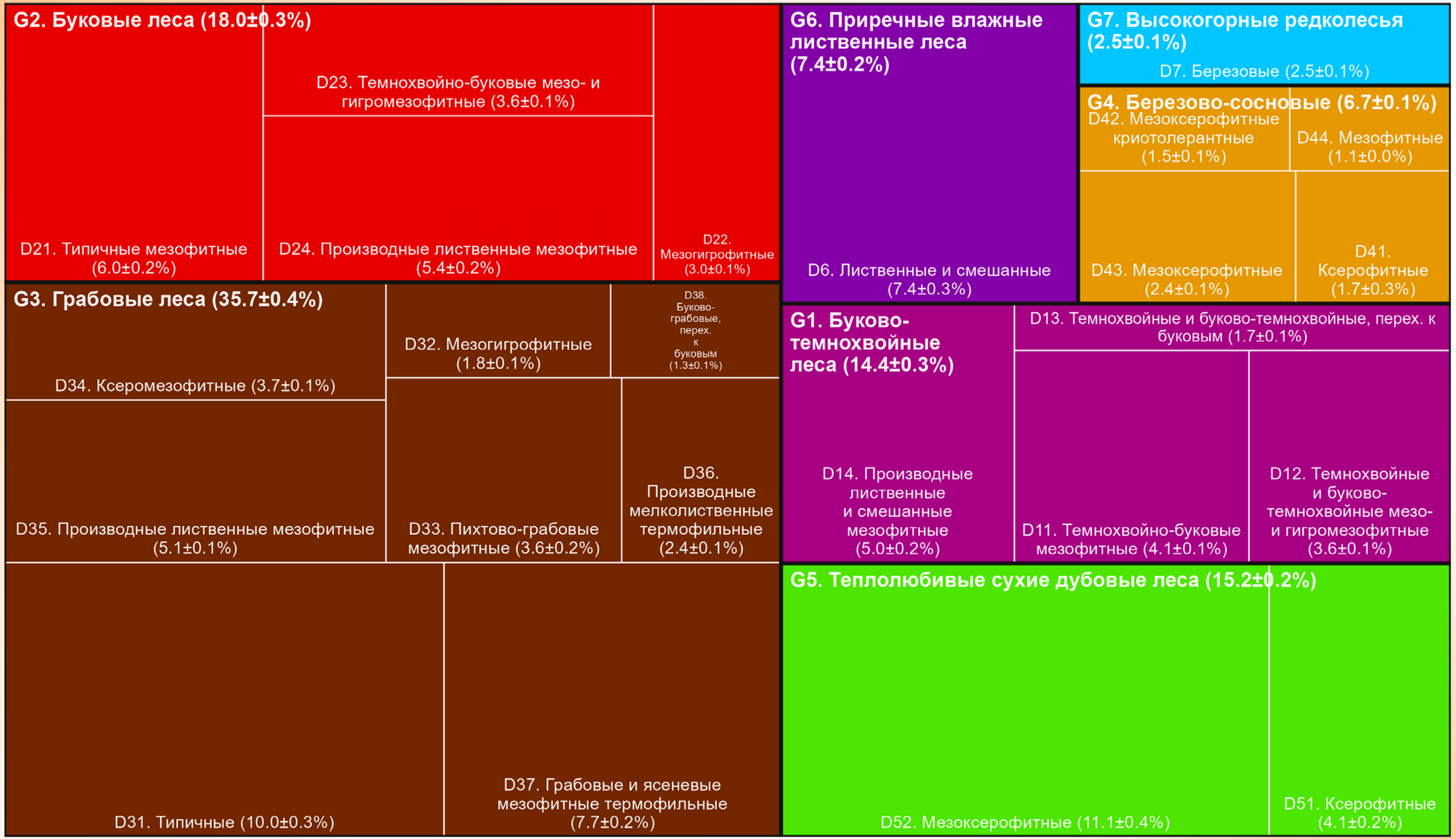

## Данные в web-приложении Google Earth Engine




***Ссылка на GEE app:*** *https://rgee-sylbermine.projects.earthengine.app/view/nwcforests*

# ОЦЕНКИ ПЛОЩАДЕЙ СИНТАКСОНОВ






***Подписи синтаксонов:*** Обозначение (доля общей площади лесов в % ± ½ 95%-го доверительного интервала в %)

# ВЫВОДЫ



» В рамках работы было реализовано **вероятностное картографирование синтаксонов лесной растительности Северо-Западного Кавказа**. По объему задействованных наземных данных и детализации эколого-флористической классификации исследование **не имеет аналогов для данной территории**. Итогом стало создание **уникального открытого набора геопространственных данных**, объективно отражающего вероятностное распределение основных лесных формаций и ассоциаций региона.

» Методологически был показан **высокий потенциал** интеграции полевых измерений с **глубокими векторными представлениями**. Готовые спутниковые эмбеддинги в сочетании с базовыми биоклиматическими переменными позволяют строить модели, не уступающие (и даже превосходящие) по стабильности классическим ансамблям на основе многокомпонентных наборов геоданных. Такой подход избавляет от необходимости ресурсоемкой предварительной обработки и гармонизации временных серий спутниковых изображений, делая процесс моделирования более эффективным.

» Исследование обозначило естественные **физические пределы дистанционного зондирования в горных условиях**. Сложная, континуальная структура лесов делает **невозможным создание высокоточных дискретных карт детальных ассоциаций**. Однако переход к парадигме **вероятностного картографирования** и оценка алгоритмов по способности выделять ограниченный набор наиболее вероятных классов смягчают это ограничение, сохраняя **высокую аналитическую ценность результатов**.

» Важным методическим аспектом стала **разработка оригинального подхода к оценке пространственной неопределенности через мета-моделирование**. Интеграция локальных рисков ошибки в схему вероятностного агрегирования площадей позволила нивелировать систематические смещения, характерные для прямого подсчета пикселей, и извлечь корректные инвентаризационные данные с узкими доверительными интервалами даже в экотонных зонах.

# *СПАСИБО ЗА ВНИМАНИЕ!*

***Гаврилюк Е.А.,*** *Браславская Т.Ю., Шевченко Н.Е.*

**Центр по проблемам экологии и продуктивности лесов им. А.С. Исаева РАН**
г. Москва, Россия

e-mail: **egor@ifi.rssi.ru**

**Работа выполнена за счет средств гранта Российского научного фонда № 25-24-00169 «Геопространственное моделирование синтаксонов хвойно-широколиственных и широколиственных лесов Северо-Западного Кавказа».**