



Цифровое картографирование содержания органического углерода в почвах Кузнецко-Салаирской геоморфологической провинции с использованием онлайн-платформы Google Earth Engine

© 2025 Н. В. Гопп

ФГБУН Институт почвоведения и агрохимии СО РАН, проспект Академика Лаврентьева, 8/2,
г. Новосибирск, 630090, Россия. E-mail: natalia.gopp@gmail.com

Цель исследования. Провести цифровое картографирование содержания почвенного органического углерода (ПОУ) в 0–30 см слое почв с использованием 11 наборов предикторов и алгоритма Random Forest (RF), реализованного на онлайн-платформе Google Earth Engine.

Место и время проведения. Исследование проведено в Кузнецко-Салаирской геоморфологической провинции (в пределах Искитимского района Новосибирской области).

Методы. Цифровое картографирование проведено с использованием алгоритма Random Forest (RF) с настройками 200 и 1000 «деревьев». Реализация алгоритма выполнена на онлайн-платформе Google Earth Engine в виде скрипта (кода). В процессе картографирования содержания ПОУ были использованы растровые карты предикторов, характеризующих следующие факторы почвообразования: климат, рельеф, растительность, пространственное положение, почвенные свойства. Моделирование проведено на основе различного количества предикторов в наборе в зависимости от коэффициента корреляции (R) между содержанием ПОУ и значениями предикторов: 1) 80 (все); 2) 50 (все с $R > \pm 0,1$); 3) 42 (все с $R > \pm 0,2$); 4) 26 (все с $R > \pm 0,3$); 5) 15 все с $R > \pm 0,4$). Дополнительно проведено картографирование содержания ПОУ с использованием пяти наборов предикторов, характеризующих отдельные факторы почвообразования (рельеф, климат, растительность, почва, пространственное положение), и одного набора, объединяющего по пять доминирующих предикторов от каждого фактора почвообразования. Таким образом, всего было использовано 11 наборов предикторов, представленных как в сочетании показателей от разных факторов почвообразования, так и по отдельным группам. Почвенная база данных создана на основе архивных материалов Института землеустройства «ЗапсибНИИгипрозем» за период с 1984 по 1994 годы и содержит данные по содержанию ПОУ в 0–30 см слое почв ($n=261$). В работе использован обучающий (ОНД, $n=206$) и валидационный (ВНД, $n=55$) наборы почвенных данных.

Основные результаты. Лучший результат картографирования содержания ПОУ в 0–30 см слое почв алгоритмом RF (200 «деревьев») получен на основе 42 предикторов. Показатели эффективности моделирования следующие: $R^2_{ОНД}=0,83$; $R^2_{ВНД}=0,55$; $RMSE_{ВНД}=1,25\%$. По фактическим данным, содержание ПОУ в 0–30 см слое исследуемых почв варьировало от 1,2% до 13,9%. Согласно составленной карте, наибольшее содержание ПОУ выявлено в почвах юго-западной части исследуемой территории.

Заключение. Использование растровых карт предикторов с низким исходным разрешением (1×1 км – WorldClim, 250×250 м – SoilGrids) и улучшенным до 30×30 м в сочетании с картами предикторов среднего разрешения (рельеф, растительность, пространственное положение) позволяет достичь лучших результатов в картографировании содержания ПОУ. Одни группы предикторов (климат, почва, пространственное положение) позволяют отобразить на итоговой карте глобальный тренд в изменении содержания ПОУ, а другие (рельеф, растительность) – дифференцирующие/детализирующие тренды, учитывающие особенности рельефа и растительности. Таким образом, предикторы, характеризующие разные факторы почвообразования, взаимно дополняют друг друга в их совмещённом наборе и позволяют получить более детальные/дифференцированные карты при проведении картографирования содержания ПОУ в почвах.

Ключевые слова: цифровое картографирование; содержание органического углерода; почвы; Западная Сибирь; WorldClim; FABDEM; SoilGrids; Landsat 5; SAGA GIS.

Цитирование: Гопп Н.В. Цифровое картографирование содержания органического углерода в почвах Кузнецко-Салаирской геоморфологической провинции с использованием онлайн-платформы Google Earth Engine // Почвы и окружающая среда. 2025. Том 8. № 3. e335. DOI: [10.31251/pos.v8i3.335](https://doi.org/10.31251/pos.v8i3.335)

ВВЕДЕНИЕ

Картографические материалы, полученные в результате изучения пространственной изменчивости содержания почвенного органического углерода (ПОУ) с использованием данных дистанционного зондирования Земли, могут быть полезны, в первую очередь, для землепользователей и ученых, изучающих изменения свойств почв и климата. Снижение содержания ПОУ является показателем интенсивности использования почвы и происходящих в ней процессов деградации, локализацию которых можно выявить с использованием картографического подхода, в дальнейшем

приняв меры по их устранению. Наряду с этим, при протекании различных процессов (минерализации, секвестрации и др.) почвы вносят значительный вклад в углеродный обмен между наземными экосистемами и атмосферой, являясь источниками выбросов и поглощения парниковых газов, которые оказывают как положительное, так и отрицательное воздействие на климат Земли (IPCC, 2006; FAO and ITPS, 2020; Шарков, Антипина, 2022; Романова и др., 2024). Нарастание проблемы выбросов парниковых газов привело к формированию рынка торговли квотами, который стимулирует их сокращение (Zhang, Wen, 2022). Изучение различных процессов, связанных с изменением содержания ПОУ, как в пространственном, так и временном аспектах, неинформативно без картографического подхода, который позволяет визуализировать пространственную изменчивость изучаемых свойств почв и произвести необходимые расчёты, поэтому исследования в данном направлении являются актуальными.

Анализ исследований показал, что алгоритм машинного обучения Random Forest (**RF**) чаще всего использовался для цифрового картографирования свойств почв (Гопп и др., 2023; Duarte et al., 2022; Suleymanov et al., 2023). Этот алгоритм реализован в программном обеспечении R (R Core Team, 2022) и на онлайн-платформе Google Earth Engine (Gorelick et al., 2017). Google Earth Engine (GEE) предоставляет готовые к использованию наборы геопространственных данных и несколько серверов для их обработки. Возможность распределять вычисления по серверам делает платформу GEE привлекательной для выполнения трудоемких задач (Gorelick et al., 2017; Gandhi, 2023). Алгоритм машинного обучения Random Forest обладает рядом преимуществ перед другими алгоритмами классификации и регрессии. К числу этих преимуществ относятся следующие: 1) значительное повышение точности за счет разработки ансамбля деревьев регрессии (Breinman, 2001); 2) новый подход к определению значимости переменных; 3) возможность моделировать сложные взаимодействия переменных; 4) возможность выполнять различные виды статистического анализа, включая регрессию и классификацию; 5) использование алгоритмов для вычисления недостающих данных (Cutler et al., 2007); 6) использование дополнительной информации с помощью различных типов энтропии (Ignatenko et al., 2024); 7) устойчивость к уменьшению объема обучающих данных и шумам (Rodrigues-Galiano et al., 2012); 8) отсутствие необходимости в предположении о нормальности распределения данных; 9) возможность получения приемлемого прогноза за пределами области полевых исследований. Потенциал методов машинного обучения для цифрового картографирования свойств почв трудно переоценить, поскольку эти методы позволяют обрабатывать большое количество данных (качественных/количественных/категориальных) и точечных лабораторно-полевых измерений без ограничений на их статистические параметры.

Предикторы играют важную роль в процессе создания цифровых карт пространственной изменчивости свойств почв при наличии ограниченного количества геопривязанных пунктов пробоотбора лабораторно-полевых данных, необходимых для обучения моделей при использовании алгоритмов машинного обучения. Однако проблема выбора информативных предикторов до сих пор остается дискуссионной и в большей степени зависит от предпочтений экспертов или их профессиональных навыков работы с теми или иными геопространственными данными. Порой эксперты вынуждены отдавать предпочтение предикторам, характеризующим один или два фактора почвообразования, при этом отсутствие в модели других очень важных факторов может привести к значительному снижению показателей эффективности моделирования. Зачастую из-за отсутствия необходимых геопространственных данных, некоторые предикторы вообще не используются в моделировании. Например, при картографировании содержания ПОУ, такого важного предиктора в виде растровых карт как количество поступивших растительных остатков за весь период почвообразования не существует, а получить эти данные с использованием дистанционного зондирования Земли (**ДЗ**) невозможно. Безусловно, отсутствие растровых карт подобных важных показателей негативно сказывается на точности геопространственного моделирования содержания ПОУ. Поэтому, чаще всего, в геопространственном моделировании свойств почв используют показатели, определяемые по данным **ДЗ** или архивным источникам, которые оказывают косвенное/опосредованное влияние на процессы, приводящие к изменению содержания ПОУ.

Стратегия выбора информативных предикторов, в первую очередь, основана на знании как самих предикторов (статические, сезонные, среднегодовые и т.д.), так и характеристик геопространственных данных (разрешение, количество спектральных диапазонов, наличие облаков, атмосферной коррекции и т.д.). В частности, дискуссионным является вопрос об использовании отдельных мультиспектральных снимков или их композиций для сезона или года, сезонных композиций за разные годы. Например, в исследовании (Luo et al., 2022) продемонстрировано, что точность картографирования содержания органического вещества в почве на основе снимков Landsat-8 разных

временных периодов существенно различается; в частности, наилучшие результаты картографирования были получены на основе использования снимков, усредненных за разные годы и сделанных в периоды отсутствия растительности на почве. В других исследованиях (Zhou et al., 2023) показано, что совместное использование снимков Sentinel-1 и Sentinel-2 повысило точность картографирования содержания ПОУ.

Размер области исследования имеет важное значение при выборе предикторов, рассчитываемых по мультиспектральным снимкам. Если территория большая, то это может привести к получению пространственных данных, которые будут соответствовать разным фазам развития растительности. Это связано с трудностями получения снимков большой территории для одной и той же фазы роста растений, поскольку спутниковая съемка таких территорий не может быть произведена в одно и то же время. Чаще всего при региональном картографировании больших территорий составляют мозаику из разновременных снимков, поэтому такая процедура может привести к сложностям интерпретации полученных результатов моделирования.

Очень часто в геопространственном моделировании используют климатические прогнозы с низким пространственным разрешением, например, WorldClim (разрешение 1×1 км) (Fick, Hijmans, 2017). Эти прогнозы основаны на долгосрочных гидрометеорологических данных с удаленных друг от друга постов. Однако применение этих данных даже на небольших территориях с использованием процедур, улучшающих разрешение и качество растровых карт, показывает, что переменные климата часто оказываются в числе ведущих предикторов (Гопп, Нечаева, 2024; Moquedace et al., 2024). Безусловно, улучшение разрешения растровых карт не способствует повышению точности или детальности представления климатических данных. Однако положительным моментом является то, что, согласно концепции SCORPAN (Arrouays et al., 2020), такой важный фактор почвообразования как климат (C) учитывается в совокупном анализе данных; использование этих данных позволяет отразить на итоговых картах глобальные тренды в изменении свойств почв.

Цель исследования – провести цифровое картографирование содержания ПОУ в 0–30 см слое почв с использованием 11 наборов предикторов и алгоритма Random Forest, реализованного на онлайн-платформе Google Earth Engine.

Картографирование содержания ПОУ в почвах основывалось на концепции **SCORPAN** (Arrouays et al., 2020), которая подразумевает использование имеющихся пространственных данных о факторах почвообразования, характеризующих свойства почв (S), климат (C), организмы (O), рельеф (R), почвообразующие породы (P), возраст (A) и пространственное положение (N).

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Исследуемый регион площадью 4384 км^2 расположен в западной части Кузнецко-Салаирской геоморфологической провинции Алтае-Саянской горной страны (Российская Федерация, Новосибирская область, Искитимский район) (рис. 1).

Климат региона резко континентальный; среднегодовые температуры колеблются от $-0,3^\circ\text{C}$ до $-0,6^\circ\text{C}$; средняя температура января составляет -19°C , а средняя температура июля – $+19^\circ\text{C}$. Годовое количество осадков варьирует от 400 до 450 мм. Территория исследования относится к лесостепной зоне и имеет денудационно-аккумулятивный рельеф.

Почвенная база данных (рис. 1), содержащая описание почвенных профилей ($n=261$), создана на основе архивных материалов Института землеустройства «ЗапсибНИИгипрозем» за период с 1984 по 1994 годы (Гопп, 2021). В базе данных содержится описание почв, в основном, пахотных угодий, доля почв других угодий (сенокосы и пастбища) составляет 1,5%. Таксономические названия почв представлены по классификации и диагностике почв СССР (Классификация ..., 1977) и мировой реферативной базе данных почвенных ресурсов WRB (IUSS Working Group ..., 2015). Черноземы, темно-серые лесные, серые и светло-серые лесные почвы сформировались на лёссовидных карбонатных суглинках; на оглеенных лёссовидных суглинках – лугово-черноземные и луговые почвы; на аллювиальных отложениях рек (супесях и песках) – аллювиальные луговые почвы. Гранулометрический состав почвообразующих пород и почв в основном представлен средними и тяжелыми суглинками (иловато-пылеватыми).

Содержание ПОУ определено методом Тюрина (Аринушкина, 1970). Согласно рекомендациям (FAO ..., 2021), полученные результаты по содержанию ПОУ были умножены на коэффициент 1,15, который позволяет учесть ошибки, связанные с неполным окислением органического углерода, определяемого методом Тюрина.

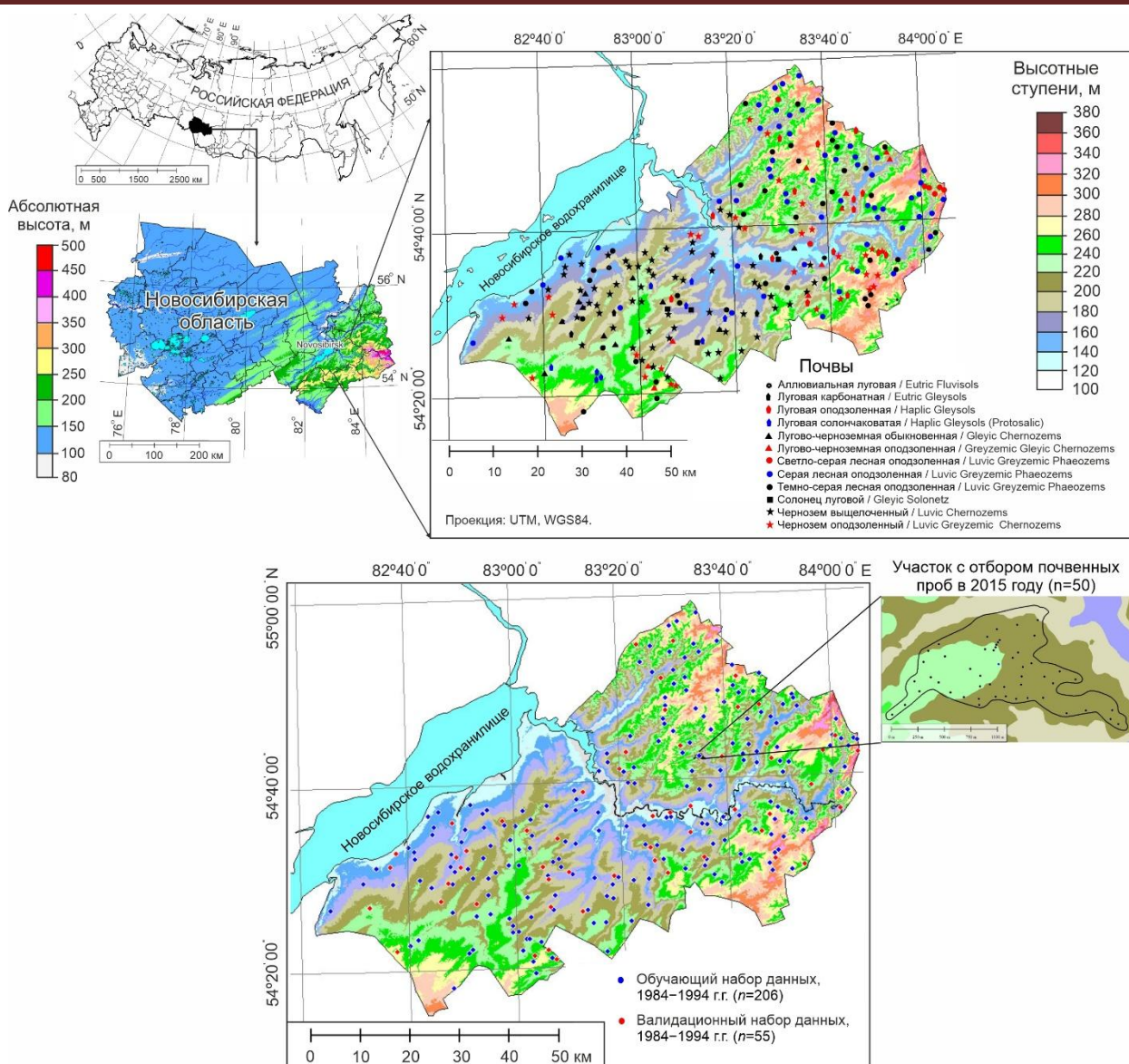


Рисунок 1. Карта высотных ступеней исследуемой территории и расположение почвенных разрезов ($n=261$). Названия почвенных таксономических единиц приведены по классификации и диагностике почв СССР (Классификация ..., 1977), через косую черту – по мировой реферативной базе данных почвенных ресурсов WRB (IUSS Working Group ..., 2015).

В процессе картографирования содержания ПОУ были использованы растровые карты предикторов, характеризующих следующие факторы почвообразования: климат, рельеф, растительность, пространственное положение, почвенные свойства (табл. 1).

Растровые карты предикторов, характеризующих рельеф, построены на основе цифровой модели высот FABDEM V1-2 (Forest and Buildings removed Copernicus DEM, разрешение 30×30 м), в которой произведена корректировка высоты на участках, занятых лесами и зданиями (Hawker et al., 2022). Данные FABDEM можно использовать для получения информации об общих топографических характеристиках местности, как на уровне хозяйства, так и на уровне отдельных полей (Erupova, Yakubailik, 2024).

Усредненные значения отражения по каждому каналу (Band 1, 2, 3, 4, 5, 7) получены на основе безоблачных снимков со спутника Landsat 5 TM (разрешение 30×30 м, Level-2 Data Products – Surface Reflectance) за следующие даты съемки в вегетационном периоде: 04.07.1989 г., 20.05.1990 г., 21.06.1990 г., 07.05.1994 г. На основе полученных усредненных данных по необходимым каналам вычислены различные вегетационные индексы (см. табл. 1) в программном обеспечении SAGA GIS (Conrad et al., 2015). В качестве категориального предиктора, характеризующего наземный покров, был использован продукт ESA CCI LC (European Space Agency Climate Change Initiative Land Cover) за 1992 год (Defourny et al., 2017; Defourny et al., 2023).

Таблица 1

Перечень предикторов, использованных в цифровом картографировании содержания ПОУ

Сокращенное название / интерпретация	Интерпретация на русском языке
Предикторы, характеризующие климат (на основе данных WorldClim и архивных карт). Исходное разрешение 1×1 км, в настоящей работе преобразовано к 30×30 м	
BIO1/Annual Mean Temperature	Среднегодовая температура
BIO2 / Mean Diurnal Range (Mean of monthly (max temp - min temp))	Среднесуточная разность (среднее значение за месяц (макс.температура - мин. температура))
BIO3 / Isothermality ((BIO2/BIO7) ×100)	Изотермичность (BIO2/BIO7) (×100)
BIO4 / Temperature Seasonality (stand. dev.×100)	Сезонность температуры (стандартное отклонение ×100)
BIO5 / Max Temperature of Warmest Month	Максимальная температура самого теплого месяца
BIO6 / Min Temperature of Coldest Month	Минимальная температура самого холодного месяца
BIO7 / Temperature Annual Range (BIO5-BIO6)	Годовой диапазон температур (BIO5-BIO6)
BIO8 / Mean Temperature of Wettest Quarter	Средняя температура самого влажного квартала
BIO9 / Mean Temperature of Driest Quarter	Средняя температура самого сухого квартала
BIO10 / Mean Temperature of Warmest Quarter	Средняя температура самого теплого квартала
BIO11 / Mean Temperature of Coldest Quarter	Средняя температура самого холодного квартала
BIO12 / Annual Precipitation	Годовое количество осадков
BIO13 / Precipitation of Wettest Month	Количество осадков в самый влажный месяц
BIO14 / Precipitation of Driest Month	Количество осадков в самый засушливый месяц
BIO15 / Precipitation Seasonality (Coefficient of Variation)	Сезонность осадков (коэффициент вариации)
BIO16 / Precipitation of Wettest Quarter	Количество осадков в самом влажном квартале
BIO17 / Precipitation of Driest Quarter	Количество осадков в самом сухом квартале
BIO18 / Precipitation of Warmest Quarter	Количество осадков в самом теплом квартале
BIO19 / Precipitation of Coldest Quarter	Количество осадков в самом холодном квартале
SRmean / Mean Solar Radiation per year (sum by month divided by 12)	Среднее солнечное излучение за год (сумма по месяцам, деленная на 12)
SAT10 / Sum of Active Temperatures above 10°C	Сумма активных температур выше 10°C
CC / Coefficient of Continentality of Climate	Коэффициент континентальности климата
EVAP / Evaporation	Испаряемость
CH / Coefficient of Humidity	Коэффициент увлажнения
AgPot / Agroecological potential	Агроэкологический потенциал
Предикторы, характеризующие рельеф (на основе FABDEM V1-2, разрешение 30×30 м)	
ELEV / Elevation	Высота над уровнем моря
Slope	Крутизна склонов
LS / LS-Factor	Коэффициент соотношения длины и крутизны склона
Aspect	Экспозиция склонов
CrosSecCurv / Cross-Sectional Curvature	Кривизна поперечного сечения
FlowLineCurv / Flow Line Curvature	Кривизна линии потока
GenCurv / General Curvature	Главная кривизна
LongCurv / Longitudinal Curvature	Продольная кривизна
MinCurv / Minimal Curvature	Минимальная кривизна
MaxCurv / Maximal Curvature	Максимальная кривизна
ProfCurv / Profile Curvature	Профильная кривизна
PlanCurv / Plan Curvature	Плановая кривизна
TanCurv / Tangential Curvature	Тангенциальная кривизна
TotCurv / Total Curvature	Общая кривизна
ConvIndex / Convergence Index	Индекс конвергенции
Texture / Terrain Surface Texture	Рельефная текстура поверхности
Convexity / Terrain Surface Convexity	Индекс выпуклости
MassBalans / Mass Balance Index	Индекс баланса массы
AnalitHill / Analytical Hillshading	Аналитическая затененность холмов
VallDepth / Valley Depth	Глубина долин
TWI / Topographic Wetness Index	Топографический индекс влажности
TCA / Total Catchment Area	Общая площадь водосбора
RSP / Relative Slope Position	Относительное положение на склоне
TPI / Topographic Position Index	Топографический индекс положения
TRI / Terrain Ruggedness Index	Индекс расчлененности рельефа

CND / Channel Network Distance	Расстояние до водотоков (дренажной сети)
CNBL / Channel Network Base Level	Базовый уровень водотоков
VisSky / Visible Sky	Видимость неба
SkyViewFact / Sky View Factor	Фактор видимости неба
Предикторы, характеризующие пространственное положение (разрешение 30×30 м)	
LONG / Longitude	Географическая долгота
LAT / Latitude	Географическая широта
Предикторы, характеризующие растительность (на основе Landsat 5 TM, разрешение 30×30 м)	
NDVI / Normalized Difference Vegetation Index	Нормализованный разностный вегетационный индекс
CTVI / Corrected Transformed Vegetation Index	Скорректированный трансформированный вегетационный индекс
DVI / Difference Vegetation Index	Разностный вегетационный индекс
NRVI / Normalized Ratio Vegetation Index	Нормализованный относительный вегетационный индекс
RVI / Ratio Vegetation Index	Относительный вегетационный индекс
SAVI / Soil Adjusted Vegetation Index	Вегетационный индекс с коррекцией по почве
TVI / Transformed Vegetation Index	Трансформированный вегетационный индекс
TTVI / Tiam's Transformed Vegetation Index	Трансформированный вегетационный индекс Тиамы
Band 1 / Visible Blue	Канал B1 – Видимый синий диапазон
Band 2 / Visible Green	Канал B2 – Видимый зеленый диапазон
Band 3 / Visible Red	Канал B3 – Видимый красный диапазон
Band 4 / Near-Infrared	Канал B4 – Ближний инфракрасный диапазон
Band 5 / Near-Infrared	Канал B5 – Ближний инфракрасный диапазон
Band 7 / Mid-Infrared	Канал B7 – Средний инфракрасный диапазон
Предикторы, характеризующие почву (SoilGrids, для свойств почв в слое 0-30 см) (Poggio et al., 2021). Исходное разрешение 250×250 м, в настоящей работе преобразовано к 30×30 м	
pH / Acidity	pH водной суспензии (кислотность)
NITRO / Total nitrogen	Содержание общего азота
SOSC / Soil Organic Carbon Stocks	Запасы органического углерода в почве
CEC / Cation Exchange Capacity	Емкость катионного обмена
CLAY / Clay content	Содержание ила (глины)
SILT / Silt content	Содержание пыли
SAND / Sand content	Содержание песка
BD / Soil Bulk Density	Плотность сложения почв
Soil Groups	Почвенные группы по WRB 2006

Для предикторов, характеризующих свойства почв по данным SoilGrids (табл. 1), было выполнено усреднение путём суммирования растров по глубинам (0–5, 5–15, 15–30 см) и деления на 3.

Растровые карты таких климатических показателей как SAT10, EVAP, CC, CH, AgPot (табл. 1) составлены с использованием архивных картографических данных (Сляднев, 1978; Гопп, 2023).

Большая часть растровых карт предикторов имеет разрешение 30×30 м, поэтому разрешение карт с отличным от этого разрешением приведено к 30×30 м. Преобразование разрешения растровых карт касалось климатических данных WorldClim с разрешением 1×1 км (Fick, Hijmans, 2017) и глобальных почвенных данных SoilGrids с разрешением 250×250 м (Poggio et al., 2021). Такая практика использования глобальных данных является допустимой в работах подобного плана, так как более детальных карт в лучшем разрешении не существует. Трудоемкость получения данных среднего и высокого разрешения с необходимым временным усреднением для большой территории приводит к необходимости использования растровых карт с низким пространственным разрешением, которое преобразовывают согласно техническому требованию проекта. В качестве примера совместного использования геоданных с низким и средним разрешением можно привести недавно опубликованную работу ученых из Западной Амазонии (Moquedace et al., 2024). В настоящей работе при изменении разрешения растровых карт климатических показателей WorldClim применена интерполяция B-Spline, которая позволяет сделать плавный переход между значениями пикселей. При изменении разрешения глобальных почвенных данных SoilGrids применена интерполяция методом ближайшего соседа (Nearest Neighbour), которая присваивает вновь созданным пикселям значение исходного пикселя.

Размеры пикселя определяют предел детальности снимка, а окончательный масштаб выбирается исходя из целей картографирования. Снимки с разрешением 30×30 м подходят для составления среднемасштабных и мелкомасштабных карт с масштабом от 1:50 000 и мельче. Однако, согласно руководству по почвенной съемке (Евдокимова, 1987), при работе с почвами, для которых обязательно

должны быть представлены полевые данные, масштаб карты зависит не только от разрешения снимков, но и от детальности полевой почвенной съемки, а именно от количества гектаров, приходящихся на один почвенный разрез. Аналогичный подход можно применить при составлении карт почвенных свойств. Таким образом, если площадь территории в гектарах разделить на количество разрезов, то мы получим количество гектаров, приходящихся на 1 разрез. В настоящем исследовании на 1 почвенный разрез приходится 1679 га, что, согласно руководству (Евдокимова, 1987), определяет предел детальности карт, соответствующий масштабу 1:200 000. Это означает, что публикация карт в аналоговой (бумажной) форме в данном масштабе является научно обоснованной, поскольку её содержание полностью обеспечено достаточным количеством полевых данных.

Для моделирования содержания ПОУ с использованием алгоритма RF созданы обучающий (ОНД, $n=206$) и валидационный (ВНД, $n=55$) наборы почвенных данных. ВНД использовался для оценки эффективности моделирования содержания ПОУ и сформирован с учётом минимальных, средних и максимальных значений ПОУ в каждом типе и подтипе почв. Гармонизация данных по содержанию ПОУ из разных интервалов глубин почвенного профиля до заданной мощности слоя (0–30 см) выполнена с использованием пакета *mpspline2* в программном обеспечении R (O'Brien, 2022).

Математические операции с геопривязанными растровыми картами выполнены на облачной онлайн-платформе Google Earth Engine (Gorelick et al., 2017) и в программном обеспечении SAGA GIS (Conrad et al., 2015). Карты содержания ПОУ созданы с использованием алгоритма Random Forest (RF), реализованного в виде скрипта (кода) на онлайн-платформе Google Earth Engine, со следующими настройками: 200 и 1000 «деревьев», остальные значения гиперпараметров по умолчанию. Термин «дерево» в концепции алгоритма RF означает независимый алгоритм, обучающийся на случайной подвыборке данных и подмножестве признаков. Алгоритм RF позволяет создавать множество «деревьев»; дальнейшее объединение результатов работы разных независимых «деревьев» или алгоритмов даёт более точные и устойчивые предсказания для зависимой/целевой переменной. Конечный результат – среднее значение предсказаний по всем созданным «деревьям» (Breiman, 2001).

Моделирование с использованием алгоритма RF проводилось на основе разного количества предикторов в наборе в зависимости от коэффициента корреляции (**R**) между содержанием ПОУ и значениями предикторов: 1) все 80; 2) 50 (все с $R \geq \pm 0,1$); 3) 42 (все с $R \geq \pm 0,2$); 4) 26 (все с $R \geq \pm 0,3$); 5) 15 (все с $R \geq \pm 0,4$). Также было проведено картографирование содержания ПОУ с использованием пяти наборов предикторов, характеризующих отдельные факторы почвообразования согласно модели SCORPAN (почва (S), климат (C), растительность (O), рельеф (R), пространственное положение (N)), и одного набора, объединяющего по 5 доминирующих показателей от каждого фактора почвообразования. Таким образом, всего было использовано 11 наборов предикторов, представленных как в сочетании показателей от разных факторов почвообразования, так и по отдельным группам.

Эффективность моделирования оценена с использованием коэффициента детерминации (R^2), который определен по графику зависимости между фактическими и предсказанными значениями для двух наборов данных: ОНД ($R^2_{\text{онд}}$) и ВНД ($R^2_{\text{внд}}$). В общем случае $R^2=1$ означает, что модель идеально эффективна для моделирования, а $R^2=0$ – модель объясняет вариацию изучаемого показателя не лучше, чем среднее значение. $R^2_{\text{онд}}$ показывает, насколько качественно модель аппроксимирует фактические данные, на которых она обучалась. $R^2_{\text{внд}}$ отражает предсказательную способность модели для новых данных, т.е. тех данных, которые не использовали в обучающем наборе данных. $R^2_{\text{внд}}$ рассчитан путем наложения ВНД на карту содержания ПОУ, построенную на основе ОНД, с последующим извлечением предсказанных значений в геопривязанных точках расположения профилей почв и построением графика зависимости между предсказанными и фактическими значениями. Кроме этого, эффективность моделирования оценена на основе ВНД с использованием следующих показателей: корень из среднеквадратической ошибки (**RMSE**, *англ.* Root Mean Square Error), средняя абсолютная процентная ошибка (**MAPE**, *англ.* Mean Absolute Percentage Error) и средняя абсолютная ошибка (**MAE**, *англ.* Mean Absolute Error).

Корень из среднеквадратической ошибки рассчитан следующим образом (Hengl, 2007):

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=0}^n [\hat{z}(sj) - z(sj)]^2}$$

где n – количество наблюдений в анализируемом наборе данных, $z(sj)$ – прогнозируемое значение, $\hat{z}(sj)$ – наблюдаемое значение. Данные представлены в единицах измерения содержания ПОУ (%).

Значения RMSE, близкие к нулю, указывают на идеальное соответствие модели фактическим данным. Значение RMSE, составляющее менее половины стандартного отклонения измеренных данных, указывает на высокий уровень точности прогноза (Singh et al., 2004).

Средняя абсолютная процентная ошибка MAPE рассчитана следующим образом (Афанасьев, Цыпин, 2008):

$$MAPE = \frac{1}{n} \sum \frac{|y_t - \bar{y}_t|}{y_t} \times 100,$$

где y_t – фактическое значение, $\bar{y}_t$ – прогнозируемое значение, n – количество наблюдений в анализируемом наборе данных.

Точность моделирования оценивалась по значениям MAPE следующим образом (Афанасьев, Цыпин, 2008): <10% – высокая; 10–20% – хорошая; 20–50% – удовлетворительная; >50% – неудовлетворительная точность.

Среднее отклонение прогнозируемых значений от фактического значения было рассчитано с использованием средней абсолютной ошибки MAE:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_t - \bar{y}_t|,$$

где y_t – фактическое значение; $\bar{y}_t$ – прогнозируемое значение; n – количество наблюдений в анализируемом наборе данных. $|y_t - \bar{y}_t|$ – абсолютное значение (без отрицательного знака).

Блок-схема методологии цифрового картирования содержания ПОУ представлена на рисунке 2.

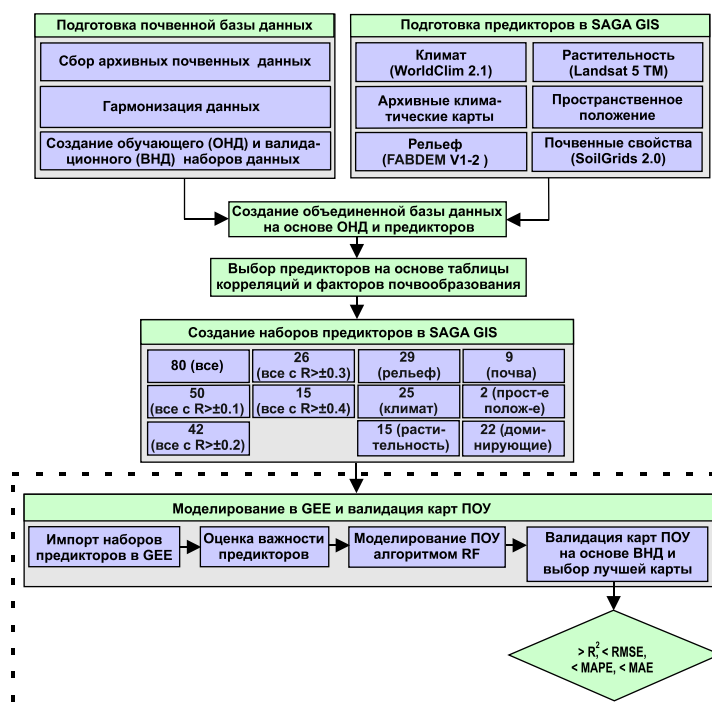


Рисунок 2. Блок-схема методики цифрового картографирования содержания ПОУ в почвах.

На рисунке 1 показано поле площадью 112 га, на котором в 2015 году отобраны почвенные образцы ($n=50$) из слоя 0–30 см. Полученные данные по содержанию ПОУ в этих образцах почв не использовались при обучении моделей и картографировании и были предназначены для визуального сравнения результатов моделирования (1984–1994 гг.) с современными (2015 г.) данными путём наложения последних на карту ПОУ. Такое наложение позволяет приблизительно оценить возможные изменения в содержании ПОУ с течением времени. Важно отметить, что данные 2015 года нельзя использовать для количественной оценки точности карты, построенной по данным 30-летней давности, так как за это время свойства почвы могли существенно измениться.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ОБСУЖДЕНИЕ

По фактическим данным содержание ПОУ варьировало в пределах 1,2–13,9% (табл. 2). Коэффициент вариации во всех наборах данных составил более 40%, что указывает на сильную

вариабельность. Значения коэффициента асимметрии (> 1) показывают, что общий и обучающий наборы данных имеют сильную правостороннюю асимметрию, а валидационный – слабую, что указывает на более однородные данные или отсутствие значительного количества экстремальных значений (выбросов) по сравнению с общим и обучающим наборами данных. Значения коэффициента эксцесса для общего и обучающего наборов данных указывают на существенное отклонение от нормального распределения; для валидационного набора данных распределение близко к нормальному. При моделировании алгоритмом RF допускается использование данных с аномальным распределением, поэтому логарифмирование данных обучающего набора не проводилось.

Таблица 2

Статистика по содержанию ПОУ для общего, обучающего и валидационного наборов данных

Показатель	Наборы данных		
	Общий ($n=261$)	Обучающий ($n=206$)	Валидационный ($n=55$)
Среднее, %	4,6	4,6	4,5
Стандартное отклонение, %	$\pm 2,0$	$\pm 2,1$	$\pm 1,8$
Минимум, %	1,2	1,3	1,2
Медиана, %	4,4	4,4	4,4
Максимум, %	13,9	13,9	10,0
Коэффициент вариации, %	43,9	44,8	40,0
Коэффициент асимметрии	1,32	1,45	0,45
Коэффициент эксцесса	3,49	3,84	0,17

Важность предикторов для цифрового картографирования содержания ПОУ оценили с использованием корреляционного анализа (табл. 3) и показателя относительной степени важности (индекс IncNodePurity/Джинни) предикторов (рис. 3).

Таблица 3

Коэффициенты корреляции Спирмена (R) между содержанием ПОУ и значениями предикторов, характеризующих факторы почвообразования

Предиктор	R	Предиктор	R	Предиктор	R
BIO1	0,38**	LS	–	LONG	-0,35**
BIO2	-0,22**	Aspect	–	LAT	-0,61**
BIO3	-0,32**	CrosSecCurv	–	NDVI	–
BIO4	0,50**	FlowLineCurv	–	CTVI	–
BIO5	0,41**	GenCurv	–	DVI	–
BIO6	0,23**	LongiCurv	-0,20**	NRVI	–
BIO7	0,24**	MinCurv	–	RVI	–
BIO8	0,42**	MaxCurv	–	SAVI	–
BIO9	-0,26**	ProfCurv	-0,19**	TVI	–
BIO10	0,42**	PlanCurv	–	TTVI	–
BIO11	–	TanCurv	–	Band 1	-0,24**
BIO12	-0,40**	TotCurv	–	Band 2	-0,31**
BIO13	-0,27**	ConvIndex	–	Band 3	-0,26**
BIO14	-0,42**	Texture	-0,28**	Band 4	-0,26**
BIO15	0,31**	Convexity	-0,22**	Band 5	-0,18**
BIO16	-0,41**	MassBalans	–	Band 7	–
BIO17	-0,33**	AnalitHill	–	pH	0,17*
BIO18	-0,41**	ValDepth	–	NITRO	0,47
BIO19	-0,24**	TWI	0,26**	SOCS	–
SRmean	0,51**	TCA	0,14*	CEC	0,34**
SAT10	0,23**	RSP	-0,14*	CLAY	-0,38**
CC	0,49**	TPI	-0,17	SILT	–
EVAP	0,41**	TRI	-0,19**	SAND	0,31**
CH	-0,42**	CND	-0,36**	BD	-0,34**
AgPot	-0,55**	CNBL	–	SoilGroups	–
ELEV	-0,21**	VisSky	–	LandCov	–
Slope	-0,20**	SkyViewFact	–		

Примечание: ** корреляция значима при $p < 0,01$; * $p < 0,05$; прочерк – корреляция отсутствует или не значима. Интерпретация сокращённого названия предикторов приведена в табл. 1.

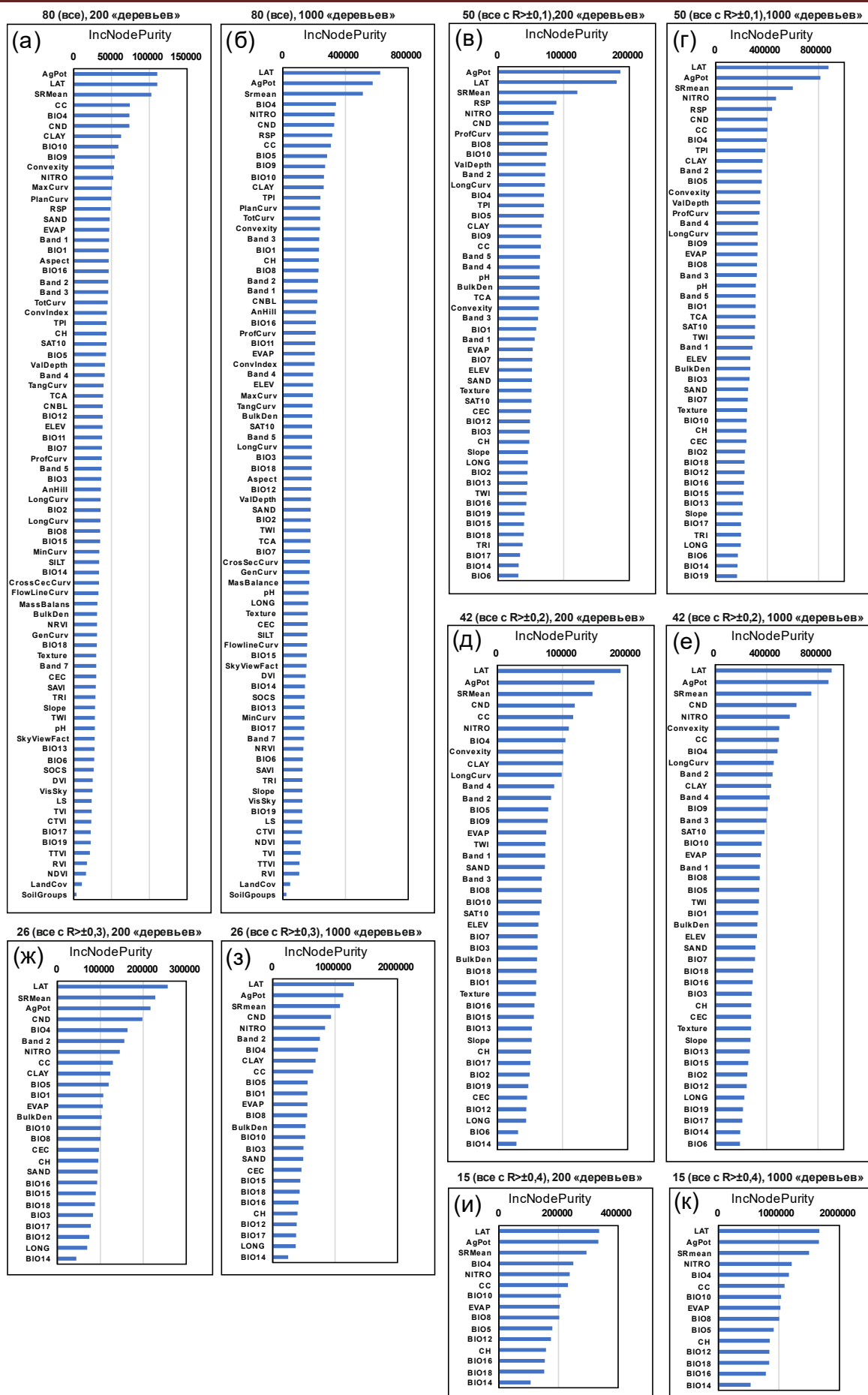


Рисунок 3. Относительная степень важности (индекс IncNodePurity) предикторов в моделях с разным количеством «деревьев» и предикторов в наборе. Обозначения предикторов см. в табл. 1.

Корреляционный анализ показал, что из 80 потенциальных переменных 25 были наиболее значимыми ($R > 0,3$) для картографирования содержания ПОУ (табл. 3). Умеренная положительная корреляция (R в диапазоне от 0,3 до 0,5) выявлена между содержанием ПОУ и следующими переменными: BIO1, BIO4, BIO5, BIO8, BIO10, SRmean, CC, EVAP, NITRO, CEC, SAND. Умеренная отрицательная корреляция (R в диапазоне от -0,3 до -0,5) установлена со следующими переменными: BIO3, BIO12, BIO14, BIO16, BIO17, BIO18, CH, AgPot, CND, LONG, LAT, Band 2, CLAY, BD (табл. 3). Обращает на себя внимание отсутствие корреляции между содержанием ПОУ и его запасами по данным SoilGrids (табл. 3). Анализ карты запасов ПОУ (SoilGrid) показал, что данные сильно усреднены, а карта имеет слабую дифференциацию по значениям запасов ПОУ; следовательно, эта особенность является причиной отсутствия корреляции между вышеупомянутыми показателями. При моделировании алгоритмом RF количество используемых предикторов не ограничено. Для сравнительного анализа были выбраны предикторы с разным их количеством в наборе (табл. 3): 1) 80 (все); 2) 50 (все с $R > \pm 0,1$); 3) 42 (все с $R > \pm 0,2$); 4) 26 (все с $R > \pm 0,3$); 5) 15 (все с $R > \pm 0,4$). Относительная степень важности отдельных переменных в моделях с вышеуказанными наборами предикторов показана на рисунке 3. Чем выше значение чистоты узла в «деревьях», определяемого по индексу IncNodePurity/Джинни, тем больше вклад переменной в объясняющую способность модели. Анализ графиков показывает, что ранжирование значимости предикторов определяется не по более высоким значениям коэффициентов корреляции между исследуемыми показателями (табл. 3, рис. 3). Например, предикторы PlanCurv и MaxCurv не коррелировали с содержанием ПОУ, однако на графике видно, что степень их важности выше, чем у CH, у которого коэффициент корреляции с ПОУ равен -0,42 (рис. 3 а, б). Также стоит отметить, что предикторы с R менее $\pm 0,3$ (Convexity, MaxCurv, RSP, Band 1 и др.) по степени важности были выше тех, у которых R был намного больше $\pm 0,3$ (рис. 3 а, б). Данная ситуация может указывать на наличие более значимых зависимостей между изучаемыми переменными в отдельных «деревьях» моделей; однако, эксперту невозможно определить силу их корреляционной зависимости, в связи с отсутствием понимания устройства модели, которую в машинном обучении называют «черный ящик», то есть она рассматривается как система, имеющая некий вход для ввода информации и выход для отображения результатов работы, при этом принципы взаимодействия её компонентов и происходящие в ходе работы системы процессы эксперту неизвестны. Увеличение количества «деревьев» до 1000 привело к незначительному изменению положения отдельных предикторов на графиках (рис. 3 б). Во всех рассматриваемых наборах предикторов доминировали предикторы, характеризующие климат и пространственное положение (AgPot, LAT, SRmean); затем располагались предикторы, характеризующие почвенные свойства, рельеф и растительность (рис. 3).

В порядке убывания $R^2_{\text{ОНД}}$, модели с разным количеством предикторов расположились следующим образом (количество в наборе): 80 – 50 – 42 – 26 – 15 (табл. 4).

Таблица 4

Эффективность моделирования содержания ПОУ алгоритмом RF с различным количеством «деревьев» и наборов предикторов

Наборы предикторов (см. табл. 1, 3)	Показатели эффективности моделирования				
	$R^2_{\text{ОНД}}$	$R^2_{\text{ВНД}}$	RMSE _{ВНД} , %	MAPE _{ВНД} , %	MAE _{ВНД} , %
200 «деревьев» в модели					
80 (все предикторы)	0,84	0,52	1,29	28,5	0,96
50 (все с $R > \pm 0,1$)	0,84	0,51	1,29	28,7	0,98
42 (все с $R > \pm 0,2$)	0,83	0,55	1,25	27,6	0,95
26 (все с $R > \pm 0,3$)	0,82	0,54	1,25	27,6	0,95
15 (все с $R > \pm 0,4$)	0,78	0,46	1,33	26,7	0,97
1000 «деревьев» в модели					
80 (все предикторы)	0,85	0,54	1,27	28,5	0,95
50 (все с $R > \pm 0,1$)	0,84	0,53	1,28	28,4	0,97
42 (все с $R > \pm 0,2$)	0,83	0,56	1,24	27,3	0,94
26 (все с $R > \pm 0,3$)	0,83	0,54	1,25	27,6	0,94
15 (все с $R > \pm 0,4$)	0,78	0,46	1,33	26,5	0,96

Примечание. $R^2_{\text{ОНД}}$ – коэффициент детерминации, рассчитанный по обучающему набору данных (ОНД); RMSE_{ВНД} – корень из среднеквадратической ошибки, рассчитанный по валидационному набору данных (ВНД); MAPE_{ВНД} – средняя абсолютная ошибка в процентах, рассчитанная по ВНД; MAE_{ВНД} – средняя абсолютная ошибка, рассчитанная по ВНД.

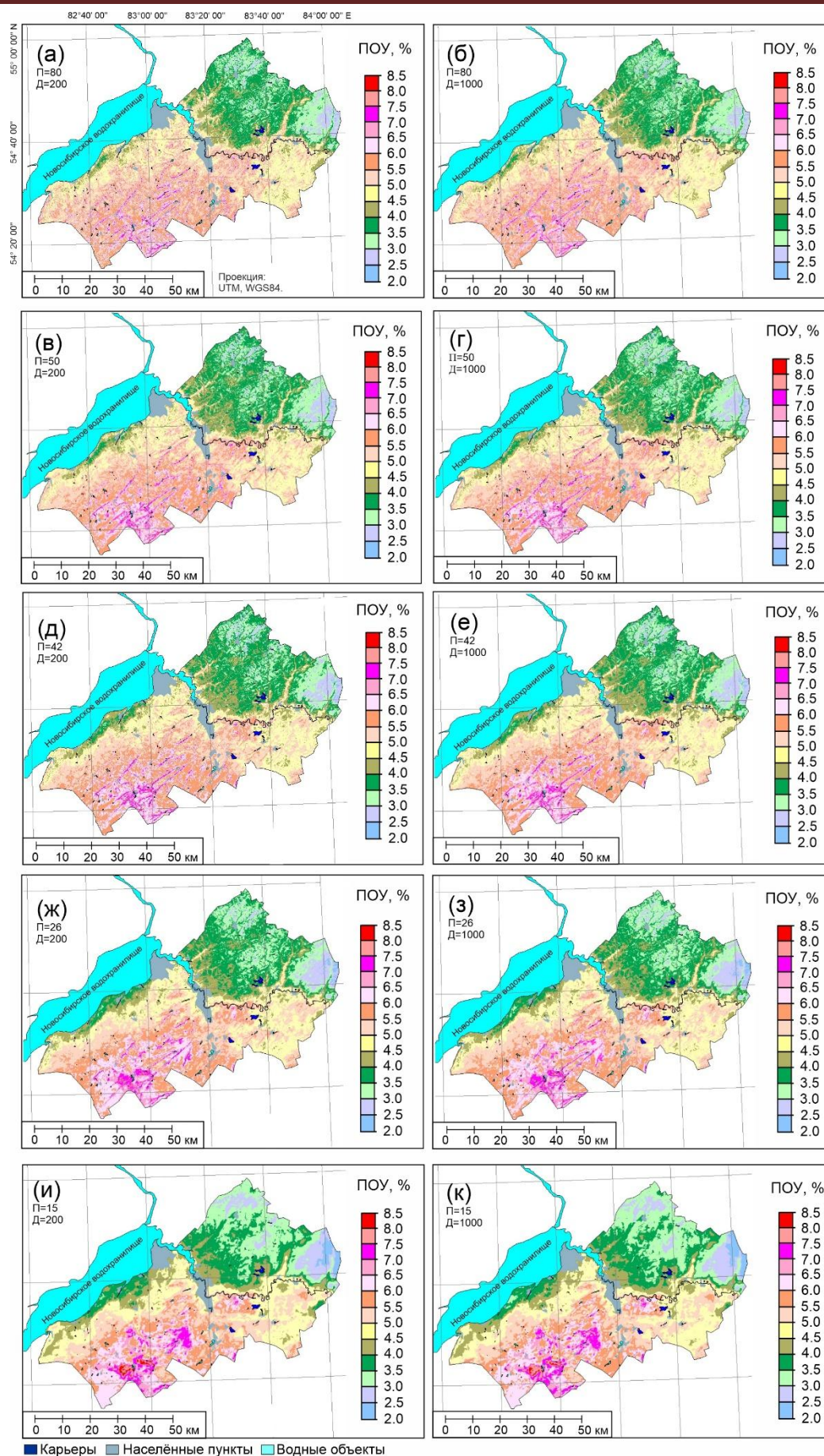


Рисунок 4. Карты содержания ПОУ (по обучающим данным 1984–1994 гг.) в слое 0–30 см, составленные на основе различного количества предикторов и «деревьев». Сокращения: П – количество предикторов, Д – количество «деревьев». Разрешение карт 30×30 м.

Распределение моделей в порядке убывания $R^2_{\text{внд}}$ было следующим (количество предикторов в наборе): 42 – 26 – 80 – 50 – 15 (табл. 4). На основании более высоких значений $R^2_{\text{внд}}$ (0,55) и более низких значений RMSE (1,25), карта содержания ПОУ, составленная с использованием 42 предикторов, выбрана в качестве лучшей (табл. 4, рис. 4 д). Увеличение в моделях числа деревьев до 1000 не привело к значительному увеличению $R^2_{\text{онд}}$ и $R^2_{\text{внд}}$ (табл. 4).

Сравнение фактических и предсказанных данных показало, что диапазон размаха без выбросов по содержанию ПОУ отличался значительно, однако отличия средних не выявлено (рис. 5). Предсказанный ПОУ во всех моделях имел меньший диапазон вариации по сравнению с фактическими данными (рис. 5).

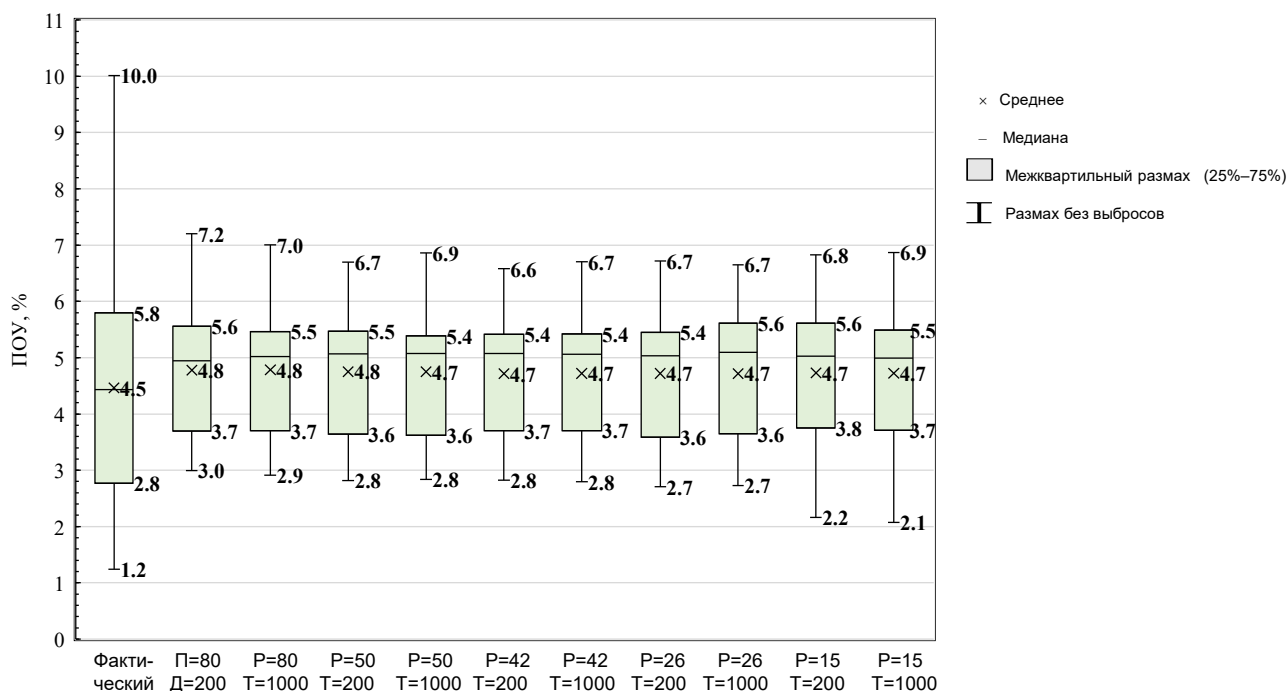


Рисунок 5. Сравнение фактического и предсказанного алгоритмом RF содержания ПОУ (по данным 1984–1994 гг.) в валидационном наборе данных ($n=55$). Сокращения: П – количество предикторов; Д – количество «деревьев» в модели.

Увеличение масштаба построенной карты (рис. 4д) на участке, где имелся набор по содержанию ПОУ за 2015 год с близким расположением точек отбора проб, и наложение этих данных на карту показало приемлемое сходство с моделью, построенной по данным 1984–1994 годов, так как рядом с точкой отбора (красный крест на карте) имелись схожие/близкие значения по содержанию ПОУ (черные точки на карте) (рис. 6).

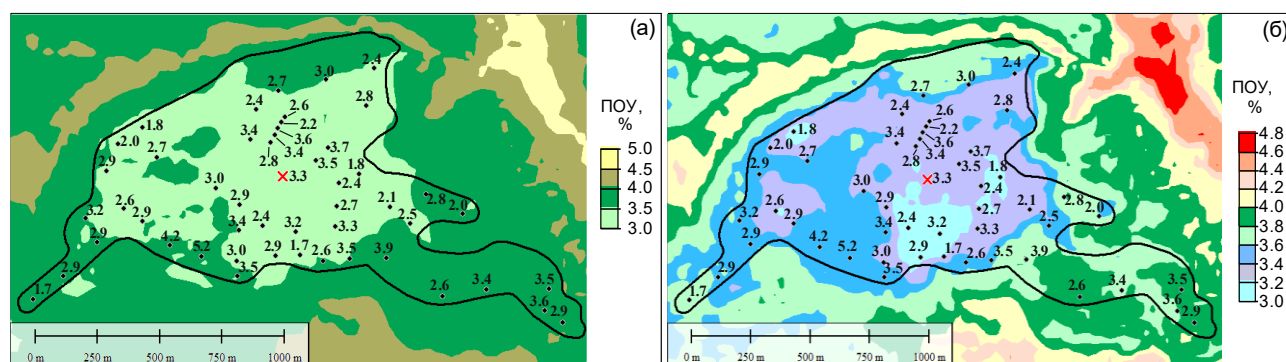


Рисунок 6. Сравнение предсказанного (по обучающим данным 1984–1994 гг.) и фактического (данные за 2015 г.) содержания ПОУ при увеличенном масштабе карты: (а) – шкала с шагом 0,5 по рис. 4д; (б) – шкала с шагом 0,2 по рис. 4д (для более детального рассмотрения). Примечание: цифры на картах возле чёрных точек представляют фактическое содержание ПОУ в 2015 году ($n=50$), красный крест – содержание ПОУ по базе данных 1984–1994 гг.

Однако полученные данные не позволяют доказать произошедшие изменения в содержании ПОУ с течением времени в связи с отсутствием установленного репера в 1984–1994 годах в том месте, где отбиралась почвенная проба (красный крест на карте). Иные варианты получения современных данных пробоотбора из этой точки будут некорректны из-за ошибки определения местоположения с использованием системы глобального позиционирования GPS (Global Positioning System), т.е. несмотря на то, что координаты точки известны, найти это место на местности без ошибки невозможно. Необходимо подчеркнуть, что данные 2015 года не использовались в процессе обучения моделей и картографирования содержания ПОУ алгоритмом RF. Они использовались только для наложения на карту ПОУ, созданную по данным 1984–1994 гг., чтобы качественно/визуально оценить объясняющую способность модели, построенную с помощью алгоритма RF, и возможные изменения в содержании ПОУ.

Содержание ПОУ увеличивается в юго-западном направлении (рис. 4д). Предсказанное среднее значение по содержанию ПОУ в 1,1–1,2 раза превышало среднее фактическое для следующих типов и подтипов почв: 1 – луговые аллювиальные / Eutric Fluvisols; 2 – луговые карбонатные / Eutric Gleysols; светло-серые лесные оподзоленные / Luvic Greyzemic Phaeozems; 8 – серые лесные оподзоленные / Luvic Greyzemic Phaeozems (рис. 7).

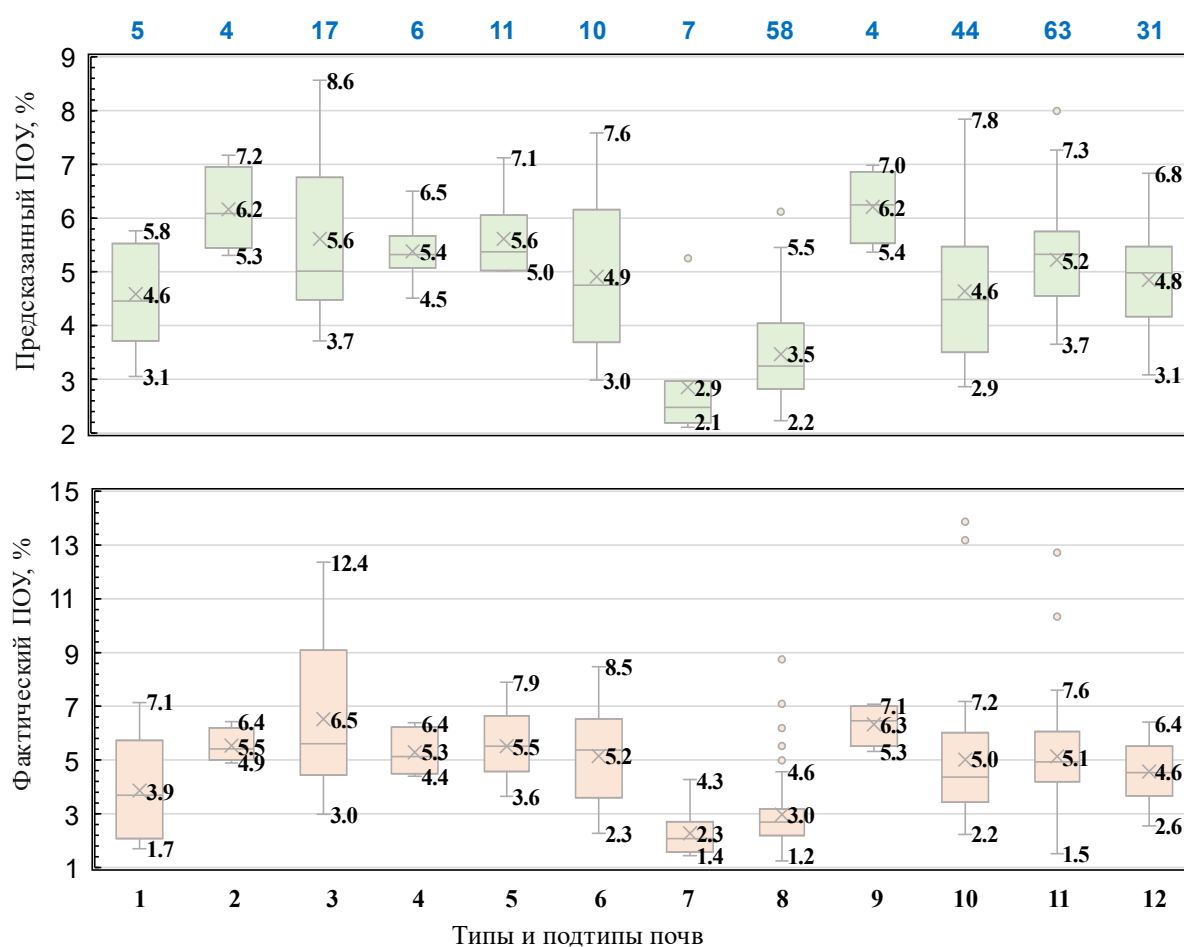


Рисунок 7. Вариабельность предсказанного (по рис. 4д) и фактического содержания ПОУ (по данным 1984–1994 гг.) в слое 0–30 см в зависимости от таксономических единиц почв ($n=261$). Условные обозначения почвенных таксономических единиц по классификации и диагностике почв СССР (Классификация ..., 1977) / мировой реферативной базе данных почвенных ресурсов WRB (IUSS Working Group ..., 2015): 1 – луговые аллювиальные / Eutric Fluvisols; 2 – луговые карбонатные / Eutric Gleysols; 3 – луговые подзолистые / Haplic Gleysols; 4 – луговые солончаковатые / Haplic Gleysols (Protosalic); 5 – лугово-черноземные обыкновенные / Gleyic Chernozems; 6 – лугово-черноземные оподзоленные / Greyzemic Gleyic Chernozems; 7 – светло-серые лесные оподзоленные / Luvic Greyzemic Phaeozems; 8 – серые лесные оподзоленные / Luvic Greyzemic Phaeozems; 9 – солонцы луговые / Gleyic Solonetz; 10 – темно-серые лесные оподзоленные / Luvic Greyzemic Phaeozems; 11 – чернозёмы выщелоченные / Luvic Chernozems; 12 – чернозёмы оподзоленные / Luvic Greyzemic Chernozems. Примечание: объем выборки обозначен синими цифрами сверху над диаграммами размаха.

Предсказанные средние значения по содержанию ПОУ были занижены в 1,1–1,2 раза по сравнению со средними фактическими для следующих типов и подтипов почв: 3 – луговые подзолистые / Haplic Gleysols; 6 – лугово-черноземные оподзоленные / Greyzemic Gleyic Chernozems; 10 – тёмно-серые лесные оподзоленные / Luvis Greyzemic Phaeozems (рис. 7). Для остальных типов и подтипов почв соответствие предсказанных и фактических средних значений удовлетворительное. Сравнение содержания ПОУ разных таксономических единиц почв показало, что, исходя из фактических данных, самые низкие средние значения характерны для светло-серых (№ 7) и серых лесных оподзоленных (№ 8) почв (рис. 7). Относительно высокие средние значения (>5,5%) ПОУ выявлены в следующих почвах: 3 – луговые подзолистые / Haplic Gleysols; 9 – солонцы луговые / Gleyic Solonchets; 2 – луговые карбонатные / Eutric Gleysols; 5 – лугово-черноземные обыкновенные / Gleyic Chernozems.

Учитывая высокую важность предикторов в цифровом картографировании содержания ПОУ, дополнительно проведен анализ их относительной степени важности по отдельным факторам почвообразования для того, чтобы понять какой вклад в итоговое изображение они вносят (рис. 8). Данный анализ также необходим для исключения некоторых климатических показателей, несмотря на то, что корреляции с большинством из них высокие и умеренные (табл. 3). Преобладание в модели предикторов с низким исходным разрешением может привести к недооценке других, не менее важных предикторов, имеющих лучшее исходное разрешение.

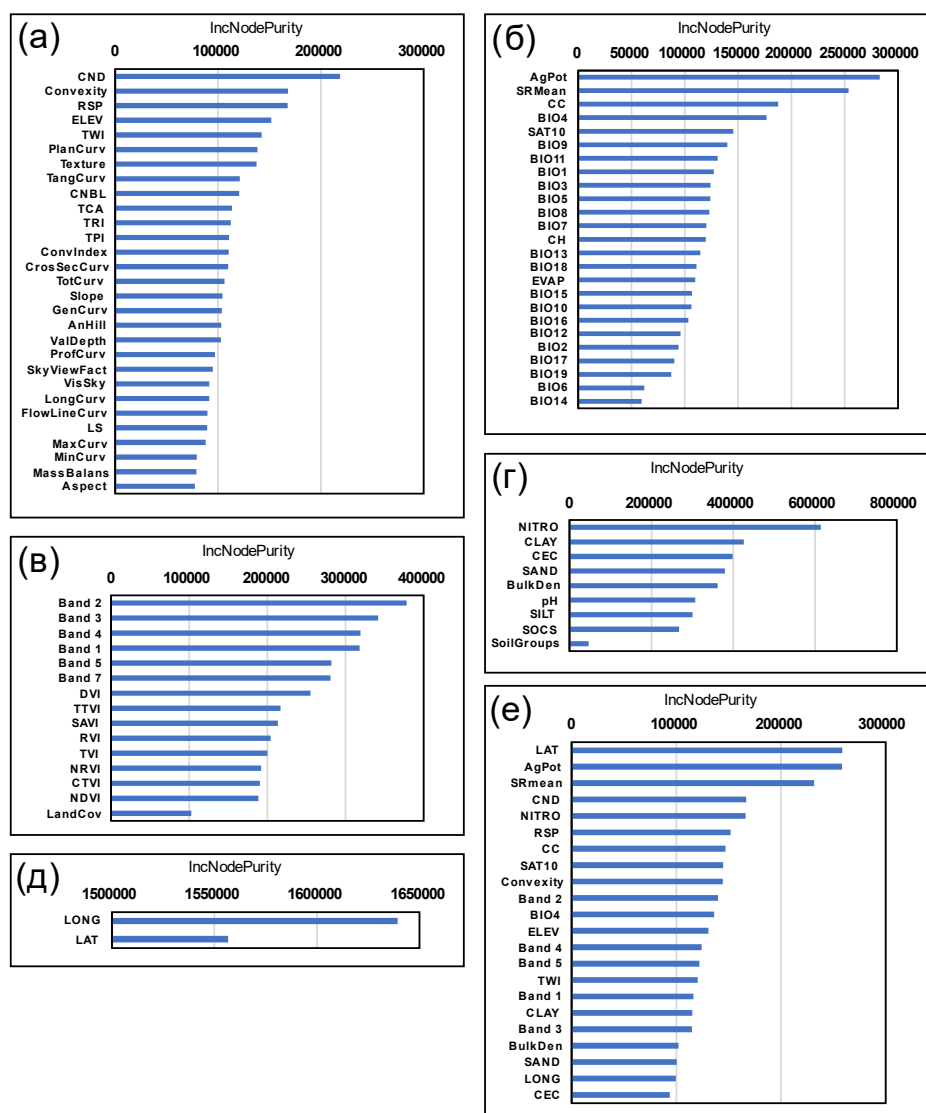


Рисунок 8. Относительная степень важности (индекс IncNodePurity) предикторов в моделях, построенных с использованием алгоритма RF и наборов предикторов, характеризующих разные группы факторов почвообразования (количество и в скобках название группы): а – 29 (рельеф); б – 25 (климат); в – 15 (растительность); г – 9 (почва); д – 2 (пространственное положение); е – 22 (доминирующие предикторы: по 5 первых из каждой группы (на основе рис. 8 а, б, в, г) + 2 (рис. 8 д)).

По 5 первых (доминирующих) предикторов из разных групп факторов почвообразования для картографирования содержания ПОУ выбраны следующие: рельеф – CND, Convexity, RSP, ELEV, TWI (рис. 8а); климат – AgPot, SRmean, CC, BIO4, SAT10 (рис. 8б); растительность – Band 2, Band 3, Band 4, Band 1, Band 5 (рис. 8в); почва – NITRO, CLAY, CEC, SAND, BulkDen (рис. 8г); пространственное положение – LONG, LAT (рис. 8д). Относительная степень важности вышеперечисленных предикторов в совмещенном наборе показывает, что лидирующее положение занял предиктор, характеризующий пространственное положение (LAT); далее по списку представлены предикторы, характеризующие климат, рельеф, почву (рис. 8е).

Обращают на себя внимание предикторы, характеризующие пространственное положение, а именно, при совместном использовании разных групп предикторов (рис. 3, 9 е) такой показатель, как LAT, занимал лидирующие позиции по сравнению с LONG. Однако, степень важности LAT, оцененная только в индивидуальной группе (пространственное положение), была ниже LONG (рис. 9 д, е). Таким образом, в зависимости от состава предикторов в наборе, относительная степень их важности может кардинально меняться.

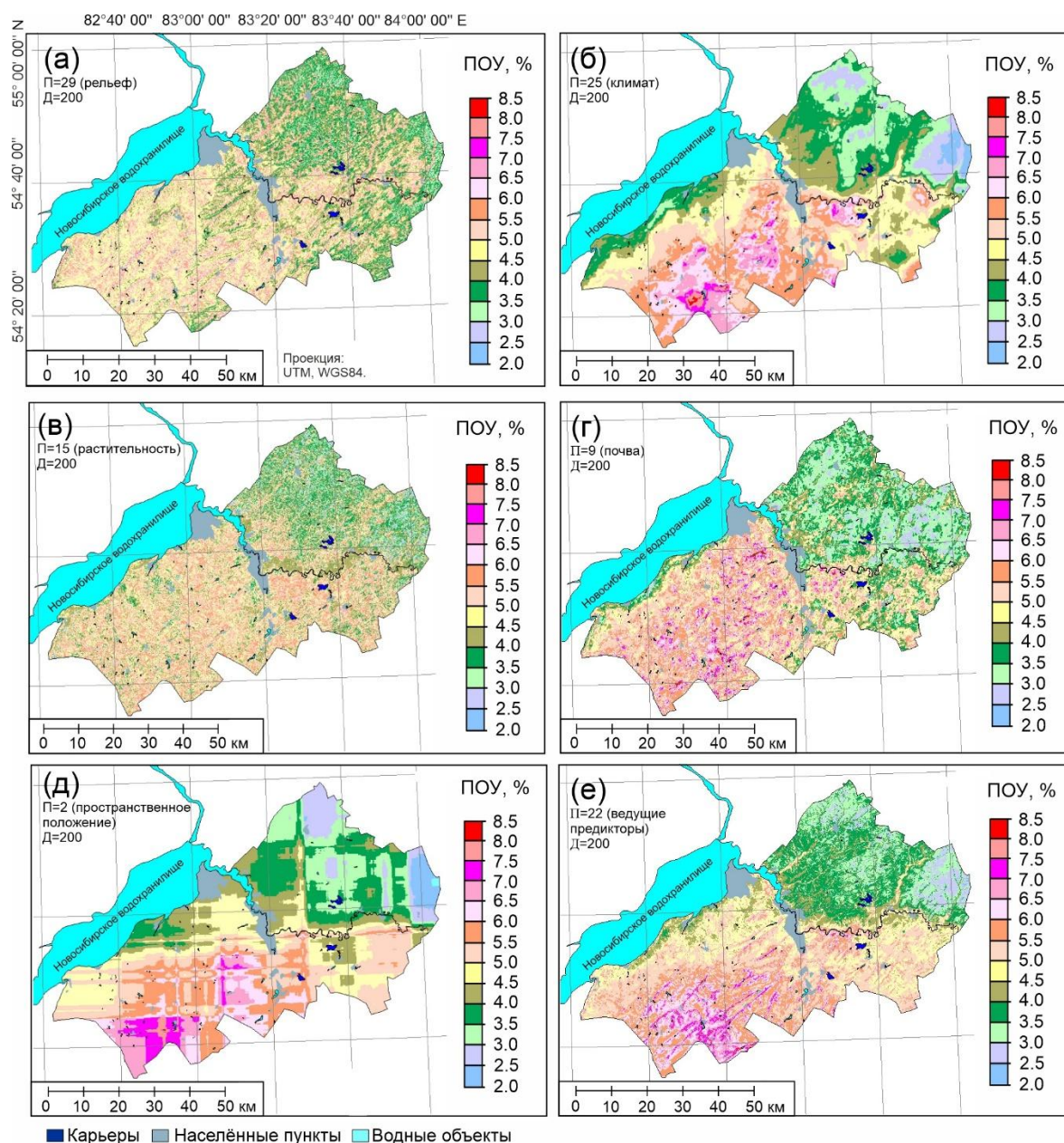


Рисунок 9. Карты содержания ПОУ (по данным 1984–1994 гг.) в слое 0–30 см, составленные на основе предикторов, характеризующих отдельные и совмещенные факторы почвообразования: а – 29 (рельеф); б – 25 (климат); в – 15 (растительность); г – 9 (почва); д – 2 (пространственное положение); е – 22 (все доминирующие предикторы: по 5 первых из каждой группы (рис. 8 а, б, в, г + 2 (рис. 8 д)). Сокращения: П – количество предикторов, Д – количество деревьев в моделях.

Среди климатических предикторов, интегральные показатели (AgPot, CC) занимали доминирующие позиции, в то время как среднегодовое количество осадков (BIO12) по степени важности находилось на более низких позициях в списке (рис. 8б). У предикторов, характеризующих растительность, отражательная способность в различных каналах занимала доминирующие позиции, а вегетационные индексы по степени важности находились ниже (рис. 8в).

На основе перечисленных выше групп предикторов (рис. 8) построены карты содержания ПОУ с использованием алгоритма RF (рис. 9).

Эффективность моделирования содержания ПОУ алгоритмом RF с использованием наборов предикторов, указанных на рис. 8, представлена в табл. 5. Модели с разным количеством предикторов в порядке убывания $R^2_{\text{онд}}$ расположились следующим образом: 29 (рельеф) – 22 (доминирующие) – 9 (почва) – 25 (климат) – 15 (растительность) – 2 (пространственное положение). Распределение моделей в порядке убывания $R^2_{\text{внд}}$ было следующим (количество предикторов в наборе): 22 (доминирующие) – 25 (климат) – 9 (почва) – 2 (пространственное положение) – 29 (рельеф) – 15 (растительность) (табл. 5). $R^2_{\text{внд}}$ по сравнению с $R^2_{\text{онд}}$ у модели «29 (рельеф)» снизился в 11 раз, а у модели «15 (растительность)» – в 24 раза, что говорит о сильном снижении эффективности моделирования с использованием этих наборов предикторов, несмотря на то, что значения $R^2_{\text{онд}}$ у этих моделей были относительно высокими (табл. 5). Значение $\text{RMSE}_{\text{внд}}$ у модели «15 (растительность)» больше стандартного отклонения (1,8) в валидационном наборе данных (табл. 2, 5), что свидетельствует о неудовлетворительном качестве моделирования.

Таблица 5

Эффективность моделирования алгоритмом RF (200 «деревьев») с наборами предикторов, характеризующих отдельные и совмещенные факторы почвообразования

Количество предикторов (см. табл. 1)	Показатели эффективности моделирования				
	$R^2_{\text{онд}}$	$R^2_{\text{внд}}$	$\text{RMSE}_{\text{внд}}$	$\text{MAPE}_{\text{внд}}$	$\text{MAE}_{\text{внд}}$
			%		
29 (рельеф)	0,85	0,08	1,71	40,9	1,29
25 (климат)	0,79	0,47	1,32	26,5	0,96
15 (растительность)	0,71	0,03	1,85	45,7	1,52
9 (почва)	0,81	0,41	1,39	31,2	1,06
2 (пространственное положение)	0,68	0,39	1,44	29,9	1,07
22 (доминирующие предикторы: по 5 первых из каждой группы (рис. 8 а, б, в, г + 2 (рис. 8 д))	0,82	0,54	1,26	28,2	0,94

Наиболее характерно глобальный тренд в изменении содержания ПОУ выражен на картах, построенных с использованием предикторов, характеризующих климат (рис. 9б), почву (рис. 9г), пространственное положение (рис. 9д). На картах содержания ПОУ, составленных на основе наборов предикторов (рельеф, растительность), глобальный тренд выражен с отсечением минимальных и максимальных значений содержания ПОУ (рис. 9 а, в). При этом $R^2_{\text{онд}}$ у этих моделей высокие (0,85 и 0,71), а ошибки, оцененные по показателям эффективности модели ($\text{RMSE}_{\text{внд}}$, $\text{MAPE}_{\text{внд}}$, $\text{MAE}_{\text{внд}}$), значительно больше, чем у моделей, имеющих более низкий $R^2_{\text{онд}}$ (табл. 5). В совмещенном наборе предикторов (22) показатели эффективности моделирования ($\text{RMSE}_{\text{внд}}$, $\text{MAE}_{\text{внд}}$) были самыми лучшими и сравнимы с показателями модели, построенной на основе 42 предикторов (табл. 4, 5). Таким образом, подход с выбором только доминирующих предикторов в отдельных группах факторов почвообразования можно считать приемлемым для картографирования содержания ПОУ.

Предикторы, характеризующие пространственное положение (LAT, LONG), можно использовать только в сочетании с предикторами из других групп факторов почвообразования, так как карта ПОУ, составленная с использованием только этих двух предикторов, выглядит неправдоподобно из-за наличия вертикальных и горизонтальных полос (рис. 10д). Однако, в сочетании с другими предикторами, такой предиктор, как LAT, значительно улучшает показатели эффективности моделирования, так как занимает лидирующую позицию по степени важности (рис. 3, 9е).

Следующий этап нашего исследования состоял в увеличении на экране компьютера масштаба построенной карты (рис. 4д) на участке, где имелся набор фактических данных по содержанию ПОУ за 2015 год с близким расположением точек отбора проб (рис. 10). Наложение этих более детальных данных на итоговые карты ПОУ позволяет лишь приблизительно оценить качество моделирования и произошедшие временные изменения в содержании ПОУ. Такой подход относится к типу разведочного анализа данных или проверке предположений. Увеличение масштаба карт и наложение

более подробных данных, пусть даже не соответствующих необходимому временному периоду, позволяет сделать важные выводы относительно дальнейшего использования предикторов, характеризующих разные факторы почвообразования. На увеличенных картах (рис. 10 б, г, д) видно, что предикторы из отдельных групп (климат, почва, пространственное положение) не способны предсказывать аккумулятивный эффект (увеличение содержания ПОУ в различного рода понижениях рельефа). Тогда как на картах содержания ПОУ, составленных с использованием предикторов от других факторов почвообразования (рельеф, растительность), аккумулятивный эффект отчетливо просматривается в виде ареалов, повторяющих формы понижений (за границами поля со всех сторон имеются балки, лога, долины рек и др.), в которых наблюдается увеличение содержания ПОУ (рис. 10 а, в). Данное исследование не позволяет визуально показать (нет точек отбора проб за пределами поля в понижениях), что такое увеличение содержания ПОУ в этих ареалах имеется на самом деле, но тот факт, что при моделировании (рис. 3, 8е) такие предикторы как CND, Convexity, TWI присутствуют не в самых нижних позициях, всё же указывает на то, что эти закономерности накопления ПОУ в понижениях были выявлены по данным 1984–1994 гг.

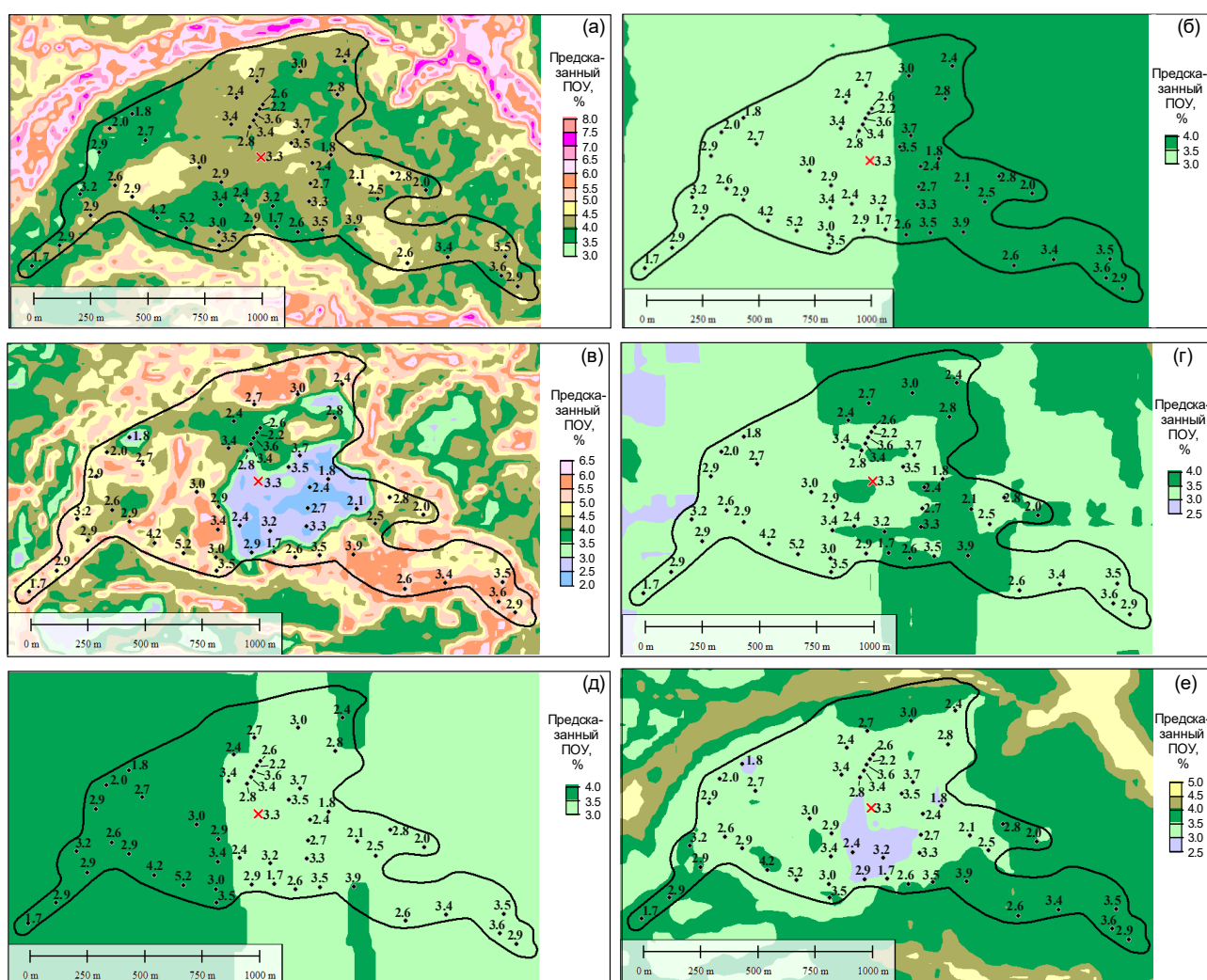
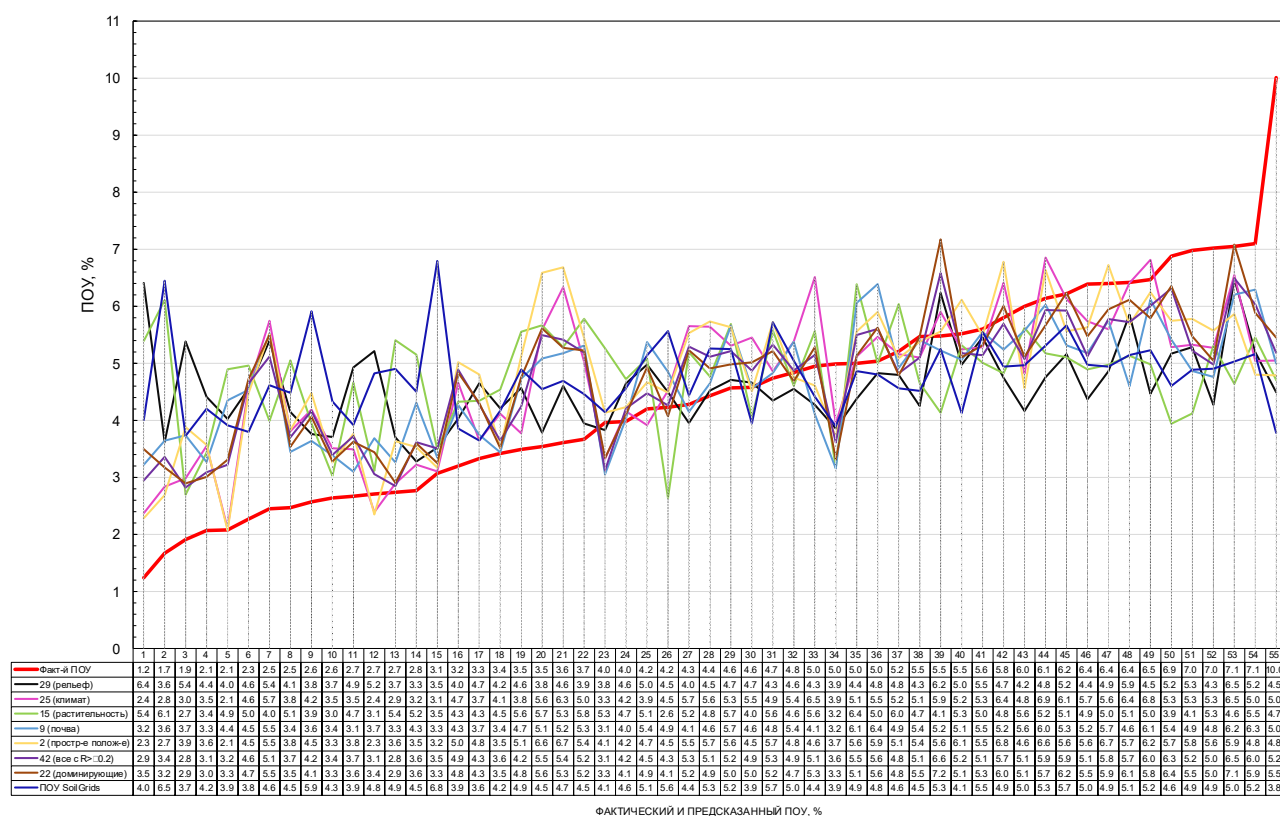


Рисунок 10. Сравнение предсказанного по отдельным и совмещённым факторам почвообразования (по данным 1984–1994 гг.) и фактического (данные за 2015 г.) содержания ПОУ в слое 0–30 см при увеличенном масштабе карты: а – 29 (рельеф, рис. 10а); б – 25 (климат, рис. 10 б); в – 15 (растительность, рис. 10 в); г – 9 (почва, рис. 10 г); д – 2 (пространственное положение, рис. 10 д); е – 22 (все доминирующие предикторы: по 5 первых из каждой группы (рис. 8 а, б, в, г + 2 (рис. 8 д)). Примечание: цифры над черными точками на картах представляют фактическое содержание ПОУ в слое 0–30 см в 2015 году ($n=50$), красный крест – фактическое содержание ПОУ по базе данных 1984–1994 гг.; черным контуром выделено поле (пахотное с/х угодье).

Анализ карт показал, что предикторы от разных групп факторов почвообразования вносят определенные закономерности в конечный результат картографирования содержания ПОУ и

совместное их использование в плане улучшения показателей эффективности моделирования ($RMSE_{внд}$, $MAPE_{внд}$, $MAE_{внд}$) позволяет получить лучший результат (табл. 5, рис. 9е, 10е), чем при использовании предикторов по отдельности (табл. 5, рис. 9 а–д, 10 а–д). Одни группы предикторов (климат, почва, пространственное положение) позволяют отобразить на итоговой карте глобальный тренд в изменении содержания ПОУ, а другие (рельеф, растительность) – дифференцирующие/детализирующие тренды, учитывающие особенности рельефа и растительности. Таким образом, использование растровых карт предикторов с низким исходным разрешением (1×1 км – WorldClim, 250×250 м – SoilGrids) и улучшенным до 30×30 м, в сочетании с картами предикторов среднего разрешения (рельеф, растительность, пространственное положение) позволяет достичь лучших результатов в картографировании содержания ПОУ.

На рисунке 11 показано, что наилучшим образом ход кривой фактических данных повторяют кривые, построенные по картам ПОУ, составленным с использованием следующих наборов предикторов: 42 (все с $R > \pm 0,2$), 22 (доминирующие) и 25 (климат). В целом, все модели занизили относительно высокие значения ПОУ и завысили относительно низкие значения ПОУ, т.е. алгоритм RF уравнивает/усредняет предсказания по содержанию ПОУ (рис. 11). Обучающий и валидационный наборы данных содержали выбросы, которые все модели не предсказали. В пунктах отбора проб валидационной выборки прогноз для лучшей модели 42 (все с $R > \pm 0,2$) был неточным в 2,4 и 1,9 раза для минимальных и максимальных значений по содержанию ПОУ, соответственно (точки 1 и 55). Предсказанные значения ПОУ имели меньший диапазон вариации (рис. 11).



ФАКТИЧЕСКИЙ И ПРЕДСКАЗАННЫЙ ПОУ, %

Рисунок 11. Сравнение предсказанных на основе разных наборов предикторов и фактических данных по содержанию ПОУ (1984–1994 гг.) в 0–30 см слое почв (валидационный набор данных, $n=55$). Примечание: номера точек валидационного набора данных указаны в первой (верхней) строке.

Ход кривой фактических данных хуже всего совпадает с кривой, построенной по карте ПОУ, составленной с использованием 15 предикторов, характеризующих только растительность (рис. 11). Использование этих предикторов при картографировании на региональном уровне, возможно, будет приводить к получению недостаточно качественных карт. Однако использование большего количества растительных показателей в наборе доминирующих предикторов (22) привело к появлению на карте ареалов с более низкими значениями ПОУ (от 2,5 до 3,0), которых нет на картах, составленных с использованием других предикторов (климат, рельеф, почва, пространственное положение); таких

ареалов нет и на карте, составленной с использованием 42-х предикторов. Предсказанные значения по содержанию ПОУ в этих ареалах лучше совпадают с фактическими данными 2015 года (рис. 10 в, е).

Таким образом, можно сделать вывод, что предикторы из разных групп факторов почвообразования взаимно дополняют друг друга в совмещенном наборе доминирующих предикторов и позволяют получить лучший результат при картографировании содержания ПОУ, чем их использование по отдельности. Необходимо отметить, что на карте ПОУ, составленной с использованием 42 предикторов, нет ареалов со значениями ПОУ от 2,5 до 3,0%, которые есть на карте, составленной с использованием 22 предикторов. По всей видимости, это говорит о том, что у модели 42 (все с $R \geq 0,2$) используется избыточное количество других предикторов, к которым, например, можно отнести климатические переменные, находящиеся по степени важности в нижних позициях списка (рис. 3 д). Это можно связать и с использованием большего количества растительных показателей в модели с 22 предикторами по сравнению с набором, состоящим из 42 предикторов. Следовательно, можно сделать вывод, что избыточное количество предикторов из одной группы факторов почвообразования может привести, например, к снижению детальности итоговой карты или наоборот. Кроме того, это свидетельствует, что доля/количество предикторов из каждой группы оказывает влияние на детальность итоговой карты ПОУ.

В нашем исследовании в качестве предикторов для картографирования содержания ПОУ были использованы глобальные почвенные карты SoilGrids (табл. 1), поэтому следующий этап данной работы состоял в визуальном сравнении карты ПОУ от SoilGrids с лучшими картами, построенными нами (рис. 12).

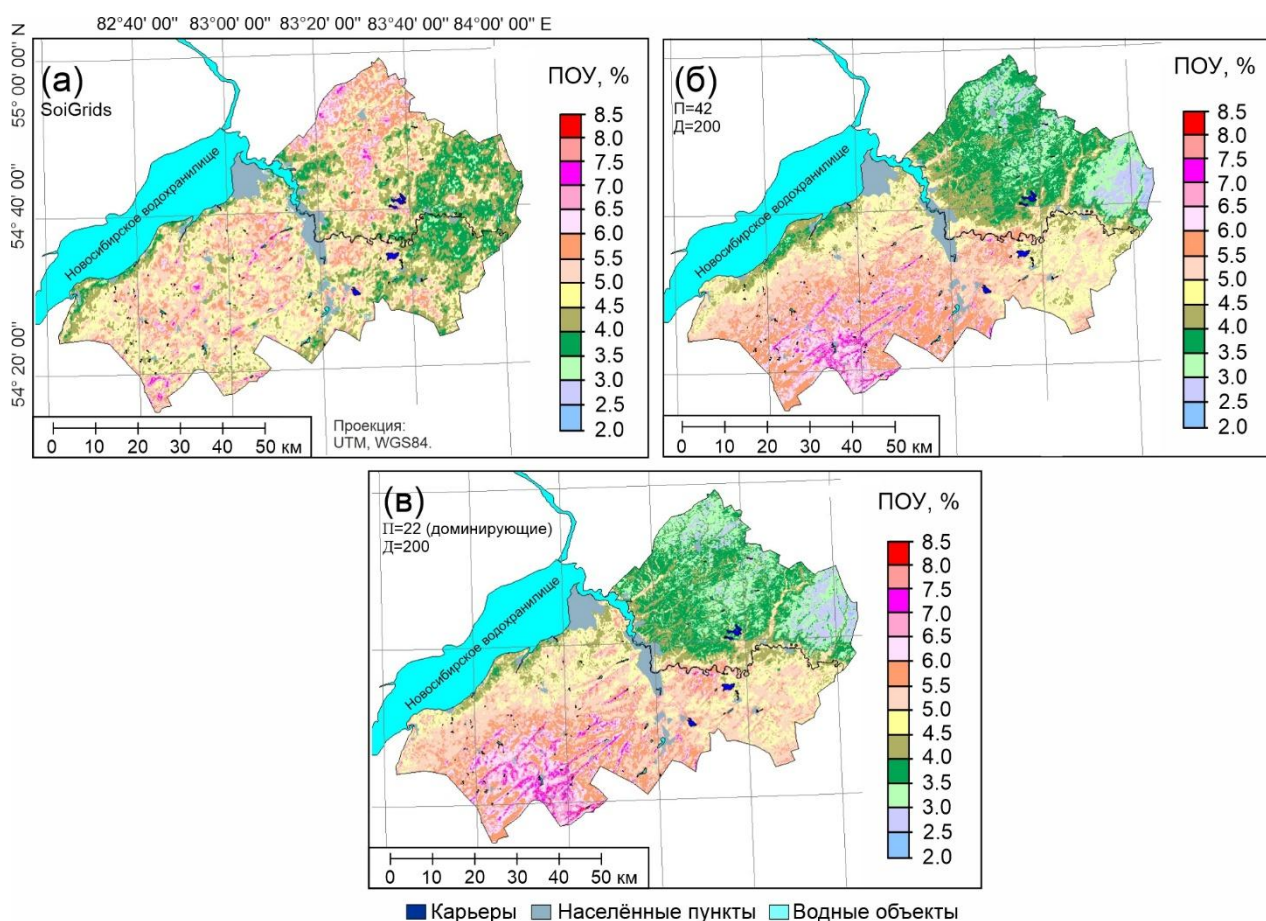


Рисунок 12. Карты содержания ПОУ: а – SoilGrids; б – 42 (все с $R \geq 0,2$); в – 22 (доминирующие).

Следует отметить, что глобальные карты имеют определенные минусы, например, в некоторых местах могут встречаться артефакты, локализацию которых необходимо выявлять, и, по всей видимости, принимать меры по инициированию исправлений. Другой минус этих карт состоит в том, что нет информации о местоположении почвенных профилей на исследуемой территории, поэтому судить о репрезентативности обучающей выборки, использованной при построении карт ПОУ SoilGrids, очень сложно. Подчеркнем, что подобное сравнение карт следует считать исключительно разведочным, так как усреднение растровых карт SoilGrids в нашей работе проведено не по стандарту.

По действующим правилам (FAO and ITPS, 2020) для получения значений содержания ПОУ для заданной мощности слоя (0–30 см) необходимо предварительно провести гармонизацию значений из разных интервалов глубин почвенного профиля, так как растровой карты содержания ПОУ для слоя 0–30 см в проекте SoilGrids нет (<https://soilgrids.org/>). Поэтому усреднение готовых растровых карт содержания ПОУ, построенных для глубин 0–5, 5–15, 15–30 см, и последующее сравнение с полученными в нашем исследовании картами позволяет сделать выводы исключительно о трендах в изменении содержания ПОУ; при этом вести речь о точности этой карты будет некорректно, так как первоначальная гармонизация данных профилей почв для необходимой мощности не проведена, в связи с отсутствием у автора данной работы исходных данных SoilGrids. Однако, несмотря на эти сложности, карта содержания ПОУ SoilGrids в настоящее время является чуть ли не единственной глобальной картой и её определенно необходимо использовать для разведочного и сравнительного анализа в исследованиях. Сравнительный анализ карт показал, что на карте SoilGrids имеется явно выраженный тренд увеличения содержания ПОУ в северо-западной части исследуемой нами территории (рис. 12а). Такого тренда нет на построенных в нашем исследовании лучших картах содержания ПОУ (рис. 12б, в). Репрезентативность выборки в проведенном исследовании, несомненно, выше, поэтому можно сделать заключение о наличии локальных завышений значений содержания ПОУ на картах SoilGrids, что, по всей видимости, обусловлено наличием недостаточного количества исходных данных, использованных в качестве обучающей выборки при моделировании с использованием алгоритма машинного обучения. Завышение значений ПОУ на карте SoilGrids по отношению к фактическим данным валидационной выборки заметно на рисунке 11 (первые 16 точек валидационного набора данных).

Таким образом, сравнение карт показало, что потенциал алгоритмов машинного обучения, используемых в картографии, может снижаться в зависимости от объема обучающего набора и наличия в нём вариативных почвенных данных, охватывающих, по возможности, всю территорию исследования.

ЗАКЛЮЧЕНИЕ

Основная задача нашего исследования заключалась в определении показателей эффективности моделирования содержания ПОУ с использованием 11 наборов предикторов и алгоритма Random Forest, реализованного на онлайн-платформе Google Earth Engine. Из 11 наборов предикторов для моделирования содержания ПОУ в слое 0–30 см был выбран лучший набор, состоящий из 42 предикторов. Использование этого набора предикторов показало лучшую эффективность моделирования ($R^2_{\text{онд}}=0,83$; $R^2_{\text{внд}}=0,55$; $\text{RMSE}_{\text{внд}}=1,25$). Обучающий и валидационный наборы данных содержали выбросы, которые все модели не предсказали. В пунктах отбора проб валидационной выборки прогноз для лучшей модели (42 (все с $R>\pm 0,2$) был неточным в 2,4 и 1,9 раза для минимальных и максимальных значений содержания ПОУ, соответственно. Предсказанные значения ПОУ имели меньший диапазон вариации. Сравнение средних значений в выборках показало, что увеличение количества «деревьев» с 200 до 1000 не привело к существенному улучшению показателей эффективности моделирования.

Содержание ПОУ в изученных почвах варьирует от 1,2 до 13,9% по фактическим данным. Согласно построенной карте, содержание ПОУ увеличивается в юго-западном направлении. Предсказанные моделью средние значения по отношению к среднему фактическому содержанию ПОУ были в 1,1–1,2 раза выше для четырех типов и подтипов почв (луговые аллювиальные, луговые карбонатные, светло-серые лесные оподзоленные, серые лесные оподзоленные), в то время как для трех типов почв они были в 1,1–1,2 раза ниже (луговые подзолистые, лугово-черноземные оподзоленные, темно-серые лесные оподзоленные). Для пяти типов и подтипов почв соответствие предсказанных и фактических средних значений удовлетворительное (луговые солончаковатые, лугово-черноземные обыкновенные, солонцы луговые, черноземы выщелоченные, черноземы оподзоленные).

Использование растровых карт предикторов с низким исходным разрешением (1×1 км – WorldClim, 250×250 м – SoilGrids) и улучшенным до 30×30 м, в сочетании с картами предикторов среднего разрешения (рельеф, растительность, пространственное положение), позволяет достичь лучших результатов в картографировании содержания ПОУ. Одни группы предикторов (климат, почва, пространственное положение) позволяют отобразить на итоговой карте глобальный тренд в изменении содержания ПОУ, а другие (рельеф, растительность) – дифференцирующие/детализирующие тренды, учитывающие особенности рельефа и растительности.

Согласно исследованию (Argouays et al., 2020), для создания цифровых почвенных карт в государственном и частном секторах требуются стандарты, чтобы предотвратить получение некачественной информации, которая может ввести в заблуждение и привести к обратным результатам. В частности, необходимо ввести минимальные статистические стандарты для оценки карт, составленных с использованием методов цифрового картографирования. По нашему мнению, землепользователям необходимо предоставлять графический материал, показывающий, во сколько раз модель превысила или недооценила значения конкретного почвенного показателя в валидационном/независимом наборе данных с близким расположением точек отбора образцов. Для усовершенствования методологии оценки эффективности моделирования необходимо, чтобы валидационные/независимые наборы данных были расположены как в районах с низким содержанием ПОУ, так и в районах с его высоким содержанием, например, в юго-восточной и северо-восточной части исследуемого региона (на примере исследуемого участка в нашей работе).

В большинстве исследований используются почвенные базы данных, созданные в разные годы (Чернова и др., 2021; FAO and ITPS, 2020; Poggio et al., 2021), в которых не рассматривается вопрос о проверке достоверности результатов моделирования по устаревшим и современным данным, которые часто используются совместно. Трудоёмкость проведения полевых исследований почв не позволяет в короткие сроки создать полноценные базы данных для проведения цифрового картографирования свойств почв, а также независимые наборы данных для проверки точности карт, в связи с чем, проблемы цифрового картографирования свойств почв сохраняют свою актуальность и в настоящее время.

ФИНАНСОВАЯ ПОДДЕРЖКА

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

ЛИТЕРАТУРА

- Аринушкина Е.В. Руководство по химическому анализу почв. Москва: Издательство МГУ, 1970. 488 с.
- Афанасьев В.Н., Цыпин А.П. Эконометрика в пакете STATISTICA: учебное пособие по выполнению лабораторных работ. Оренбург: ГОУ ОГУ, 2008. 204 с.
- Гопп Н.В. Использование почвенно-геоморфологической базы данных для изучения пространственной изменчивости содержания гумуса, физической глины и ила в почвах Кузнецко-Салаирской геоморфологической провинции // Почвоведение. 2021. № 7. С. 783–796. <https://doi.org/10.31857/S0032180X21070054>
- Гопп Н.В. Агроэкологический потенциал западной части Кузнецко-Салаирской геоморфологической провинции: методика цифрового картографирования, геопространственный анализ, корреляция с содержанием органического углерода в почвах // Почвы и окружающая среда. 2023. Том 6. № 3. e224. <https://doi.org/10.31251/pos.v6i3.224>
- Гопп Н.В., Мешалкина Ю.Л., Нарыкова А.Н., Плотникова А.С., Чернова О.В. Картографирование содержания и запасов органического углерода почв на региональном и локальном уровнях: анализ современных методических подходов // Вопросы лесной науки. 2023. Том 6. № 1. С. 14–73. <https://doi.org/10.31509/2658-607x-202361-120>
- Гопп Н.В., Нечаева Т.В. Цифровое картографирование содержания и запасов органического углерода в почвах склона (на примере ключевого участка в Предсалаирье) // Материалы Международной научно-практической конференции «Проблемы плодородия почв в современном земледелии», посвященной 70-летию освоения целинных и залежных земель (Красноярск, 24–28 июня 2024 г.). Красноярск: ФИЦ КНЦ СО РАН, 2024. С. 403–407.
- Евдокимова Т.И. Почвенная съемка. Москва: Изд-во МГУ, 1987. 270 с.
- Классификация и диагностика почв СССР / Составители: В.В. Егоров, В.М. Фридланд, Е.Н. Иванова, Н.Н. Розова, В.А. Носин, Т.А. Фриев. Москва: Колос, 1977. 224 с.
- Романова А.И., Маханцева В.А., Волкова Т.Ю., Лебедева Т.Н., Журавлева А.И., Митрохина Е.С., Курганова И.Н., Иващенко К.В. Пространственно-временной анализ почвенных показателей углеродного цикла в постагрогенных экосистемах южного Подмоскovie // Почвы и окружающая среда. 2024. Том 7. № 3. e261. <https://doi.org/10.31251/pos.v7i3.261>
- Сляднев А.П. Почвенно-климатический атлас Новосибирской области. Новосибирск: Наука, 1978.

- Чернова О.В., Голозубов О.М., Алябина И.О., Щепашенко Д.Г. Комплексный подход к картографической оценке запасов органического углерода в почвах России // Почвоведение. 2021. № 3. С. 273–286. <https://doi.org/10.31857/S0032180X21030047>
- Шарков И.Н., Антипина П.В. Некоторые аспекты углерод-секвестрирующей способности пахотных почв // Почвы и окружающая среда. 2022. Том 5. № 2. с175. <https://doi.org/10.31251/pos.v5i2.175>
- Arrouays D., McBratney A., Bouma J., Libohova Z., Richer-de-Forges A.C., Morgan C.L.S., Roudier P., Poggio L., Mulder V.L. Impressions of digital soil maps: The good, the not so good, and making them ever better // Geoderma Regional. 2020. Vol. 20. e00255. <https://doi.org/10.1016/j.geodrs.2020.e00255>
- Breiman L. Random forests // Machine learning. 2001. Vol. 45. No. 1. P. 5–32. <https://doi.org/10.1023/A:1010933404324>
- Conrad O., Bechtel B., Bock M., Dietrich H., Fischer E., Gerlitz L., Wehberg J., Wichmann V., Böhner J. System for automated geoscientific analyses (SAGA) v. 2.1.4 // Geoscientific model development. 2015. Vol. 8. No. 7. P. 1991–2007. <https://doi.org/10.5194/gmd-8-1991-2015>
- Cutler D.R., Edwards T.C.Jr., Beard K.H., Cutler A., Hess K.T., Gibson J., Lawler J.J. Random forests for classification in Ecology // Ecology. 2007. Vol. 88. No. 1. P. 2783–2792. <https://doi.org/10.1890/07-0539.1>
- Defourny P., Lamarche C., Bontemps S., De Maet T., Van Bogaert E., Moreau I., Brockmann C., Boettcher M., Kirches G., Wevers J., Santoro M., Ramoino F., Arino O. Land Cover Climate Change Initiative – Product User Guide v2. 2017. Issue 2.0. http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf
- Defourny P., Lamarche C., Brockmann C., Boettcher M., Bontemps S., De Maet T., Duveiller G. L., Harper K., Hartley A., Kirches G., Moreau I., Peylin P., Ottlé C., Radoux J., Van Bogaert E., Ramoino F., Albergel C., Arino O. Observed annual global land-use change from 1992 to 2020 three times more dynamic than reported by inventory-based statistics, in preparation, 2023. URL: <https://maps.elie.ucl.ac.be/CCI/viewer/index.php> (дата обращения 11.11.2024 г.).
- Duarte E., Zagal E., Barrera J.A., Dube F., Casco F., Hernández A.J. Digital mapping of soil organic carbon stocks in the forest lands of Dominican Republic // European journal of remote sensing. 2022. Vol. 55. No. 1. P. 213–231. <https://doi.org/10.1080/22797254.2022.2045226>
- Erunova M.G., Yakubailik O.E. Geomorphometric analysis of agricultural areas based on the FABDEM digital elevation model // InterCarto. InterGIS. 2024. Vol. 30. No. 2. P. 252–262. <https://doi.org/10.35595/2414-9179-2024-2-30-252-262>
- FAO. Standard operating procedure for soil organic carbon: Tyurin spectrophotometric method. Rom: FAO, 2021. URL: <https://www.fao.org/3/cb4757en/cb4757en.pdf> (дата обращения 11.11.2024 г.).
- FAO and ITPS. Global Soil Organic Carbon Map V1.5: Technical report. Rome, FAO, 2020. P. 169.
- Fick S.E., Hijmans R.J. WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas // International Journal of Climatology. 2017. Vol. 37. No. 12. P. 4302–4315. <https://doi.org/10.1002/joc.5086>
- Gandhi U. JavaScript and the Earth Engine API. In book: Cloud-Based Remote Sensing with Google Earth Engine / J.A. Cardille, M.A. Crowley, D. Saah, N.E. Clinton (eds). Cham: Springer, 2023. P. 3–18. https://doi.org/10.1007/978-3-031-26588-4_1
- Gorelick N., Hancher M., Dixon M., Ilyushchenko S., Thau D., Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone // Remote sensing of Environment. 2017. Vol. 202. P. 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- Hawker L., Uhe P., Paulo L., Sosa J., Savage J., Sampson C., Neal J. A 30 m global map of elevation with forests and buildings removed // Environmental Research Letters. 2022. Vol. 17. No. 2. P. 024016. <https://doi.org/10.1088/1748-9326/ac4d4f>
- Hengl T.A. Practical guide to geostatistical mapping of environmental variables. EC JRC, Ispra (Italy), 2007. 165 p.
- Ignatenko V., Surkov A., Koltcov S. Random forests with parametric entropy-based information gains for classification and regression problems // PeerJ Computer Science. 2024. Vol. 10. e1775. <https://doi.org/10.7717/peerj-cs.1775>
- IPCC. Guidelines for National Greenhouse Gas Inventories, Vol. 4: Agriculture, forestry and other types of land use. Japan: IGES, 2006.
- IUSS. Working Group WRB. World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps. Update 2015. World Soil Resources Reports No. 106. FAO, Rome, 2015. <https://www.fao.org/3/i3794en/i3794en.pdf>
- Luo C., Zhang X., Meng X., Zhu H., Ni C., Chen M., Liu H. Regional mapping of soil organic matter content using multitemporal synthetic Landsat 8 images in Google Earth Engine // Catena. 2022. Vol. 209. P. 105842. <https://doi.org/10.1016/j.catena.2021.105842>

Moquedace C.M., Baldi C.G.O., Siqueira R.G., Cardoso I.M., Souza E.F.M., Fontes R.L.F., Francelino M.R., Gomes L.C., Fernandes-Filho E.I. High-resolution mapping of soil carbon stocks in the western Amazon // *Geoderma Regional*. 2024. Vol. 36. e00773. <https://doi.org/10.1016/j.geodrs.2024.e00773>

O'Brien L. Mpspline2: Mass-Preserving Spline Functions for Soil Data. 2022. Available online: <https://CRAN.R-project.org/package=mpspline2> (дата обращения 14.12.2024).

Poggio L., de Sousa L., Batjes N., Heuvelink G., Kempen B., Ribeiro E., Rossiter D. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty // *Soil*. 2021. Vol. 7. No. 1. P. 217–240. <https://doi.org/10.5194/soil-7-217-2021>

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2022. <https://www.R-project.org/>

Rodriguez-Galiano V.F., Ghimire B., Rogan J., Chica-Olmo M., Rigol-Sanchez J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2012. 67. P. 93–104. DOI: <https://doi.org/10.1016/j.isprsjprs.2011.11.002>

Singh J., Knapp H.V., Demissie M. Hydrologic modelling of the Iroquois River watershed using HSPF and SWAT. Illinois State Water Survey Contract Report 2004-08. Illinois State Water Survey, Champaign. 2004. <http://www.isws.illinois.edu/pubdoc/CR/ISWSCR2004-08.pdf>

Suleymanov A., Tuktarova I., Belan L., Suleymanov R., Gabbasova I., Araslanova L. Spatial prediction of soil properties using random forest, k-nearest neighbors and cubist approaches in the foothills of the Ural Mountains, Russia // *Modelling Earth System and Environment*. 2023. Vol. 9. No. 3. P. 3461–3471. <https://doi.org/10.1007/s40808-023-01723-4>

Zhang F., Wen N. Carbon price forecasting: a novel deep learning approach // *Environmental Science and Pollution Research*. 2022. Vol. 29. No. 36. P. 54782–54795. <https://doi.org/10.1007/s11356-022-19713-x>

Zhou T., Wenhao L.W., Geng Y., Xiao S., Chen J., Xu X., Pan J., Si B., Lausch A. National-scale spatial prediction of soil organic carbon and total nitrogen using long-term optical and microwave satellite observations in Google Earth Engine // *Computers and Electronics in Agriculture*. 2023. Vol. 210. P. 107928. <https://doi.org/10.1016/j.compag.2023.107928>

Поступила в редакцию 12.09.2025

Принята 14.10.2025

Опубликована 17.10.2025

Сведения об авторе:

Гопп Наталья Владимировна – кандидат биологических наук, ведущий научный сотрудник лаборатории географии и генезиса почв ФГБУН Институт почвоведения и агрохимии СО РАН (г. Новосибирск, Россия); natalia.gopp@gmail.com; <https://orcid.org/0000-0001-7549-2984>

Автор прочитал и одобрил окончательный вариант рукописи.



Статья доступна по лицензии [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Digital mapping of soil organic carbon content in the Kuznetsk-Salair geomorphological province using the Google Earth Engine platform

© 2025 N. V. Gopp



Institute of Soil Science and Agrochemistry, Siberian Branch of the Russian Academy of Sciences, Lavrentieva 8/2, Novosibirsk, Russia. E-mail: natalia.gopp@gmail.com

Abstract. The aim of this study was to create digital maps of soil organic carbon (SOC) content in the 0–30 cm layer using 11 different predictor datasets and the Random Forest algorithm implemented on the Google Earth Engine platform.

Location and time of the study. The study was carried out in the Kuznetsk-Salair geomorphological province, within the Iskitimsky district in the Novosibirsk Region.

Methods. Digital mapping was performed using the Random Forest (RF) algorithm with settings of 200 and 1000 "trees". RF algorithm is implemented on the Google Earth Engine online platform in the form of a script (code). SOC content was mapped using raster predictor maps that characterized the following soil-forming factors: climate, relief, vegetation, spatial position, and soil properties. The number of predictors used in the modeling varied depending on the number of available indicators and the correlation coefficient (R) between the SOC

content and the predictor values. The following numbers of predictors were used: 80 (all); 2) 50 (all with $R > \pm 0,1$); 3) 42 (all with $R > \pm 0,2$); 4) 26 (all with $R > \pm 0,3$); 5) 15 (all with $R > \pm 0,4$). Additionally, the SOC content was mapped using five sets of predictors characterizing individual soil formation factors (relief, climate, vegetation, soil, spatial position), and one dataset combining five dominant predictors from each soil formation factor. Thus, a total of 11 sets of predictors was used, the sets presented both in various combinations of indicators of soil formation factors and individually. Training (TD, $n=206$) and validation (VD, $n=55$) data sets were used in the study. The database was created using archival materials from the "Zapsibgiprozem" Land Management Institute from 1984 to 1994. It contains data on the SOC content in the 0–30 cm layer, with a sample size of 261.

Results. The optimal model for mapping SOC content in the 0–30 cm layer, based on 42 predictors and 200 trees, demonstrated the highest accuracy. The simulation performance indicators are as follows: $R^2_{TD}=0,83$, $R^2_{VD}=0,55$; $RMSE_{VD}=1,25\%$. Model performance indicators showed that the model performed well. According to the actual data, the SOC content ranged from 1,2 to 13,9%. The generated map showed that the highest SOC levels were found in the southwestern part of the study area.

Conclusions. The use of raster maps of predictors with low original resolution (1×1 km – WorldClim, 250×250 m – SoilGrids) improved to 30×30 m, combined with medium-resolution predictor maps (topography, vegetation, spatial location), allows for better results in digital mapping of SOC. Some groups of predictors (climate, soil, spatial location) provide a global trend in the variation of SOC content in the final map, while others (topography, vegetation) display differentiating/detailing trends that take into account the characteristics of the terrain and vegetation. Thus, predictors characterizing different soil formation factors complement each other in the combined predictor dataset and enable development of more detailed maps during SOC mapping.

Keywords: digital mapping; organic carbon content; soils; WorldClim; West Siberia; WorldClim; FABDEM; SoilGrids; Landsat 5; SAGA GIS.

How to cite: Gopp N.V. Digital mapping of soil organic carbon content in the Kuznetsk-Salair geomorphological province using the Google Earth Engine platform. The Journal of Soils and Environment. 2025. 8(3). e335. DOI: [10.31251/pos.v8i3.335](https://doi.org/10.31251/pos.v8i3.335) (in Russian with English abstract).

FUNDING

The work was carried out as a part of the most important innovative project of national importance «Development of a System of Ground-based and Remote Monitoring of Carbon Pools and Greenhouse Gas Flows on the Territory of the Russian Federation, ensuring the Creation of a System for Recording Data on the Fluxes of Climate-active Substances and Carbon Budget in Forests and other Terrestrial Ecological Systems» (reg. No. 123030300031-6).

REFERENCES

- Arinushkina E.V. Guidelines for the Chemical Analysis of Soils. Moscow: Moscow State University Publishing House, 1970. 488 p. (in Russian).
- Afanasyev V.N., Tsylin A.P. Econometrics in the STATISTICA package: guidelines for the laboratory work. Orenburg: OSU, 2008. 204 p. (in Russian).
- Gopp N.V. The Use of the Soil-Geomorphological Database for Studying the Spatial Variability of the Humus Content, Physical Clay, and Clay in the Soils of the Kuznetsk–Salair Geomorphological Province. Eurasian Soil Science. 2021. Vol. 54. No. 7. P. 986–998. <https://doi.org/10.1134/S106422932107005X>
- Gopp N.V. Agroecological potential of the western part of the Kuznetsk-Salair geomorphological province: digital mapping methodology, geospatial analysis, correlation with soil organic carbon content. The Journal of Soils and Environment. 2023. Vol. 6. No. 3. e224. (in Russian). <https://doi.org/10.31251/pos.v6i3.224>
- Gopp N.V., Meshalkina J.L., Narykova A.N., Plotnikova A.S., Chernova O.V. Mapping of soil organic carbon content and stocks at the regional and local levels: the analysis of modern methodological approaches. Forest Science Issues. 2023. Vol. 6. No. 1. P. 14–73. (in Russian). <https://doi.org/10.31509/2658-607x-202361-120>
- Gopp N.V., Nechaeva T.V. Digital mapping of organic carbon content and stocks in soils on the slope (using the example of a key site in the Cis-Salair drained plain). In book: Problems of soil fertility in modern agriculture. Proceedings of the International Scientific and Practical Conference Materials dedicated to the 70th anniversary of the development of virgin and fallow lands (Krasnoyarsk, 24–28 June, 2024). Krasnoyarsk: KSC SB RAS, 2024. P. 403–407. (in Russian).
- Evdokimova T.I. Soil survey. Moscow: Publishing house of Moscow State University, 1987. 270 p. (in Russian).
- Classification and diagnostics of soils of the USSR / Compiled by: V.V. Egorov, V.M. Friedland, E.N. Ivanova, N.N. Rozova, V.A. Nosin, T.A. Frieve. Moscow: Kolos Publ., 1977. 224 p. (in Russian).
- Romanova A.I., Makhantceva V.A., Volkova T.Yu., Lebedeva T.N., Zhuravleva A.I., Mitrokhina E.S., Kurganova I.N., Ivashchenko K.V. Spatio-temporal analysis of soil indicators of the carbon cycle in some postagrogenic ecosystems of

- the southern Moscow region. The Journal of Soils and Environment. 2024. Vol. 7. No. 3. e261. (in Russian). <https://doi.org/10.31251/pos.v7i3.261>
- Slyadnev A.P. Soil and climatic atlas of the Novosibirsk region. Novosibirsk: Nauka Publ., Siberian Branch, 1978. 122 p. (in Russian).
- Chernova O.V., Golozubov O.M., Alyabina I.O., Shepashchenko D.G. Integrated Approach to Spatial Assessment of Soil Organic Carbon in Russian Federation. Eurasian Soil Science. 2021. Vol. 54. No. 3. P. 325–336. <https://doi.org/10.1134/S1064229321030042>
- Sharkov I.N., Antipina P.V. Some aspects of carbon sequestration capacity of arable soils. The Journal of Soils and Environment. 2022. Vol. 5. No. 2. e175. (in Russian). <https://doi.org/10.31251/pos.v5i2.175>
- Arrouays D., McBratney A., Bouma J., Libohova Z., Richer-de-Forges A.C., Morgan C.L.S., Roudier P., Poggio L., Mulder V.L. Impressions of digital soil maps: The good, the not so good, and making them ever better. Geoderma Regional. 2020. Vol. 20. e00255. <https://doi.org/10.1016/j.geodrs.2020.e00255>
- Breiman L. Random forests. Machine learning. 2001. Vol. 45. No. 1. P. 5–32. <https://doi.org/10.1023/A:1010933404324>
- Conrad O., Bechtel B., Bock M., Dietrich H., Fischer E., Gerlitz L., Wehberg J., Wichmann V., Böhner J. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geoscientific model development. 2015. Vol. 8. No. 7. P. 1991–2007. <https://doi.org/10.5194/gmd-8-1991-2015>
- Cutler D.R., Edwards T.C.Jr., Beard K.H., Cutler A., Hess K.T., Gibson J., Lawler J.J. Random forests for classification in Ecology. Ecology. 2007. Vol. 88. No. 1. P. 2783–2792. <https://doi.org/10.1890/07-0539.1>
- Defourny P., Lamarche C., Bontemps S., De Maet T., Van Bogaert E., Moreau I., Brockmann C., Boettcher M., Kirches G., Wevers J., Santoro M., Ramoino F., Arino O. Land Cover Climate Change Initiative – Product User Guide v2. 2017. Issue 2.0. http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf
- Defourny P., Lamarche C., Brockmann C., Boettcher M., Bontemps S., De Maet T., Duveiller G. L., Harper K., Hartley A., Kirches G., Moreau I., Peylin P., Ottlé C., Radoux J., Van Bogaert E., Ramoino F., Albergel C., Arino O. Observed annual global land-use change from 1992 to 2020 three times more dynamic than reported by inventory-based statistics, in preparation, 2023. URL: <https://maps.elie.ucl.ac.be/CCI/viewer/index.php> (accessed on 11.11.2024).
- Duarte E., Zagal E., Barrera J.A., Dube F., Casco F., Hernández A.J. Digital mapping of soil organic carbon stocks in the forest lands of Dominican Republic. European journal of remote sensing. 2022. Vol. 55. No. 1. P. 213–231. <https://doi.org/10.1080/22797254.2022.2045226>
- Erunova M.G., Yakubailik O.E. Geomorphometric analysis of agricultural areas based on the FABDEM digital elevation model. InterCarto. InterGIS. 2024. Vol. 30. No. 2. P. 252–262. <https://doi.org/10.35595/2414-9179-2024-2-30-252-262>
- FAO. Standard operating procedure for soil organic carbon: Tyurin spectrophotometric method. Rom: FAO, 2021. URL: <https://www.fao.org/3/cb4757en/cb4757en.pdf> (accessed on 11.11.2024).
- FAO and ITPS. Global Soil Organic Carbon Map V1.5: Technical report. Rome, FAO, 2020. P. 169.
- Fick S.E., Hijmans R.J. WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas. International Journal of Climatology. 2017. Vol. 37. No. 12. P. 4302–4315. <https://doi.org/10.1002/joc.5086>
- Gandhi U. JavaScript and the Earth Engine API. In book: Cloud-Based Remote Sensing with Google Earth Engine / J.A. Cardille, M.A. Crowley, D. Saah, N.E. Clinton (eds). Cham: Springer, 2023. P. 3–18. https://doi.org/10.1007/978-3-031-26588-4_1
- Gorelick N., Hancher M., Dixon M., Ilyushchenko S., Thau D., Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment. 2017. Vol. 202. P. 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- Hawker L., Uhe P., Paulo L., Sosa J., Savage J., Sampson C., Neal J. A 30 m global map of elevation with forests and buildings removed. Environmental Research Letters. 2022. Vol. 17. No. 2. P. 024016. <https://doi.org/10.1088/1748-9326/ac4d4f>
- Hengl T.A. Practical guide to geostatistical mapping of environmental variables. EC JRC, Ispra (Italy), 2007. 165 p.
- Ignatenko V., Surkov A., Koltcov S. Random forests with parametric entropy-based information gains for classification and regression problems. PeerJ Computer Science. 2024. Vol. 10. e1775. <https://doi.org/10.7717/peerj-cs.1775>
- IPCC. Guidelines for National Greenhouse Gas Inventories, Vol. 4: Agriculture, forestry and other types of land use. Japan: IGES, 2006.
- IUSS. Working Group WRB. World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps. Update 2015. World Soil Resources Reports No. 106. FAO, Rome, 2015. <https://www.fao.org/3/i3794en/i3794en.pdf>

- Luo C., Zhang X., Meng X., Zhu H., Ni C., Chen M., Liu H. Regional mapping of soil organic matter content using multitemporal synthetic Landsat 8 images in Google Earth Engine. *Catena*. 2022. Vol. 209. P. 105842. <https://doi.org/10.1016/j.catena.2021.105842>
- Moquedace C.M., Baldi C.G.O., Siqueira R.G., Cardoso I.M., Souza E.F.M., Fontes R.L.F., Francelino M.R., Gomes L.C., Fernandes-Filho E.I. High-resolution mapping of soil carbon stocks in the western Amazon. *Geoderma Regional*. 2024. Vol. 36. e00773. <https://doi.org/10.1016/j.geodrs.2024.e00773>
- O'Brien L. Mpspline2: Mass-Preserving Spline Functions for Soil Data. 2022. Available online: <https://CRAN.R-project.org/package=mpspline2> (accessed on 14.12.2024).
- Poggio L., de Sousa L., Batjes N., Heuvelink G., Kempen B., Ribeiro E., Rossiter D. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil*. 2021. Vol. 7. No. 1. P. 217–240. <https://doi.org/10.5194/soil-7-217-2021>
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2022. <https://www.R-project.org/>
- Rodriguez-Galiano V.F., Ghimire B., Rogan J., Chica-Olmo M., Rigol-Sanchez J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2012. 67. P. 93–104. DOI: <https://doi.org/10.1016/j.isprsjprs.2011.11.002>
- Singh J., Knapp H.V., Demissie M. Hydrologic modelling of the Iroquois River watershed using HSPF and SWAT. Illinois State Water Survey Contract Report 2004-08. Illinois State Water Survey, Champaign. 2004. <http://www.isws.illinois.edu/pubdoc/CR/ISWSCR2004-08.pdf>
- Suleymanov A., Tuktarova I., Belan L., Suleymanov R., Gabbasova I., Araslanova L. Spatial prediction of soil properties using random forest, k-nearest neighbors and cubist approaches in the foothills of the Ural Mountains, Russia. *Modelling Earth System and Environment*. 2023. Vol. 9. No. 3. P. 3461–3471. <https://doi.org/10.1007/s40808-023-01723-4>
- Zhang F., Wen N. Carbon price forecasting: a novel deep learning approach. *Environmental Science and Pollution Research*. 2022. Vol. 29. No. 36. P. 54782–54795. <https://doi.org/10.1007/s11356-022-19713-x>
- Zhou T., Wenhao L.W., Geng Y., Xiao S., Chen J., Xu X., Pan J., Si B., Lausch A. National-scale spatial prediction of soil organic carbon and total nitrogen using long-term optical and microwave satellite observations in Google Earth Engine Computers and Electronics in Agriculture. 2023. Vol. 210. P. 107928. <https://doi.org/10.1016/j.compag.2023.107928>

Received 12 September 2025

Accepted 14 October 2025

Published 17 October 2025

About the author:

Natalya V. Gopp – Candidate of Biological Sciences, Leading Researcher in the Laboratory of Geography and Soil Genesis in the Institute of Soil Science and Agrochemistry, the Siberian Branch of the Russian Academy of Sciences (Novosibirsk, Russia); natalia.gopp@gmail.com; <https://orcid.org/0000-0001-7549-2984>

The author read and approved the final manuscript



The article is available under [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/)