ПОЧВА, 7, 217–240, 2021 https://doi.org/
10.5194/soil-7-217-2021

ПОЧВЫ Open Access EGU

SoilGrids 2.0: создание информации о почве для всего мира

с количественной пространственной неопределенностью

Лаура Поджо, Луис М. де Соуза, Нильс Х. Батжес, Жерар Б.М. Хевелинк, Бас Кемпен, Элой Рибейро, и Дэвид Росситер

ISRIC – Мировая информация о почвах, Вагенинген, Нидерланды

Переписка:Лаура Поджо (laura.poggio@wur.nl)

Получено: 14 октября 2020 г. – Начало обсуждения: 9 ноября 2020 г. Пересмотрено: 9 апреля 2021 г. – Принято: 18 апреля 2021 г. – Опубликовано: 14 июня 2021 г.

Абстрактный. SoilGrids создает карты свойств почвы для всего земного шара со средним пространственным разрешением (размер ячейки 250 м) с использованием современных методов машинного обучения для создания необходимых моделей. В качестве входных данных используются наблюдения за почвой из примерно 240 000 мест по всему миру и более 400 глобальных экологических ковариаций, описывающих растительность, морфологию местности, климат, геологию и гидрологию. Целью этой работы было создание глобальных карт свойств почвы с перекрестной проверкой, выбором гиперпараметров и количественным определением явной пространственной неопределенности, как это реализовано в продукте SoilGrids версии 2.0, включающем самые современные методы и адаптирующие их для глобального цифрового картографирования почвы с использованием устаревших данных. В статье представлена оценка глобальных прогнозов содержания органического углерода в почве, общего азота, крупные обломки, pH (вода), емкость катионного обмена, насыпная плотность и текстурные фракции на шести стандартных глубинах (до 200 см). Количественная оценка показала показатели, соответствующие предыдущим глобальным, континентальным исследованиям и исследованиям в крупных регионах. Качественная оценка показала, что крупномасштабные узоры воспроизводятся хорошо. Пространственная неопределенность в глобальном масштабе высветила необходимость большего количества наблюдений за почвой, особенно в высокоширотных регионах.

1. Введение

Здоровые почвы обеспечивают важные экосистемные услуги на местном, ландшафтном и глобальном уровнях и важны для функционирования наземных экосистем (Banwart et al., 2014; FAO and ITPS, 2015; UNEP, 2012). Информация о мировых почвенных ресурсах, основанная на «наилучших доступных» (общих) данных профиля почвы в масштабе, соответствующем потребностям пользователей, необходима для решения ряда насущных глобальных проблем. К ним относятся предотвращение и сокращение эрозии почвы за счет восстановления и освоения земель (Borrelli et al., 2017; WOCAT, 2007), смягчение последствий изменения климата и адаптация к ним (Batjes, 2019; Harden et al., 2017; Sanderman et al., 2017; Yigini and Panagos, 2016; Smith et al., 2019) и обеспечение водной безопасности (Rockstroem et al., 2012), производства продуктов питания и продовольственной безопасности (FAO et al., 2018; Soussana et al., 2017; Springmann et al. , 2018),

Наилучшие имеющиеся данные о почве необходимы для поддержки инициативы «Нейтральность деградации земель» (LDN) (Cowie et al., 2018), достижения нескольких целей в области устойчивого развития и обеспечения исходных данных, например, для моделирования земной системы МГЭИК (Dai et al. al., 2019; Luo et al., 2016; Todd-Brown et al., 2013) и моделирование сельскохозяйственных культур (Han et al., 2019; van Bussel et al., 2015; van Ittersum et al., 2013), среди многие другие приложения. Такая информация, в свою очередь, может помочь в разработке международных конвенций, таких как Рамочная конвенция Организации Объединенных Наций об изменении климата (РКИКООН), Конвенция Организации Объединенных Наций по борьбе с опустыниванием (КБО ООН) и Конвенция Организации Объединенных Наций о биологическом разнообразии (КБР ООН).

До последнего десятилетия в большинстве оценок глобального масштаба, требующих данных о почве, использовалась Цифровая карта почв мира (DSMW) ФАО (1995 г.), обновленная версия оригинала, напечатанная в масштабе 1:5.×10⁶масштабная карта почв мира (SMW) (ФАО-Юнеско, 1971–1981). Почвенно-географические данные из DSMW послужили основой для создания ряда де-

объединенные базы данных о свойствах почвы, основанные на более широком наборе данных о профилях почв, хранящихся в базе данных WISE (Batjes, 2012), и более сложных процедурах (таксотрансфер) для получения различных свойств почвы (Batjes et al., 2007). Впоследствии в результате совместных усилий, координируемых Продовольственной и сельскохозяйственной организацией Объединенных Наций (ФАО), наилучшая доступная (более новая) информация о почвах, собранная для центральной и южной части Африки, Китая, Европы, северной Евразии и Латинской Америки, была объединена в новый продукт, известный как Гармонизированная мировая база данных о почвах (HWSD) (FAO et al., 2012).

До недавнего времени HWSD была единственной базой данных приложений к цифровым картам, доступной для глобального анализа. Однако он имеет ряд ограничений (GSP and FAO, 2016; Hengl et al., 2014; Ivushkin et al., 2019; Omuto et al., 2012). Некоторые из них связаны с частично устаревшими почвенно-географическими данными, а также с использованием двухслойной модели (0–30 и 30–100 см) для получения свойств почвы. Другие касаются самих производных атрибутивных данных, в частности их неопределяемой количественно неопределенности и использования трех различных версий легенды ФАО (т.е. ФАО74, ФАО85 и ФАО90). Эти вопросы в той или иной степени были решены в различных новых глобальных наборах почвенных данных (Batjes, 2016; Shangguan et al., 2014; Stoorvogel et al., 2017), которые по-прежнему в значительной степени основаны на традиционном подходе к картографированию почв (Dai et al., 2019).).

В последнее десятилетие цифровое картографирование почв (DSM) стало широко использоваться для получения карт информации о почвах (Minasny and McBratney, 2016). DSM состоит в первую очередь в построении количественной численной модели на основе наблюдений за почвой и информации об окружающей среде, выступающей в качестве заменителей факторов почвообразования (McBratney et al., 2003; Minasny and McBratney, 2016). DSM также может интегрировать прямую информацию в качестве заменителей свойств почвы, например, измерения проксимального зондирования. Количество исследований с использованием DSM для создания карт свойств почвы постоянно растет. Рассматриваются многочисленные подходы к моделированию, от линейных моделей до геостатистики, машинного обучения и искусственного интеллекта (например, глубокого обучения). Кескин и Грюнвальд (2018) представили недавний обзор методов и приложений в области DSM. Витхарана и др., 2019 г.; Поджио и Джимона, 2017b; Kempen et al., 2019), региональный (например, Dorji et al., 2014; Moulatlet et al., 2017), континентальный (например, Grunwald et al., 2011; Guevara et al., 2018; Hengl et al., 2017a) и глобальном уровнях (например, Hengl et al., 2014, 2017b; GSP и ITPS, 2018; Stockmann et al., 2015). Витхарана и др., 2019 г.; Поджио и Джимона, 2017b; Kempen et al., 2019), региональный (например, Dorji et al., 2014; Moulatlet et al., 2017), континентальный (например, Grunwald et al., 2011; Guevara et al., 2018; Hengl et al., 2017a) и глобальном уровнях (например, Hengl et al., 2014, 2017b; GSP и ITPS, 2018; Stockmann et al., 2015).

Цель этой статьи состоит в том, чтобы представить разработку новых карт свойств почв для мира с разрешением сетки 250 м с процессом, включающим самые современные методы и адаптирующим их к задачам глобального цифрового картографирования почв с использованием устаревших данных. Он основан на предыдущих глобальных картах свойств почвы (SoilGrids250m) (Hengl et al., 2017b), интегрируя современные методы машинного обучения, дардизированные данные почвенных профилей для всего мира (Batjes et al., 2020) и экологические ковариаты (Nussbaum et al., 2018; Poggio et al., 2013; Reuter and Hengl, 2012). В частности, в этом документе рассматриваются следующие элементы в глобальном масштабе:

1. включение данных о профилях почв, полученных из Всемирной службы информации о почвах (WoSIS) IS-RIC, с расширенным количеством и пространственным распределением наблюдений (Batjes et al., 2020);

2. воспроизводимая процедура выбора ковариат, основанная на рекурсивном исключении признаков (Guyon et al., 2002);

3. улучшенная процедура перекрестной проверки, основанная на пространственной стратификации; а также

4. Количественная оценка неопределенности прогноза с использованием квантильных регрессионных лесов (Meinshausen, 2006).

2. Материалы и методы

В этом исследовании используются леса квантильной регрессии (Meinshausen, 2006), метод с ограниченным числом параметров для настройки, который оказался эффективным компромиссом между точностью и осуществимостью для больших наборов данных. Были смоделированы выбранные первичные свойства почвы, определенные и описанные в спецификациях GlobalSoilMap (Arrouays et al., 2014). В следующих разделах подробно описывается каждый шаг рабочего процесса (рис. 1). К ним относятся следующие:

1. подготовка исходных данных о почве

2. выбор ковариат

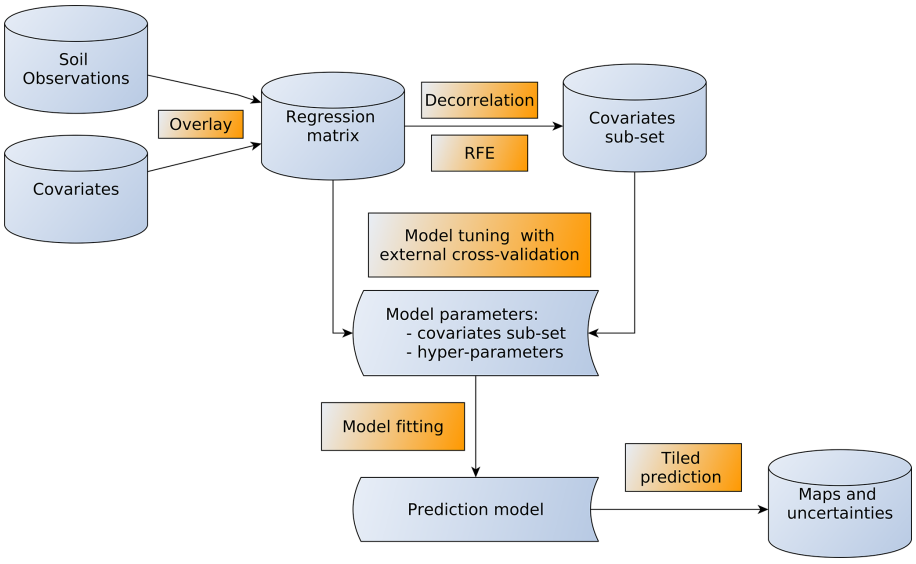
3. Настройка модели и перекрестная проверка

4. окончательная подгонка модели для предсказания

5. прогнозы с оценкой неопределенности.

2.1 Данные почвенных наблюдений

Данные о свойствах почвы для этого исследования были получены из Всемирной информационной службы почв IS-RIC (WoSIS), которая предоставляет согласованные стандартизированные данные о профилях почв для всего мира (Batjes et al., 2020). Все данные о почвах, переданные ISRIC для поддержки глобальной картографической деятельности, сначала хранятся в хранилище данных ISRIC вместе с их метаданными (включая имя владельца данных и лицензию, определяющую права доступа). Впоследствии исходные данные импортируются «как есть» в PostgreSQL, после чего они загружаются в саму модель данных WoSIS. После оценки и контроля качества данных (включая проверку согласованности широты и долготы и глубины горизонта/слоя, пометку дубликатов профилей и определение точности географических и атрибутивных данных, а также меток времени),



Фигура 1.Рабочий процесс методологического подхода.

Таблица 1.Описание и единицы измерения свойств почвы.

Свойство почвы	Акроним	Единицы	Сопоставленные единицы	Описание
Объемная плотность	БДОД	кг/дм3	кг/см3	Насыпная масса мелкоземной фракции в сухом состоянии
Катионный обмен вместимость	ЦИК	смоль(с)/кг	ммоль(с)/кг	Способность фракции мелкозема удерживать обменные катионы
Крупные фрагменты	CFVO	см3/100 см3 (объем %)	см3/дм3	Объемное содержание фрагментов крупнее 2 мм во всем грунте
Азот	N	г/кг	г/кг	Сумма общего азота (аммиак, органический и восстановленный азот), измеренная методом Кьельдаля, плюс нитрат-нитрит.
pH (вода)	pH	–	10*	Отрицательный десятичный логарифм активности ионов гидроксония (H+) в воде
Органический углерод концентрация	SOC	г/кг	дг/кг	Гравиметрическое содержание органического углерода в мелкоземной фракции почвы
Фракция текстуры почвы	СТФ	%	г/кг	Гравиметрическое содержание песка, ила и глины в мелкоземистой фракции почвы

*безразмерный.

возможные ошибочные записи самих данных почвенного анализа (Ribeiro et al., 2018). В конечном счете, после окончательной проверки согласованности стандартизированные данные становятся доступными через Центр почвенных данных ISRIC (https://data.isric.org, последний доступ: 20 мая 2021 г.) в соответствии с лицензией, указанной поставщиками данных. В результате не все данные, стандартизированные в WoSIS, находятся в свободном доступе для международного сообщества. Следовательно, в этом исследовании рассматриваются два «источника» точечных данных.

Первый — это последний общедоступный снимок WoSIS (Batjes et al., 2020). Он содержит, среди прочего, данные по химическому составу (органический углерод, общий азот, pH почвы, выделение катионов). способность к изменению) и физические свойства (механический состав почвы (песок, ил и глина), крупные обломки). Снимок включает 196 498 профилей с географической привязкой из 173 стран, представляющих более 832 000 слоев (или горизонтов) почвы, всего более 5,8 миллиона записей. Как правило, поверхностных слоев больше, чем более глубоких. Около 5 % профилей были опробованы до 1960 г., 14 % в период с 1961 по 1980 г., 32 % с 1981 по 2000 г. и 16 % с 2001 по 2020 г.; дата отбора проб неизвестна для 34 % общих профилей (Batjes et al., 2020).

Во-вторых, в дополнение к свободно распространяемым данным, несколько баз данных наблюдений за почвой в нашем репозитории имеют лицензии, предусматривающие, что ISRIC может использовать их только для приложений или визуализации SoilGrids, например, EU-LUCAS (Tóth et al., 2013) и данные о почве для штат Виктория (Австралия). Соответствующие исходные наборы данных были проверены и обработаны с использованием тех же процедур, которые используются для обычного рабочего процесса WoSIS (около 42 000 профилей). В результате около 240 000 профилей были использованы в качестве источника данных для текущего прогона SoilGrids 2020 года, включая более 920 000 наблюдаемых слоев почвы. Во время обработки данных в объединенный набор входных данных были внесены некоторые незначительные исправления, например, дополнительные проверки соответствия глубины.

2.1.1 Свойства почвы

Для целей SoilGrids «почва» — это рыхлый материал толщиной до 2 м в эпидермисе Земли, находящийся в прямом контакте с атмосферой; таким образом, подводные почвы и почвы, подверженные воздействию приливов и отливов, здесь не рассматриваются. Ни материалы глубже 2 м. Это решение имеет последствия для расчетов общих запасов, в частности органического углерода почвы.

В таблице 1 описаны свойства почвы, учитываемые в этой версии SoilGrids: содержание органического углерода, общее содержание азота, pH почвы (измеренный в воде), способность к обмену катионов, гранулометрический состав почвы и доля крупных фрагментов. Эти свойства были смоделированы для шести стандартных интервалов глубин, определенных в спецификациях GlobalSoilMap (Arrouays et al., 2014): 0–5, 5–15, 15–30, 30–60, 60–100 и 100–200 см.

«Подстилочные слои» поверх минеральных почв были исключены из дальнейшего моделирования с использованием следующих допущений. Согласованность глубины слоя (например, последовательное увеличение верхней и нижней глубины для каждого слоя вниз по профилю) в WoSIS проверялась с использованием автоматизированных процедур. В соответствии с действующими международно-признанными соглашениями такие приращения глубины даются как «измеренные от поверхности, включая органические слои и минеральные покровы» (ФАО, 2006; Schoeneberger et al., 2012). Однако до 1993 г. начало профиля (нулевая глубина) устанавливалось в верхней части минеральной поверхности (собственно солума), за исключением случаев, когда «толстые» органические слои, определенные для торфяных почв (ФАО-ИСРИК, 1986 г.), присутствующие на поверхности. Тогда за поверхность почвы принимали верхнюю часть торфяного слоя. Органические горизонты отмечены, как указано выше, а минеральные горизонты указаны ниже, относительно поверхности минералов (Schoeneberger et al., 2012) (стр. 2–6). Насколько это возможно, «поверхностный мусор» поверх минеральных слоев был помечен как вспомогательная (булева) переменная, а также со ссылкой на исходное обозначение почвенного горизонта, если оно было предоставлено, чтобы его можно было отфильтровать при вспомогательных расчетах свойств почвы.

2.1.2 Преобразование текстурных данных

Преобразование было применено к фракциям текстуры следующим образом. Относительное процентное содержание песка, ила и глины можно рассматривать как переменные состава, так как сумма компонентов всегда равна 100 %. Поэтому эти компоненты были преобразованы с использованием преобразования логарифмического отношения привыкания (ALR) с квадратурой Гаусса-Эрмита (Aitchison, 1986). ALR ранее применялся к данным о текстуре почвы (Lark and Bishop, 2007; Akpa et al., 2014; Ballabio et al., 2016; Poggio and Gimona, 2017a), и было показано (Lark and Bishop, 2007), что Преобразованные ALR переменные сохраняют информацию о пространственной корреляции и сохраняют композиционную целостность исходных компонентов. В этом исследовании в качестве переменной знаменателя использовалась глина. Следовательно, две интерполированные компоненты ALR можно определить как

$$\begin{aligned} \text{ALR1} &= \text{журнал}\left(\frac{\text{песок}}{\text{глина}}\right) \\ \text{ALR2} &= \text{журнал}\left(\frac{\text{ил}}{\text{глина}}\right). \end{aligned} \quad (1)$$

2.1.3 Пространственная стратификация наблюдений

Случайное разбиение наблюдений профиля на *n* Складки перекрестной проверки не подходят в этом контексте, учитывая высокую пространственную вариацию плотности наблюдения, поскольку это может привести к смещенным результатам (Brus, 2014). Для таких регионов, как Европа и Северная Америка, существует более четырех профилей на 10 км.2, тогда как для крупных стран Азии, таких как Казахстан, Индия или Монголия, количество доступных профилей все еще весьма ограничено (<один профиль на 100 км2) (подробнее см. Batjes et al., 2020).

Поэтому наблюдения за почвой были пространственно стратифицированы в геодезической области, чтобы гарантировать сбалансированное пространственное распределение в пределах каждой складки перекрестной проверки. Пространственные слои в форме шестиугольников были созданы с помощью икосаэдрической равновеликой сетки Снайдера (ISEAG) с апертурой 3 и разрешением 6, в результате чего было получено 7292 слоя (т.е. шестиугольных ячеек), каждый площадью около 70 000 км².2. Этот ISEAG был создан с помощьюdggridR пакет для языка R (Barnes et al., 2016).

Профили были отнесены к 1 из 10 складок, каждая из которых в равной степени представлена в каждой страте, т.е. в каждой ячейке ранее описанной сетки. Все наблюдения (слои или горизонты), относящиеся к профилю, всегда находились в одной и той же кратности как для калибровки модели, так и для оценки. кареПакет R использовался для разделения мест в складках при сохранении пространственного распределения.

2.2 Ковариаты окружающей среды

Для этой работы было доступно более 400 географических слоев в качестве переменных окружающей среды. Они были выбраны по предполагаемой связи с основными почвообразующими факторами, в том числе с многолетними почвенными условиями, т. е. с фактором «времени». Приложение

предоставляет список продуктов, используемых в качестве ковариат, и их источников. Рассмотренные слои можно сгруппировать следующим образом.

- Климат: температура, осадки, снегопад, облачность, солнечная радиация, скорость ветра;
- экология: биоклиматические зоны и эколого-физиографические районы;
- геология: почвенно-осадочная толща, типы горных пород;
- землепользование и покров: из таких источников, как Европейское космическое агентство (ESA) и Геологическая служба США (USGS);
- морфология высоты и рельефа: включая многочисленные индексы морфологии и классы рельефа;
- вегетационные индексы: такие как нормализованный разностный вегетационный индекс (NDVI), расширенный вегетационный индекс (EVI) и чистая первичная продукция (ЧПП);
- необработанные полосы продуктов Landsat и MODIS;
- гидрография: глобальный уровень грунтовых вод, затопление и протяженность ледников, а также изменение поверхностных вод.

Среднее значение и стандартное отклонение климатических переменных и вегетационных индексов за 15 лет (2001–2015 гг.) были рассчитаны на основе месячных данных, чтобы отразить их сезонную динамику.

Все ковариаты были спроецированы на общую систему отсчета координат (CRS), т. е. гомолосиновую проекцию Гуда для земельных массивов, примененную к датуму WGS84. Эта проекция была выбрана, поскольку среди проекций равной площади, поддерживаемых программным обеспечением с открытым исходным кодом, она является наиболее эффективной для минимизации искажений над сушей (de Sousa et al., 2019). Спроецированные ковариаты были импортированы в ГИС GRASS в виде нормализованной растровой структуры с ячейками 250 м на 250 м. Ковариаты и, следовательно, нанесенные на карту области были ограничены участками земли без застройки, водными и ледниковыми участками с использованием маски, созданной на основе слоя земного покрова EKA за 2015 г. (Buchhorn et al., 2020). При этом свойства городских и подводных грунтов не учитываются.

2.3 Выбор ковариат

Учитывая большое количество доступных слоев окружающей среды, была реализована стандартизированная и воспроизводимая процедура выбора ковариат, используемых для моделирования, чтобы (i) уменьшить избыточность между ковариатами, (ii) получить более экономичную и вычислительно эффективную модель, (iii) снизить риск чрезмерная подгонка (Gomes et al., 2019) и (iv) избежание предвзятой оценки переменной важности (Strobl et al., 2008).

Процедура выбора ковариат состояла из двух этапов: декорреляции и рекурсивного исключения признаков.

2.3.1 Декорреляционный анализ

В качестве начального шага был проведен декорреляционный анализ для уменьшения избыточности информации из более чем 400 слоев окружающей среды. Только ковариантные слои, которые имели парный коэффициент корреляции <знак равно0.85 со всеми другими ковариатами были включены в последующий анализ. Для каждой пары ковариат, коррелирующих выше этого порога, для включения в этап моделирования выбирался только первый в алфавитном порядке. Этот шаг сократил количество исходных ковариат примерно до 150 слоев.

2.3.2 Рекурсивное устранение признаков

Рекурсивное исключение признаков (RFE) (Guyon et al., 2002) — это методология, доказавшая свою эффективность при выборе оптимального набора ковариат для моделей деревьев регрессии (Gomes et al., 2019; Hounkpatin et al., 2018). В этом исследовании процедура RFE, реализованная вкаребыл использован пакет для языка R (Kuhn, 2015), так как он предлагает хороший компромисс между точностью и временем вычислений. Алгоритм начинается с подбора модели с использованием всех ковариат, оценки ее производительности и ранжирования важности ковариата. Затем наименее важные ковариаты удаляются из пула, и снова модель подбирается и оценивается, а наименее важные ковариаты удаляются. Процедура повторяется до пула от 0 до *н*ковариаты. Эта процедура основана на перекрестной проверке вне пакета (OOB) и не проверяет все комбинации ковариатов, но считается одним из самых надежных подходов к выбору ковариатов для таких моделей, как случайные леса (Nussbaum et al., 2018).

Процедура RFE на полном наборе наблюдений и ковариат оказалась бы непомерно сложной с вычислительной точки зрения. Чтобы улучшить вычислительную осуществимость для больших наборов данных, были разработаны дополнительные шаги. Для RFE использовались четыре набора наблюдений, каждый из которых был получен с использованием трех перекрестных проверок (дополнительные сведения см. в разделе 2.1.3): набор 1 содержал наборы с 1 по 3, набор 2 сгибов с 4 по 6, набор 3 сгибов с 7 по 9. и набор 4 содержал кратность 10 и две другие случайно выбранные кратности. На первом этапе процедура RFE откаребыл запущен независимо на каждом наборе с гиперпараметрами модели по умолчанию для алгоритма случайных лесов, как это реализовано врейнджерпакет (т.е.*дерево* как 500 и*попробовать*как округленный квадратный корень числовых переменных). В каждом наборе оптимальное количество и комбинация ковариат автоматически подбирались, когда характеристики модели переставали расти, т. е. когда функция потерь достигала своего минимума. В этом исследовании функция потерь представляла собой среднеквадратичную ошибку OOB (RMSE).

На втором этапе применялась процедура RFE со всеми наблюдениями и всеми ковариатами, выбранными по крайней мере в одном из четырех наборов, использованных на предыдущем этапе. Окончательным набором ковариат был набор, минимизирующий функцию потерь.

2.4 Выбор гиперпараметров и перекрестная проверка

На рис. 2 обобщен подход, использованный для выбора гиперпараметров модели и перекрестной проверки. Более подробная информация представлена в следующих разделах.

2.4.1 Настройка модели и численная оценка

Настройка модели выполнялась с помощью 10-кратной процедуры перекрестной проверки, применяемой к нескольким комбинациям гиперпараметров.

Разное количество деревьев решений (*дерево*параметр) были объединены с различным количеством ковариат, используемых в разбиении дерева (*попробовать*параметр). Количество деревьев постепенно увеличивалось до следующих значений: 100, 150, 200, 250, 500, 750 и 1000. Различные*попробовать*значения были кратны квадратному корню из числа ковариат. Были протестированы четыре множителя, 1 (по умолчанию врейнджер),1.5, 2 и 3. Например, если процедура RFE идентифицировала набор из 50 ковариат,*попробовать* оцениваемые значения составили 7, 11, 14 и 21.

Каждая из полученных комбинаций*дерево*а также*попробовать* параметры использовались для обучения другой модели с наблюдениями из девяти раз. Затем прогнозы оценивались по оставшейся кратности с помощью классических показателей эффективности, т. е. среднеквадратичной ошибки (RMSE) и коэффициента эффективности модели (MEC; Janssen and Heuberger, 1995). MEC равен доле объясненной дисперсии, основанной на линии 1:1 предсказанного и наблюдаемого, которая определяется как 1 минус отношение между остаточной суммой квадратов и общей суммой квадратов. Окончательный выбор гиперпараметров был основан на оптимизации производительности модели и вычислительных ограничений, в данном случае потребления памяти. Например, увеличение*дерево*параметр выше 200 обеспечивал незначительный прирост показателей (обычно менее 0,1 %, здесь не сообщается), но при этом требовал значительно больше памяти и времени вычислений.

Оценка модели основывалась на показателях производительности выбранной комбинации гиперпараметров. Прогнозы в центре шести стандартных интервалов глубины сравнивались с наблюдениями, в которых средняя точка была включена в рассматриваемый интервал.

2.5 Прогноз и количественная оценка неопределенности

2.5.1 Подгонка модели

Окончательная модель для каждого свойства почвы была приспособлена ко всем доступным наблюдениям, ковариатам и гиперпараметрам, выбранным на предыдущих этапах. Глубина наблюдения была включена в модель как ковариата. Он был рассчитан в средней точке опробованного слоя или горизонта.

Модели были получены срейнджерпакет (Wright and Ziegler, 2017), с возможностьюквантрегдля построения квантильных случайных лесов (QRF; Meinshausen, 2006). С этой опцией прогноз не является одним значением, например, средним значением прогнозов из группы деревьев решений в случайный лес, а скорее кумулятивное вероятностное распределение свойства почвы в каждом месте и на каждой глубине.

Для каждого свойства (см. Таблицу 1) и стандартной глубины из спецификации GlobalSoilMap (0–5, 5–15, 15–30, 30–60, 60–100 и 100–200 см) были рассчитаны четыре различных значения для характеристики этого распределение: медиана (0.50 квантиль, $q_{0.50}$), среднее, 0.05 квантиль ($q_{0.05}$) и 0.95 квантиль ($q_{0.95}$), т. е. нижняя и верхняя границы 90-процентного интервала предсказания. Этот интервал неопределенности соответствует описанию в спецификациях Global-SoilMap (Arrouays et al., 2014). Прогнозы рассчитывались для средней точки интервала глубин и считались постоянными для всего интервала глубин.

Чтобы вычислить неопределенность прогноза текстуры почвы, обратное преобразование было применено на уровне прогнозов отдельных деревьев и квантилей распределений прогнозов деревьев, полученных из полученных значений.

2.5.2 Неопределенность

Был рассчитан процент перекрестных проверок, содержащихся в интервале прогнозирования 0,9 (вероятность охвата интервала прогнозирования, PICP) (Шреста и Соломатин, 2006). В идеале PICP близок к 0,9, что указывает на правильную оценку неопределенности. Значение PICP, существенно превышающее 0,9, свидетельствует о том, что неопределенность была недооценена; существенно меньший PICP указывает на то, что он был завышен.

Кроме того, для визуализации неопределенности в виде карты были рассчитаны следующие показатели:

1. 90-й интервал предсказания (PI90)

$$\mathrm{PI90} = q_{0.95} - q_{0.05}; \tag{2}$$

2. отношение межквартильного размаха к медиане (отношение интервалов прогнозирования, PIR):

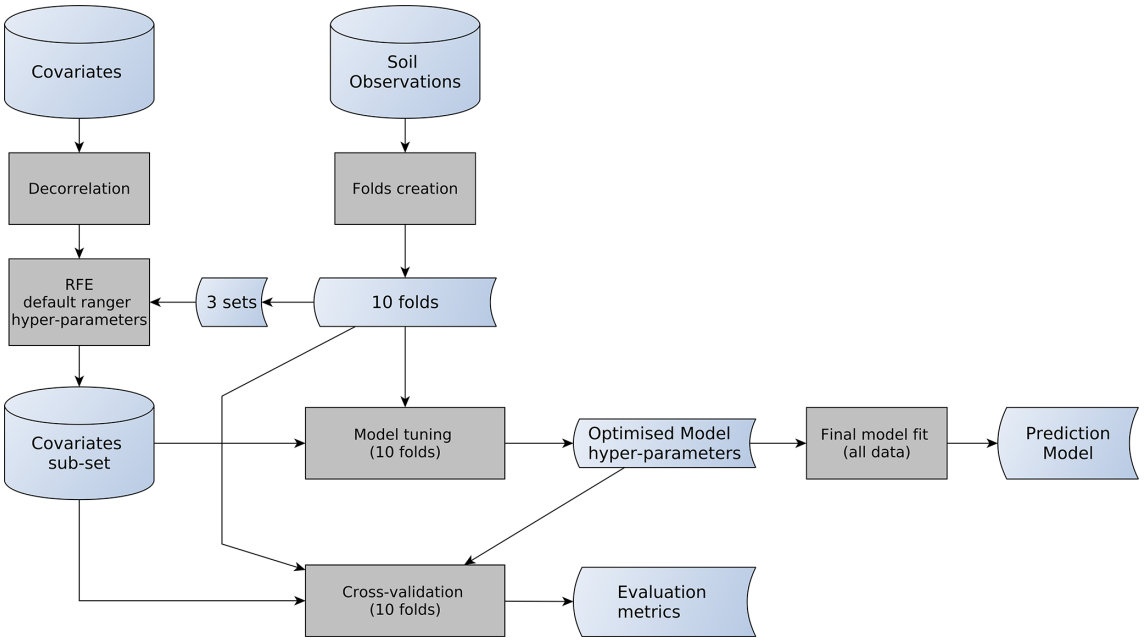
$$\mathrm{ПИР} = \frac{q_{0.95} - q_{0.05}}{q_{0.50}}. \tag{3}$$

2.6 Качественная оценка пространственных закономерностей

Экспертная оценка использовалась для оценки достоверности карт путем сравнения хорошо известных пространственных закономерностей в глобальном, региональном и локальном масштабах с прогнозами SoilGrids (см. раздел 3.4). Очевидно, что это не окончательные оценки, а лишь ориентировочные.

2.7 Программное обеспечение и вычислительная база

SoilGrids требует интенсивного вычислительного рабочего процесса с многочисленными этапами интеграции различного программного обеспечения. Soil-Grids полностью основан на программном обеспечении с открытым исходным кодом, в частности SLURM (Yoo et al., 2003) для управления заданиями, GRASS GIS (GRASS Development Team, 2020) для управления данными и листами и статистического программного обеспечения R (R Core Team). , 2020) для подбора модели и статистического анализа.



Фигура 2.Подробный рабочий процесс для выбора гиперпараметров и перекрестной проверки.

Прогнозы были рассчитаны в высокопроизводительном вычислительном кластере. Динамическая система географических листов была разработана с помощью GRASS GIS, чтобы максимально использовать память для каждого задания. Технические детали этой схемы распараллеливания приведены в de Sousa et al. (2020).

Прогнозы были умножены на коэффициент преобразования 10 или 100, чтобы сохранить требуемую точность при использовании целочисленного типа в файле geotiff для уменьшения занимаемого места на диске. Применение коэффициента преобразования привело к отображению слоев с единицами измерения, отличными от единиц входных наблюдений (см. Таблицу 1).

Общее время вычислений с выбранными ковариатами и гиперпараметрами различалось в зависимости от свойства. В среднем полное вычисление 24 карт (среднего значения и трех квантилей для каждой из шести стандартных глубин) для одного свойства, включая (i) RFE, (ii) обучение модели и (iii) прогноз, заняло около 1500 часов ЦП. . На предсказание приходилось около двух третей всего времени.

3. Результаты и обсуждение

3.1 Входные наблюдения за почвой

В таблице 2 представлено распределение традиционных наблюдений за почвой для каждого свойства почвы по интервалу глубины. Таблица B1 в Приложении B показывает количество наблюдений по биоклиматическим регионам.

На рисунках 3 и 4 показаны примеры плотности наблюдения данных калибровки почвы для двух свойств почвы, $pH_{вода}$и доля крупных фрагментов, которые показывают большую разницу в плотности.

Как указано, количество наблюдений для каждого свойства сильно различается в зависимости от глубины и биоклиматического региона, при этом более высокая плотность наблюдается для Северной Америки и Европы. Как правило, наблюдений за сельскохозяйственными угодьями больше. Кроме того, доступные профили были сопоставлены за несколько десятилетий, около 62 % данных относятся к 1960–2020 гг.; примерно для 34 % профилей время отбора проб неизвестно. Как указано Batjes et al. (2020), в принципе, возраст наблюдений следует учитывать в процессе картирования с помощью ковариационных слоев для периодов времени, соответствующих датам отбора проб, особенно для свойств почвы, на которые легко влияют изменения в землепользовании или методах управления. Однако для таких так называемых «динамических» свойств почвы, таких как pH и содержание органического вещества в почве, мы считаем, что пространственная вариация будет намного больше, чем временная вариация, так что неучет возраста наблюдений не сильно повлияет на карту. Кроме того, трудно или невозможно найти сопоставимые ковариаты, в частности ковариаты, полученные с помощью дистанционного зондирования, для каждого периода времени. В будущих оценках следует учитывать пространственно-временные отношения (Heuvelink et al., 2020).

В этом исследовании рассматриваются стандартизированные данные для примерно 240 000 профилей, полученные из WoSIS. Это более чем на 60 000 профилей больше, чем было учтено при компиляции данных, лежащих в основе предыдущих прогонов SoilGrids (Hengl et al., 2017b), что дает существенную новую информацию для калибровки новых глобальных моделей. Однако, как указано, по-прежнему существуют значительные географические пробелы (например, засушливые районы, бореальные регионы и «лесные» почвы). Некоторые из них связаны с физической удаленностью или недоступностью некоторых регионов, в то время как другие связаны с тем, что многие наборы почвенных данных до сих пор не доступны или не могут быть получены.

не могут быть переданы по разным причинам, как описано Arrouays et al. (2017).

В предыдущей версии SoilGrids (Hengl et al., 2017b) синтетические наблюдения были случайным образом размещены в регионах с небольшим количеством наблюдений или без них, например, в Сахаре и на Аравийском полуострове. Этот подход заслуживает дальнейшего изучения, включая информацию, полученную из других региональных наборов данных, мнения экспертов и перенос знаний из аналогичных областей в соответствии с *гомозем* концепции (Mallavan et al., 2010), предполагающей сходство факторов почвообразования по регионам. Тем не менее, SoilGrids уже неявно включает концепцию Homosoil, если в любой точке мира имеется достаточно наблюдений в данной почвообразующей среде. Поэтому в эту версию SoilGrids не были включены синтетические наблюдения («псевдоточки»), в том числе из-за отсутствия уверенности в точности синтетических данных.

В будущих исследованиях будет актуальным выявление заранее участков мира с низкой плотностью наблюдений, которые еще не представлены высокой плотностью наблюдений в других районах с аналогичными факторами почвообразования. Затем можно было бы создать набор синтетических профилей для описания этих территорий, проконсультировавшись с учеными-почвоведами, хорошо осведомленными о почвах и свойствах почв этих территорий.

3.2 Настройка модели и выбор гиперпараметров

Гиперпараметры модели, выбранные для каждого свойства, представлены в таблице 3.

Количество ковариат, выбранных с использованием двухэтапного подхода к выбору ковариат, было довольно небольшим по сравнению с полным набором (таблица 3), что привело к более экономным моделям. На рис. 5 показаны два примера функции потерь для RFE для двух свойств почвы с разным количеством и распределением входных наблюдений. В обоих случаях наблюдается явное улучшение производительности при использовании от 15 до 20 ковариатов. Кривая достигает минимума функции потерь, а затем остается на плато с небольшим спадом после выявленного минимума.

Все окончательные модели были обучены максимум с 200 деревьями решений, при превышении которых прирост производительности заметно не увеличивался.

The *попробовать*Параметр в основном зависел от количества ковариатов и всегда был в 1,5-2 раза больше квадратного корня из числа ковариат, что является значением по умолчанию, предоставляемым обычными пакетами случайного леса, такими как *рейнджер* (Райт и Зиглер, 2017). Это подтверждает необходимость определения оптимальных гиперпараметров модели, особенно при работе с большим количеством входных данных (Nussbaum et al., 2018), как в данном случае.

3.3 Количественная оценка

Результаты перекрестной проверки обобщены в таблице 4, где представлены среднеквадратическая ошибка (RMSE) и коэффициент эффективности модели (MEC). MEC варьируется от мини-

от 0,31 для крупных фрагментов до максимума 0,74 для BDOD. Глина моделируется хуже, чем два других класса размеров частиц. Это может быть результатом выбранной трансформации ALR, в которой в качестве знаменателя использовалась глина (Lark and Bishop, 2007). Показатели среднего всегда были лучше или равны показателям медианы для всех свойств.

В целом эти показатели соответствуют исследованиям DSM на континенте или в крупных регионах (Keskin and Grunwald, 2018). Однако они немного ниже, чем представленные Hengl et al. (2017б). Последнее различие может быть объяснено более разумным подходом к перекрестной проверке, используемым в настоящее время, с пространственно сбалансированными складками и всеми наблюдениями, относящимися к одному и тому же профилю в одной и той же складке. Это предотвращает использование данных одного и того же профиля как для калибровки, так и для числовой оценки.

Таблица 4 показывает, что модели с большим количеством сохраненных ковариат (таблица 3) имеют лучшие прогностические характеристики. Однако эти модели также являются моделями с наибольшим количеством наблюдений (табл. 2). Рассматриваемые свойства почвы также различны. Поэтому из этого наблюдения нельзя сделать общий вывод.

Таблица 5 показывает MEC для средних прогнозов по интервалу глубины. Производительность снижалась с глубиной, как и во многих других исследованиях DSM (Keskin and Grunwald, 2018). Эта закономерность может быть объяснена главным образом ослаблением связей между экологическими слоями и почвенными свойствами более глубоких слоев.

В этом исследовании вертикальное измерение изменчивости почвы учитывалось только с использованием глубины наблюдения в качестве ковариации. Недавние публикации (Ma et al., 2021; Nauman and Duniway, 2019) указывают на то, что такой подход может быть слишком упрощенным или привести к проблемам с непротиворечивостью прогнозируемой последовательности глубин. Это может быть верно для локальных наборов данных, в которых краткосрочная пространственная изменчивость имеет ту же величину, что и вертикальная изменчивость. Необходимы дальнейшие исследования для оценки влияния использования глубины в качестве ковариации на глобальные наборы данных и модели. Альтернативы, такие как 3D-сглаживатели (Poggio and Gimona, 2017b) или геостатистические модели, использующие 3D-пространственную автокорреляцию, заслуживают изучения в дальнейших исследованиях.

В Таблице 6 приведены значения PICP в глобальном масштабе и по прогнозируемым интервалам глубин. Большинство значений находится в диапазоне от 0,88 до 0,92, что указывает на то, что интервалы прогнозирования, полученные с помощью QRF, являются реалистичным представлением неопределенности прогнозирования, поскольку ожидаемое значение для 90% интервала прогнозирования составляет 0,90. Исключением являются модели для грубых фрагментов с более высокими значениями около 0,95, что указывает на завышенную оценку неопределенности прогноза. Компоненты текстуры имеют значения с большим разбросом, от 0,78 до 0,80 для песка и ближе к 0,96 для ила и глины. Это указывает на возможную недооценку интервалов прогнозирования для песка и завышение оценки для ила и глины. Эти результаты могут быть связаны с диапазоном этих свойств во входных наблюдениях. Метод преобразования, используемый для получения интервалов прогнозирования для компонентов текстуры, также может быть фактором, влияющим на результат. Целесообразно дальнейшее изучение причин.

Таблица 2.Количество наблюдений на стандартный интервал глубины для каждого свойства почвы. См. Таблицу 1 для сокращений и единиц рассматриваемых свойств почвы.

Интервал глубины	БДОД	ЦИК	CFVO	Н	pH	SOC	СТФ
0–5 см	8122	20 576	15 541	27192	44 049	48 616	42 983
5–15 см	19 817	49 463	66 833	82 856	146 677	148 918	155 302
15–30 см	17 819	40 673	35 254	39 568	91 326	91 682	98 659
30–60 см	27 146	63 444	56 755	48 804	141 812	122 338	140 353
60–100 см	23 130	58 038	50 912	36 946	131 172	102 687	12 7073
100–200 см	23 396	66 236	49 995	28 135	129 373	92 327	116 847

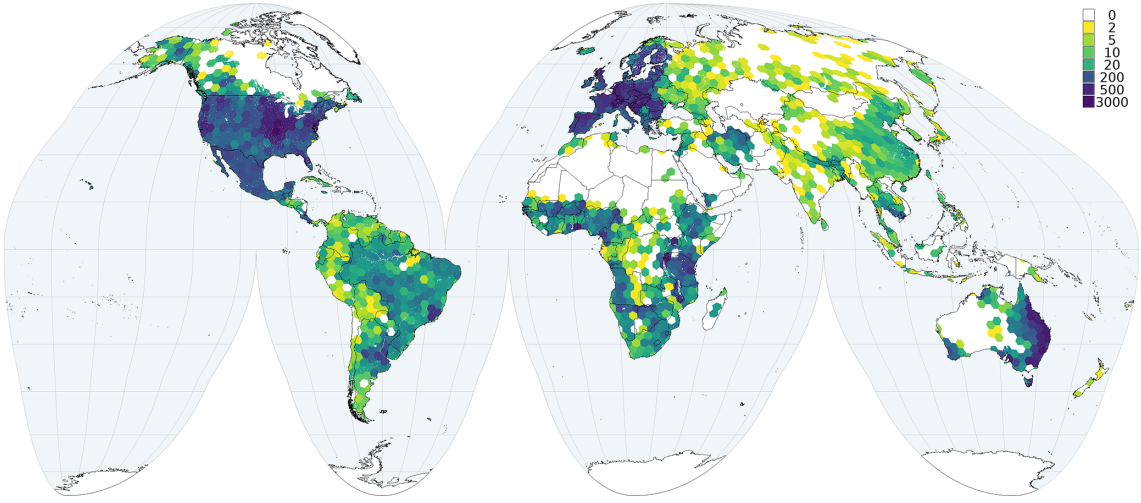


Рисунок 3.Количество наблюдений на ячейку сетки (70 000 км2) для pH почвывода.

Таблица 3.Гиперпараметры для каждого рассматриваемого свойства почвы. См. Таблицу 1 для сокращений и единиц рассматриваемых свойств почвы.

	Количество ковариаты	Количество деревья	попробовать
БДОД	40	200	12
ЦИК	25	200	10
CFVO	20	200	6
Н	30	200	10
pH	32	200	9
SOC	40	200	12
Текстура АЛР I	25	150	10
Текстура АЛР II	27	150	10

Ключевой проблемой для приложений DSM, использующих устаревшие данные о почве, является оценка результатов. Аспект оценки, который сравнивает фактические значения с прогнозируемыми, представляет собой числовую (или статистическую) оценку, часто называемую «валидацией» в литературе по DSM; Орескес (1998) и Росситер (2017) объясняют, почему термин «оценка» предпочтительнее для общего процесса оценки успешности моделей, включая модели DSM. Наилучший подход к числовой оценке состоит в том, чтобы иметь независимый набор данных, полученный с помощью вероятностной выборки с использованием известный план выборки (Brus et al., 2011; Brus, 2014). Однако это невозможно, когда доступны только старые данные. В этом случае часто используется так называемый подход «перекрестная проверка». Это необходимо настроить, чтобы избежать переоценки или недооценки показателей числовой оценки, особенно в случае больших различий в плотности наблюдений, т. е. кластеризованных пространственных наблюдений. Это особенно важно в глобальном масштабе, поскольку распределение наблюдений за почвой неравномерно по всему земному шару. Нельзя гарантировать, что числовые метрики оценки, полученные в результате перекрестной проверки, являются объективными оценками истинных числовых метрик оценки, т. е. теми, которые были бы рассчитаны на вероятностной выборке всей совокупности. Также невозможно количественно определить, насколько близки оценки метрик перекрестной проверки к истинным метрикам, поскольку невозможно получить доверительные интервалы (Brus et al., 2011). При использовании перекрестной проверки важно предотвратить завышенные или заниженные оценки. Например, вполне вероятно, что ошибки прогнозирования меньше в областях, где плотность выборки выше. Из-за высокой плотности выборки такие районы будут чрезмерно представлены в выборке, поскольку процент точек перекрестной проверки в сгруппированных районах будет выше, чем процент общей площади суши, охватываемой этими районами. На результаты стандартной перекрестной проверки сильно повлияют характеристики в кластеризованных При использовании перекрестной проверки важно предотвратить завышенные или заниженные оценки. Например, вполне вероятно, что ошибки прогнозирования меньше в областях, где плотность выборки выше. Из-за высокой плотности выборки такие районы будут чрезмерно представлены в выборке, поскольку процент точек перекрестной проверки в сгруппированных районах будет выше, чем процент общей площади суши, охватываемой этими районами. На результаты стандартной перекрестной проверки сильно повлияют характеристики в кластеризованных При использовании перекрестной проверки важно предотвратить завышенные или заниженные оценки. Например, вполне вероятно, что ошибки прогнозирования меньше в областях, где плотность выборки выше. Из-за высокой плотности выборки такие районы будут чрезмерно представлены в выборке, поскольку процент точек перекрестной проверки в сгруппированных районах будет выше, чем процент общей площади суши, охватываемой этими районами. На результаты стандартной перекрестной проверки сильно повлияют характеристики в кластеризованных такие районы будут чрезмерно представлены в выборке, поскол

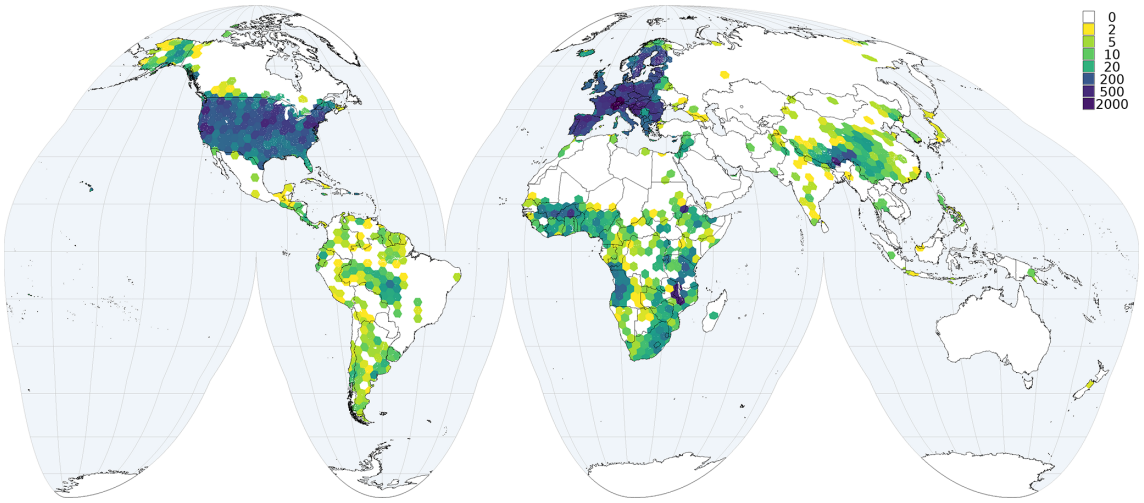


Рисунок 4.Количество наблюдений на ячейку сетки (70 000 км2) для крупных фрагментов.

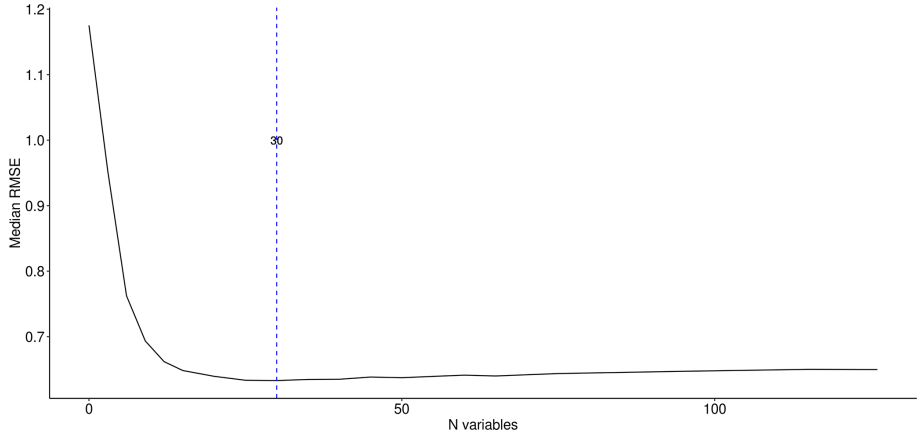


Рисунок 5.Пример функции потерь (RMSE), используемой на шаге RFE выбора ковариат.

области. Используя пространственную перекрестную проверку, предложенную Meyer et al. (например, 2018 г.), где гарантируется, что данные калибровки никогда не будут слишком близки к точке перекрестной проверки, с другой стороны, это может привести к чрезмерно пессимистичным результатам. Чтобы решить некоторые из этих проблем, в этом исследовании было принято практическое решение, в котором складки были созданы, чтобы гарантировать пространственно сбалансированное распределение между сгибами перекрестной проверки, поддерживая одинаковую плотность входных данных в каждом сгибе, чтобы они представляли приблизительно такое же население.

Хотя процедура численной оценки, используемая в этой работе, учитывает пространственное распределение наблюдений и их плотность, возможно дальнейшее улучшение как обучения модели, так и ее оценки. Например, весовые коэффициенты, присвоенные наблюдениям в районах с большим объемом выборки, могут быть уменьшены. В Соединенных Штатах и крупных регионах Европы и Австралии имеется большое количество наблюдений, которые можно было бы уменьшить, чтобы усилить пространственную надежность процедуры оценки. Методы декластеризации или устранения предвзятости (Goovaerts, 1997; Deutsch and Journel, 1998) успешно применялись в других геостатистических исследованиях и могут быть адаптированы к глобальному картографированию почвы. Создание складок также может быть изменено с учетом плотности наблюдений.

3.4 Качественная оценка пространственных закономерностей

В глобальном масштабе воспроизводятся хорошо известные закономерности, и могут быть распознаны типичные свойства, связанные со многими эталонными группами почвенных ресурсов Всемирной эталонной базы почвенных ресурсов (WRB) (Рабочая группа IUSS WRB, 2015).

Например, на карте pH выделяются большие области солонцеватых почв (Солонец, Солочак), сильно выветрелых почв (например, акрисоли, алисоли, плитосоли), кислых лесных почв (например, подзоли) и молодых почв из известняковых ледниковых отложений (например, лювисоли). Низкий pH Andosols (например, Тихоокеанский северо-запад США, Япония, Новая Зеландия) также представлен правильно. Компоненты текстуры (класс размера частиц – PSC) правильно идентифицируют более илистые дельты (например, Желто-

Таблица 4.Глобальные результаты перекрестной проверки как для средних, так и для медианных прогнозов. См. Таблицу 1 для сокращений и единиц рассматриваемых свойств почвы.

Имущество	Среднеквадратическая ошибка (медиана)	СКО (среднее)	МЭК (медиана)	МЭК (среднее)
БДОД	0,19	0,19	0,73	0,74
ЦИК	11.01	10,69	0,40	0,43
CFVO	13.46	12,69	0,22	0,31
Н	2,62	2,50	0,47	0,52
pH	0,78	0,77	0,67	0,68
SOC	39,67	36,48	0,37	0,47
Песок	0,19	0,18	0,51	0,54
Ил	0,13	0,13	0,60	0,62
Глина	0,13	0,13	0,42	0,43

Таблица 5.МЕС на слой глубины для средних прогнозов. См. Таблицу 1 для сокращений и единиц рассматриваемых свойств почвы.

Глубинный слой	БДОД	ЦИК	CFVO	Н	pH	SOC	Песок	Ил	Глина
0–5 см	0,78	0,46	0,33	0,65	0,69	0,55	0,59	0,71	0,45
5–15 см	0,74	0,42	0,35	0,41	0,66	0,39	0,58	0,64	0,42
15–30 см	0,72	0,39	0,33	0,44	0,68	0,38	0,57	0,68	0,42
30–60 см	0,70	0,42	0,31	0,46	0,68	0,38	0,54	0,62	0,41
60–100 см	0,61	0,41	0,29	0,48	0,68	0,42	0,50	0,57	0,40
100–200 см	0,59	0,45	0,29	0,49	0,67	0,59	0,48	0,54	0,40

Янцзы, Ганг-Брахмапутра), широкие речные равнины (например, По, Дунай, Миссисипи, плита Рио, верхняя часть Амазонки), лёссовые районы (например, Средний Запад США, Северо-Западная Европа, Украина) и песчаные равнины Северной Германии и Польши. Карта емкости катионного обмена (CEC) четко определяет большие регионы сильно выветрелых глин (например, юго-восток США и Китая, центральная Бразилия) и глин с высоким CEC 2:1 (например, Vertisols «черного хлопка» на плато Декан и в Судане). . Эта карта вместе с картами концентрации органического углерода в почве (SOC) идентифицирует большие регионы Histosols (например, северная Канада, Шотландия, Сибирь, Борнео). Карта запасов SOC идентифицирует глубокие Histosols и прохладные, влажные регионы (например, тихоокеанское северо-западное побережье Северной Америки, Ирландия, южная часть Чили).

Ясны также многие региональные закономерности, например, переход pH от Сахары через Сахель к побережью Западной Африки и переходы PSC от ледниковой доли Де-Мойна к прогляциальным лёссовым отложениям в Айове (США), а также переход PSC от глинистых морские отложения вдоль побережья Северного и Балтийского морей через песчаные равнины до лессового пояса центральной Германии. На карте CEC обозначены контрастирующие области Vertisols (например, «черные пояса» в Алабаме-Миссисипи и Техасе, США). Карта с грубыми фрагментами показывает подробную схему бассейна и хребта на западе Соединенных Штатов и региона хребтов и долин в Аппалачах.

Однако в местном масштабе предварительная оценка SoilGrids в Соединенных Штатах по сравнению с версией национальной подробной почвенно-географической базы данных gSSURGO (NRCS National Soil Survey Center, 2016), основанной на подробном полевом обследовании, показывает, что SoilGrids может не учитывают локальные переходы материнского материала, например, осадочные фации прибрежных равнинных морских отложений, а также ледниковые особенности, такие как прогляциальные озерные отложения и реликтовые береговые линии, так что локальная картина PSC неточна, иногда порядка 20 %. –30 % класса крупности.

Например, на рис. 6a показана прогнозируемая концентрация песка в слое 0–5 см примерно на 50×Площадь 50 км в центральной Пенсильвании (США), варьируется примерно от 25 % (самый темный красный) до 80 % (самый темно-синий). Важные локальные различия очевидны: низкие концентрации песка в глинистых почвах в известняковых долинах юго-восточного простирания и высокие концентрации в почвах от ледниковых до развитых на песчаниках на севере, а также в остаточных почвах на устойчивых песчаниковых грядах провинция Аппалачи на юге хребта и долины. Они в целом согласуются с детальным исследованием почвы.

В детальном масштабе (250 м пикселей) SoilGrids обычно показывает мелкие детали, которые не всегда кажутся связанными с очевидными различиями в ландшафте или землепользовании, когда карта рассматривается как наложение на землю в Google Earth. Например, на рис. 6b показаны детали прогнозируемой концентрации песка в слое 0–5 см примерно на 3×3 км площади предыдущего рисунка. Влияние одних ковариат на разрешение 1 км, а других на 250 м очевидно, но причина

Таблица 6.Вероятность охвата интервала прогнозирования, глобальная и по прогнозируемому интервалу глубины. См. Таблицу 1 для сокращений и единиц рассматриваемых свойств почвы.

Имущество	Глобальный	[0, 5]	[5, 15]	[15, 30]	[30, 60]	[60, 100]	[100, 200]
БДОД	0,90	0,89	0,91	0,91	0,91	0,90	0,88
ЦИК	0,88	0,89	0,90	0,88	0,88	0,88	0,87
CFVO	0,95	0,96	0,95	0,95	0,95	0,94	0,94
H	0,92	0,91	0,92	0,93	0,92	0,92	0,92
pH	0,90	0,91	0,91	0,90	0,91	0,90	0,89
SOC	0,92	0,91	0,92	0,92	0,92	0,92	0,92
Песок	0,79	0,82	0,82	0,80	0,78	0,78	0,78
Ил	0,96	0,95	0,96	0,96	0,96	0,96	0,96
Глина	0,96	0,96	0,96	0,96	0,95	0,95	0,96

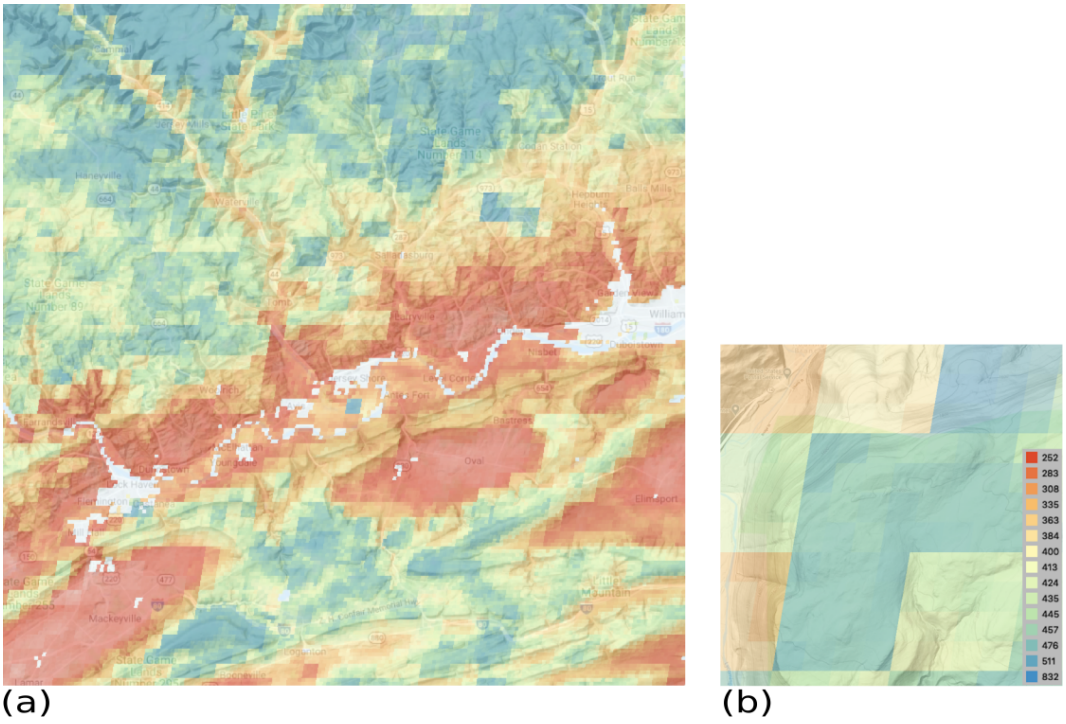


Рисунок 6.Прогнозируемая концентрация песка, %, 0–5 см, наложение на грунт в © Google Earth. (а)Обзор; центр≈−77◦14′E, 41◦14′N, недалеко от Джерси-Шор, Пенсильвания. (б)Деталь; центр≈−76◦56′E, 41◦33′H.

мелкомасштабных различий нет. Этот район имеет схожий литологический состав, рельеф и растительный покров (второстепенный густой лес), за исключением узкой долины на северо-западной окраине, но прогнозы совсем другие.

В этом контексте следует понимать, что прогнозы SoilGrids250m не предназначены для использования в детальном масштабе, т. е. на субнациональном или местном уровне, поскольку национальные поставщики данных часто имеют доступ к более подробным наборам точечных данных и ковариативным слоям для своей страны, чем это было ранее. предоставлены точечному набору данных, на котором основан SoilGrids250m (Chen et al., 2020; Roudier et al., 2020; Vitharana et al., 2019; Liu et al., 2020).

3.5 Неопределенность прогноза

Как правило, районы с наименьшим количеством выборок представляют самые высокие неопределенности прогноза, выраженные PICP. На рисунках 7 и 8 показан пример для двух свойств и глубин (карты для всех свойств и глубин доступны по адресу https://data.isric.org). На рисунках 9 и 10 показан пример, представляющий квантили для pH.водадля слоя 60–100 см. Север России, а также центр и северо-запад Канады представляют собой большие регионы, для которых имеется мало почвенных наблюдений; поэтому распределения предсказаний шире, чем в более плотно отобранных областях. Однако эти шаблоны различны для разных свойств. Например, засушливые районы на самом деле имеют самые узкие диапазоны предсказания pH.вода. Диапазон неопределенности

десять широких для свойств и регионов с более широким диапазоном моделируемого свойства. Это можно объяснить тем, что подход к моделированию работает более точно в ограниченном диапазоне вариантов. Эти регионы также имеют большую локальную пространственную изменчивость с большими трудностями для прогнозов.

Сообщение о неопределенности является открытой проблемой (Arrouays et al., 2020). Неопределенность должна предоставлять информацию политикам и другим заинтересованным сторонам, а не только ученым и разработчикам моделей. Карты, рассчитанные с помощью уравнения. (3) являются первым шагом в этом направлении, но необходимо понимать их ограничения. Для свойств, которые имеют значения, равные нулю или близкие к нулю, например крупные фрагменты, они не обеспечивают полностью точную оценку неопределенности. Использование классов неопределенности может стать еще одним шагом в помощь заинтересованным сторонам в предметной области.

3.6 Ограничения и перспективы

Это исследование представляет собой значительную попытку предоставить глобально согласованный продукт с использованием набора точечных данных, доступного для IS-RIC, большого количества соответствующих ковариат и некоторой оптимизации хорошо зарекомендовавшего себя метода машинного обучения в пределах практических вычислений. Тем не менее ясно, что этот продукт имеет некоторые ограничения, которые будут учтены в дальнейшей работе.

Во-первых, существует постоянно расширяющаяся группа новых ковариат, которые могут помочь объяснить и смоделировать пространственное изменение свойств почвы. Продукция, полученная в результате наблюдения Земли, особенно актуальна в этом отношении и значительно улучшилась за последнее десятилетие. Например, миссии Sentinel Европейского космического агентства (как оптические, так и радиолокационные) предоставляют данные с высоким разрешением, которые, как было показано, улучшают характеристики модели DSM.

Во-вторых, фундаментальной проблемой является отсутствие хорошо распределенных точечных наблюдений в пространстве географических свойств и признаков почвы. Для возможного рассмотрения в рабочем процессе WoSIS, который предоставляет точечные данные, будут запрашиваться дополнительные данные о почве для пока недопредставленных регионов, например северных бореальных регионов, которые сопоставляются Международной сетью почвенного углерода (Malhotra et al., 2019). поддержка усилий по картированию SoilGrids. Этим усилиям будет способствовать предоставление несколькими поставщиками данных хотя бы репрезентативной части своих точечных данных в WoSIS по соответствующей лицензии. Также важно учитывать распределение наблюдений в ковариативном пространстве, чтобы свести к минимуму проблемы, связанные с прогнозами в неизвестных областях пространства признаков (Meyer and Pebesma, 2020).

В-третьих, методы DSM находятся в стадии активной разработки, как новые методы, так и улучшения существующих методов. Использование моделей на основе дерева решений в DSM стало довольно распространенным в последние годы. Такие модели, как случайные леса, XGBoost или Cubist, как правило, дают лучшие результаты, чем большинство методов множественной линейной регрессии с разумными затратами на вычисления (Khaledian and Miller, 2020). Однако такие методы, как искусственные нейронные сети, обещают дальнейшее улучшение характеристик модели, если количество и распределение данных поддерживают эти очень сложные модели. В частности, это касается сверточных или рекурсивных нейронных сетей (глубокое обучение). Однако эти методы сопряжены с вычислительными трудностями из-за количества обучающих данных, необходимых для достаточно точного выполнения DSM, особенно при работе в глобальном масштабе со средним и высоким разрешением.

В-четвертых, правильный метод перекрестной проверки является еще одним важным аспектом при рассмотрении вопроса о том, как оценить и улучшить характеристики модели. В частности, необходимо дополнительно изучить пространственную перекрестную проверку и декластеризацию данных.

В-пятых, это исследование рассматривало только моделирование некоторых первичных свойств почвы, определенных и описанных в спецификациях GlobalSoilMap. Необходима дополнительная работа для получения карт мощности почвы (зоны корнеобразования, почвообразующего слоя или реголита), свойств почвы, полученных с помощью функций переноса педообразования, например, гидрологических свойств почвы, таких как насыщенная гидравлическая проводимость (Пачепски и Роулз, 2004), и комплексных свойств. которые зависят от множества первичных свойств, например запасов углерода. Эти слои являются важными исходными данными для моделирования и картирования функций почвы в настоящем и будущем, а также для поддержки моделирования системы Земля (Luo et al., 2016; Dai et al., 2019).

В-шестых, рекомендуется количественная оценка неопределенности, которая становится все более распространенной в исследованиях DSM. Насколько нам известно, эта работа впервые представила его в глобальном масштабе. Хотя предоставление квантилей упоминается в спецификациях GlobalSoilMap, представление и передача неопределенности конечным пользователям и заинтересованным сторонам остается важной областью исследований, требующей дальнейшего изучения. Соответствующие интервалы неопределенности, как с точки зрения приемлемости для пользователя, так и с точки зрения возможности моделирования, также должны быть исследованы.

Наконец, следует дополнительно изучить интеграцию высокоавтоматизированных рабочих процессов с мнением экспертов. Продукты DSM используют статистические модели для описания почв, и важно учитывать знания и опыт почвоведов, по крайней мере, в цикле оценки, если не в рамках самого моделирования. Мы предприняли первую попытку сделать это в разделе «Качественная оценка» выше, но у нас нет метода эффективного включения экспертных наблюдений в рабочий процесс.

4. Выводы

В этом исследовании представлено и обсуждается создание глобальных карт свойств почвы, реализованных в продукте SoilGrids 2.0, с перекрестной проверкой, выбором гиперпараметров и количественной оценкой неопределенности с использованием наилучших доступных (общих) данных профиля почвы для мира. В частности, в исследовании описывается надежный и воспроизводимый рабочий процесс DSM, решающий задачи моделирования глобальных данных:

1. неоднородное пространственное распределение входных почвенных наблюдений;