УДК 631.421

КРУПНОМАСШТАБНОЕ ЦИФРОВОЕ КАРТОГРАФИРОВАНИЕ СОДЕРЖАНИЯ ОРГАНИЧЕСКОГО УГЛЕРОДА ПОЧВ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

© 2018 г. А. В. Чинилин[1], И. Ю. Савин[2,*]

[1]РГАУ-МСХА им. К.А. Тимирязева,
Россия, 127550 Москва, ул. Тимирязевская, 49
[2]Почвенный институт им. В.В. Докучаева,
Россия, 119017 Москва, Пыжевский пер., 7
[*]e-mail: savin_iyu@esoil.ru

Приведены результаты цифрового картографирования содержания органического углерода в пахотных горизонтах почв и оценки точности получаемых моделей с использованием методов машинного обучения для участка Среднерусской возвышенности Воронежской области. Цифровое картографирование основывалось на 22 точках почвенного опробования, используемых для обучения и проверки моделей, а также на нескольких наборах переменных-предикторов, в качестве которых выступали цифровая модель рельефа, производные от нее и данные дистанционного зондирования различного пространственного разрешения. Для построения моделей пространственного варьирования исследуемого свойства использовали несколько методов, основанных на деревьях решений: ансамбль деревьев решений, бустинг регрессионных деревьев и байесовские регрессионные деревья. Оценку точности полученных картографических моделей определяли методом перекрестной проверки, при этом в качестве показателей точности использовали коэффициент детерминации, среднюю абсолютную ошибку и корень среднеквадратичной ошибки. По результатам моделирования выявлено, что с использованием переменных-предикторов, представленных цифровой моделью рельефа, ее производными и данными Landsat 8 удалось получить более устойчивые модели, причем коэффициент детерминации изменяется от 0.6 до 0.7, $RMSE_{cv}$, т.е. ошибка прогноза от 0.5791 до 0.6520. Лучшая модель получена с помощью метода байесовских регрессионных деревьев; тогда как для переменных-предикторов, представленных цифровой моделью рельефа, ее производными и данными Sentinel 2 – от 0.47 до 0.55, ошибка прогноза от 0.7031 до 0.7909. Выявлено, что в описанных моделях по различным наборам данных наиболее значимыми оказывались разные переменные-предикторы.

Ключевые слова: пространственное прогнозирование, цифровая модель рельефа, метод ансамблей деревьев решений, бустинг
DOI: 10.19047/0136-1694-2018-91-46-62

ВВЕДЕНИЕ

Для большинства сельскохозяйственных коллективов, арендных и фермерских хозяйств материалы крупномасштабных почвенных обследований были составлены более 30–40 лет назад областными проектными институтами по земельным ресурсам и землеустройству (ГИПРОЗЕМ) и сейчас нуждаются в обновлении, корректировке и актуализации. За последние годы накоплены новые знания о распространении и генезисе почв, появились новые методы и технологии цифровой почвенной картографии (**ЦПК**), геоморфометрии и педометрики, что создает предпосылки по переработке и обновлению устаревших материалов крупномасштабных почвенных обследований. Методология ЦПК развивает гипотезу В.В. Докучаева о почве, как функции от факторов почвообразования (Флоринский, 2012). В качестве центральной идеи ЦПК лежит моделирование пространственной дифференциации почвенного покрова на основе анализа почвенно-ландшафтных связей. На сегодняшний день наиболее распространенной моделью ЦПК является модель *scorpan* (McBratney et al., 2003):

$$S = f(s, c, o, r, p, a, n) + \varepsilon, \quad (1)$$

где S – класс почв/свойство, s – другие характеристики почвы, c – климат (локальные климатические характеристики), o – организмы, растительность, фауна, r – рельеф (морфометрические характеристики), p – характер почвообразующей породы, литология, a – возраст, время, n – пространственное положение, ε – ошибка прогноза/предсказания. Эмпирическая функция f в модели *scorpan* – это вероятностно-статистические модели, методы интерполяции и машинного обучения. Модель *scorpan* – это расширенное выражение модели *clorpt* (Jenny, 1941).

В последние годы использование методов машинного обучения, таких как деревья решений или их ансамбль, бустинг, нейронные сети и иные приобретает популярность применительно к методологии ЦПК. Такие методы подходят для исследования сложных и нелинейных связей между свойствами почв или таксационными единицами почв и независимыми переменными (предикторами, ковариатами) (Bui et al., 2006; Hengl et al., 2015, 2017). Многие наиболее популярные методы машинного обучения представляют собой ансамбли. Например, ансамбль деревьев решений или бустинг

регрессионных деревьев – методы, которые получают набор “слабых учителей” и создают “сильного учителя”.

Деревья решений все чаще используются при картографировании классов почв/свойств почв из-за их потенциала и преимуществ в пространственной дифференциации (Lagacherie, Holmes, 1997; Grinand et al., 2008; Taghizadeh-Mehrjardi et al., 2012; Arrouays et al., 2014, 2018; Жоголев, 2016). Бустинг регрессионных деревьев (деревьев классификации) работает таким образом, что генерирует множество деревьев решений (каждое дерево строится с использованием информации по ранее построенным деревьям) из одного набора данных, рассчитывает вес для каждого дерева и объединяет их в единый прогноз. Бустинг показал различные результаты при построении глобальных почвенных карт и набора карт почвенных свойств в проекте SoilGrids (Hengl et al., 2017a) и карт содержания элементов питания на территории Африки (Hengl et al., 2017b).

В настоящей статье представлен опыт использования некоторых методов машинного обучения для моделирования пространственной вариабельности содержания органического углерода пахотных горизонтов почв тестовых участков Среднерусской возвышенности Воронежской области.

ОБЪЕКТ ИССЛЕДОВАНИЯ

Объектом исследования выступают почвенный покров тестовых участков СХП “Белогорье” ЗАО “Агрофирма Апротек – Подгоренская” Подгоренского района Воронежской области. Отделение СХП “Белогорье” ЗАО “Агрофирма Апротек – Подгоренская” расположено на юго-западе Воронежской области (рис. 1а) в восточной части Подгоренского района (рис. 1б). Центральная усадьба расположена в с. Белогорье.

Территория исследования находится в южной части Среднерусской возвышенности, на правобережье р. Дон, в пределах Калитвянского волнисто-балочного южно-лесостепного района. Почти все землепользование занимает межбалочные водоразделы р. Дон со склонами различной крутизны, и только небольшая восточная его часть расположена в пойме реки. Абсолютные отметки высот водоразделов находятся в пределах 250–330 м.

На большей части участка преобладающие формы мезорельефа – холмы и увалы, между которыми располагается разветвленная

овражно-балочная сеть. Территория землепользования характеризуется сильной расчлененностью, с коэффициентами горизонтальной расчлененности достигающими 1.1–1.15 км/км^2. Степень вертикального расчленения территории (средние относительные превышения водоразделов над днищами долин) составляет 50–80 м (5-я ступень – высшая для равнинного рельефа). Средняя длина склонов 0.35–0.5 км. Коэффициент расчленения территории оврагами 1.7–1.8 км/км^2.

Тестовые участки расположены в западном направлении от с. Белогорье, в центральной части хозяйства. В состав тестовых участков входят три поля площадью 55, 16 и 57 га (рис. 1в).

Территория исследования характеризуется умеренно-континентальным климатом с жарким и сухим летом, прохладной зимой и устойчивым снежным покровом, продолжительным вегетационным периодом, умеренным дефицитом и неустойчивостью атмосферного увлажнения. Среднегодовая сумма осадков составляет 517 мм, но они выпадают неравномерно по годам, изменяясь от 200–250 до 600–700 мм.

Согласно почвенно-географическому районированию, территория исследования входит в Центральную лесостепную и степную область, зону обыкновенных и южных черноземов, Южно-Русскую провинцию (Добровольский, Урусевская, 2015). На большей части территории исследования распространены черноземы обыкновенные, которые представлены большим количеством разностей по гранулометрическому составу – от супесчаного до глинистого.

В пределах тестовых участков установлено формирование следующих родов обыкновенных черноземов: обычные, карбонатные, остаточно-карбонатные и бескарбонатные.

По пониженным элементам рельефа: неглубоким ложбинам и депрессиям водоразделов, выположенным участкам вогнутых склонов формируются лугово-черноземные почвы.

Почвообразующие и подстилающие породы представлены покровными отложениями, элюво-делювием коренных меловых пород, неогеновыми песчаными и неогеновыми глинистыми отложениями. Особо выделяются почвы на двучленных породах, где покровные отложения подстилаются неогеновыми песчаными отложениями (глубина подстилания варьирует от 40 до 80 см), неогеновыми глинистыми отложениями (глубина подстилания от 150 см и более).

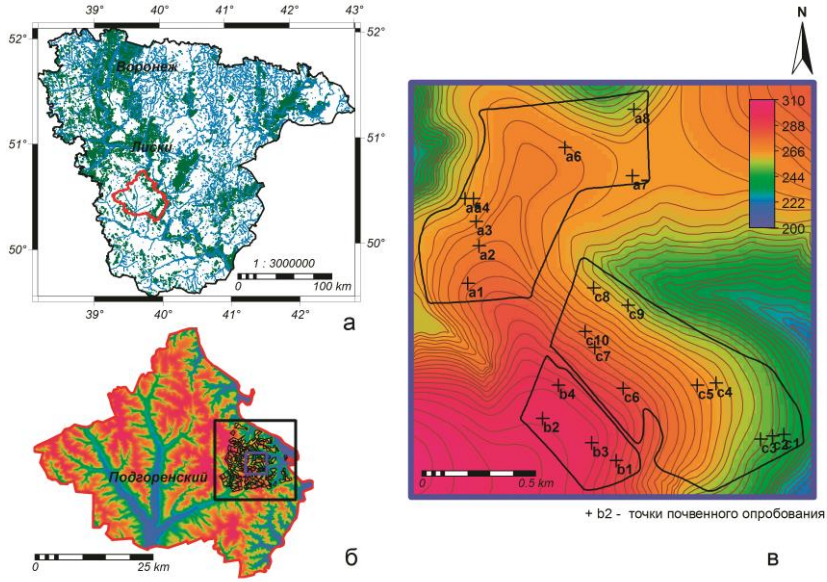

Рис. 1. Расположение объекта исследования.

ОБЪЕКТЫ И МЕТОДЫ

В ходе полевого почвенно-ландшафтного обследования заложены 22 почвенные выработки и отобраны образцы пахотных горизонтов (рис. 1в). Каждая почвенная выработка зарегистрирована прибором GPS, описана в полевом дневнике и сфотографирована. В образцах почв определяли содержание органического углерода почв по методу И.В. Тюрина в модификации В.Н. Симакова. В табл. 1 представлена описательная статистика по исследуемому свойству.

Таблица 1. Описательная статистика по анализируемому свойству

Свойство	Число наблюдений	Минимальное значение	1ый квартиль	Медиана	Среднее	3ий квартиль	Максимальное значение
$C_{орг}$, %	22	0.74	1.91	2.67	2.49	3.20	3.71

Как видно из табл. 1, для почв тестовых участков характерным является большая разница в содержании органического углерода гумуса. Это связано, в первую очередь, со сложным геологическим строением и маломощностью чехла четвертичных отложений, из-за чего территория исследования характеризуется сложным и комплексным почвенным покровом, представленным по большей части контрастными комбинациями.

В качестве предикторов для моделирования пространственной вариабельности содержания органического углерода почв тестовых участков использовали большой набор переменных, в основном представленных данными дистанционного зондирования. Набор включает в себя:

– цифровые модели рельефа и производные от них – крутизну склонов, их экспозицию, топографический индекс влажности, водосборную площадь, различную кривизну поверхности;

– данные дистанционного зондирования, представленные двумя сценами (25 марта 2014 г. и 24 апреля 2014 г.) спутника Landsat 8 (Level-2 Data Products – Surface Reflectance) (Vermote et al., 2016) – каналы видимой и ближней инфракрасной областей спектра и сценой (9 апреля 2016 г.) спутника Sentinel 2 – каналы видимой и ближней инфракрасной областей спектра, прошедшие этапы атмосферной и радиометрической коррекции с помощью модуля “Semi-Automatic Classification Plugin” ГИС QGIS;

– спектральные индексы (отношения каналов), рассчитанные с использованием каналов вышеперечисленных сцен.

Спутники Landsat 8 и Sentinel 2 выбраны по ряду причин: открытый доступ к данным (https://glovis.usgs.gov/), большие архивы данных (пополняются в настоящее время), обширный охват территории.

Так как используемые данные дистанционного зондирования различаются по своему пространственному разрешению (30 м/пиксель для Landsat 8 и 10 м/пиксель для Sentinel 2), построено несколько цифровых моделей рельефа, разрешение которых подбирали соответственно разрешению снимков спутниковых систем. ЦМР построены интерполяцией ординарным кригингом высотных отметок топографической карты М 1 : 10 000.

Использование каналов съемки одной сцены иногда малоприменимо для почвенного картографирования, так как и открытая

поверхность почвы, и растительный покров динамичны во времени и претерпевают сезонные изменения (Савин, Прудникова, 2014; Hengl et al. 2017a). В связи с этим принято решение поканально рассчитать средние значения отражения для двух сцен спутника Landsat 8 и от них определять спектральные индексы.

Таким образом, использовали два набора данных переменных-предикторов. Первый включал производные от ЦМР с разрешением 30 м/пиксель, средние значения отражения 4 каналов (видимая и ближняя инфракрасная области спектра) для двух сцен спутника Landsat 8, спектральные индексы (всего 19 переменных). Второй – производные от ЦМР с разрешением 10 м/пиксель, значения отражения 4-х каналов для сцены спутника Sentinel 2, спектральные индексы (всего 19 переменных).

Для устранения возможной мультиколлинеарности, т.е. наличия коррелирующих между собой предикторов, использовали факторный анализ для расчета главных компонент, которые в свою очередь являются нескоррелированными и стандартизованными переменными. Главные компоненты использовали вместо исходных предикторов в моделировании пространственной вариабельности содержания органического углерода почв. Gobin (2000) доказала, что использование главных компонент вместо исходных, оригинальных предикторов увеличило точность получаемых моделей.

В качестве моделей, описывающих зависимость содержания органического углерода почв от трансформированных переменных-предикторов, использовали методы, основанные на построении деревьев решений: ансамбль деревьев решений (Random Forest) (Breiman, 2001), бустинг регрессионных деревьев (XGBoost) (Chen et al., 2016), байесовские регрессионные деревья (BART) (Chipman et al., 2010). Помимо моделей, полученных для исследуемого свойства по трем рассматриваемым методам, получена модель объединенного прогноза (как взвешенное среднее) для уменьшения эффекта пере- или недообучения отдельных моделей (Sollich, Krogh, 2005; Hengl et al., 2017a).

Для построения моделей, связь между трансформированными переменными-предикторами и содержанием углерода в почвах тестовых участков исследовали путем соотнесения существующих точек наблюдений и значений предикторов в этих точках (создание регрессионной матрицы).

Для оценки точности полученных моделей использовали 5-кратную перекрестную проверку, где каждая модель переподбиралась 5-кратно с использованием 90% значений и предсказанные значения от полученной модели сравнивались с оставшимися 10% значений (Kuhn, 2008). Для модели получены коэффициенты детерминации (R^2_{cv} – доля дисперсии, объясняемая моделью), средняя абсолютная ошибка (MAE_{cv}), корень среднеквадратичной ошибки ($RMSE_{cv}$). Доля дисперсии, объясняемая моделью, оценивалась как:

$$R^2_{cv} = \left[1 - \frac{SSE}{SST}\right], \qquad (2)$$

где *SSE* – сумма квадратов ошибки при перекрестной проверке или остаточная сумма квадратов, *SST* – полная сумма квадратов.

После построения модели, убедившись в ее достоверности, применяли правила классификации полученной модели к пространственному предсказанию содержания органического углерода гумуса почв для каждого элемента (пикселя) сетки (растра).

Для работы по подготовке переменных-предикторов (построение ЦМР, расчет морфометрических характеристик, расчет спектральных индексов), построению моделей и проверке их на устойчивость использовали связку открытого программного обеспечения SAGA GIS (Conrad et al., 2015) и R (Core..., 2016) (пакеты “raster”, “sp”, “rgdal”, “caret”). Этапы моделирования, проверки моделей на устойчивость, функции визуализации расположены в свободном доступе на интернет-странице github-аккаунта[1].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

По результатам факторного анализа для первого набора данных переменных-предикторов, представленных ЦМР, ее производными и данными Landsat 8 были рассчитаны 4 главных компоненты, для второго, включающего в себя ЦМР, ее производные, а также данные Sentinel 2–5 главных компонент, описывающих 95% вариабельности исходных данных. Результативность полученных моделей и перекрестной проверки по первому набору данных показана на рис. 2 и в табл. 2.

[1] https://github.com/chinilin/SOC

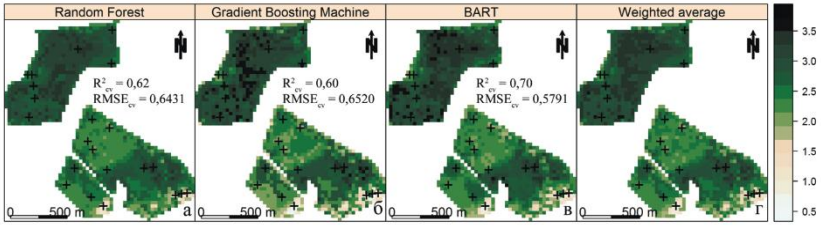

Рис. 2. Результаты пространственного прогнозирования содержания органического углерода почв на основе разных моделей по первому набору предикторов (а – на основе ансамбля деревьев решений, б – на основе бустинга регрессионных деревьев, в – на основе байесовских регрессионных деревьев, г – объединенный прогноз).

Таблица 2. Статистика результатов перекрестной проверки (пятикратная перекрестная проверка)

Модель	Мин. значение	1ый квартиль	Медиана	Среднее	3ий квартиль	Макс. значение	Интерквартильный размах
				MAE_{cv}			
RF	0.1546	0.4045	0.5059	0.5426	0.6769	0.9366	0.2724
XGBoost	0.1288	0.3936	0.5315	0.5474	0.7118	0.973	0.3182
BART	0.129	0.3942	0.4629	0.4920	0.6012	0.8451	0.2070
				$RMSE_{cv}$			
RF	0.2091	0.4701	0.6404	0.6431	0.8001	1.0338	0.3300
XGBoost	0.1554	0.4555	0.6677	0.6520	0.8566	1.1498	0.4011
BART	0.1631	0.4689	0.6039	0.5791	0.6885	0.9500	0.2196
				R^2_{cv}			
RF	0.0113	0.4349	0.6636	0.6209	0.8364	0.9992	0.4015
XGBoost	0.0660	0.4054	0.6456	0.6041	0.8300	0.9996	0.4246
BART	0.0014	0.5600	0.7563	0.7045	0.9021	0.9967	0.3421

Для анализируемого свойства с имеющимися переменными-предикторами получены хорошие результаты – R^2_{cv} изменяется от 0.60 до 0.70, ошибка прогноза, т.е. $RMSE_{cv}$ – от 0.5791 до 0.6520, причем лучшая модель получена с помощью метода байесовских

регрессионных деревьев (BART, рис. 2в), что также видно и по статистике результатов перекрестной проверки (табл. 2).

Наименьшие средние значения средней абсолютной ошибки (MAE_{cv}), корня из среднеквадратичной ошибки ($RMSE_{cv}$), значения интерквартильного размаха и наибольшее значение коэффициента детерминации позволяют выбрать модель, составленную с помощью метода байесовских регрессионных деревьев, как лучшую из представленных. Для лучшей модели, наиболее значимыми предикторами для прогнозирования содержания исследуемого свойства явились (в порядке убывания влияния): 2-я главная компонента, 1-я главная компонента и 3-я главная компонента; 4-я главная компонента не вносила вклад в модель.

Конечный результат прогнозирования, представленный на рис. 2, показывает, что модель, составленная с помощью метода бустинга регрессионных деревьев (рис. 2б), возможно, переобучена, так как содержит достаточно большое количество пикселей с высоким содержанием исследуемого свойства. Напротив, модель, составленная с помощью метода ансамблей деревьев решений (рис. 2а), возможно, недообучена, так как не содержит пикселей с высоким содержанием исследуемого свойства. В то время как модель, составленная с помощью метода байесовских регрессионных деревьев находится где-то между двумя вышеперечисленными. Так как все модели характеризуются близкими значениями ошибки прогноза была получена модель объединенного прогноза (взвешенное среднее) (рис. 2г). Объединенный прогноз, вероятно, является самым удовлетворительным результатом, учитывая, что все три модели имеют схожие и аналогичные результаты.

На рис. 3 и 4 представлены результаты полученных моделей и перекрестной проверки по второму набору данных. Для моделей получены менее точные результаты, хотя и неплохие: R^2_{cv} изменяется от 0.47 до 0.55, ошибка прогноза от 0.7031 до 0.7909. Если оценивать модель по наименьшему значению ошибки прогноза ($RMSE_{cv}$), то лучшей считается модель на основе байесовских регрессионных моделей (наиболее значимыми переменными являются (в порядке убывания) – 2-я главная компонента, 1-я главная компонента, 5-я главная компонента и 3-я главная компонента), если же оценивать по наибольшему значению коэффициента детерминации, то лучшая – модель на основе ансамблей деревьев

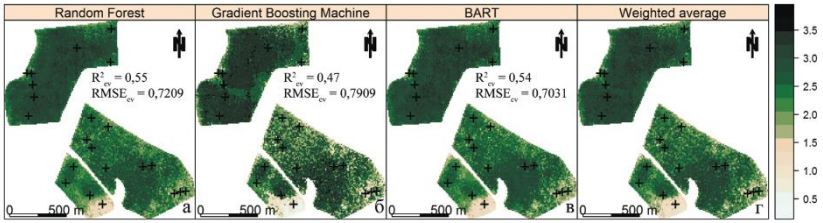

Рис. 3. Результаты пространственного прогнозирования содержания органического углерода почв на основе разных моделей по второму набору предикторов (а – на основе ансамбля деревьев решений, б – на основе бустинга регрессионных деревьев, в – на основе байесовских регрессионных деревьев, г – объединенный прогноз).

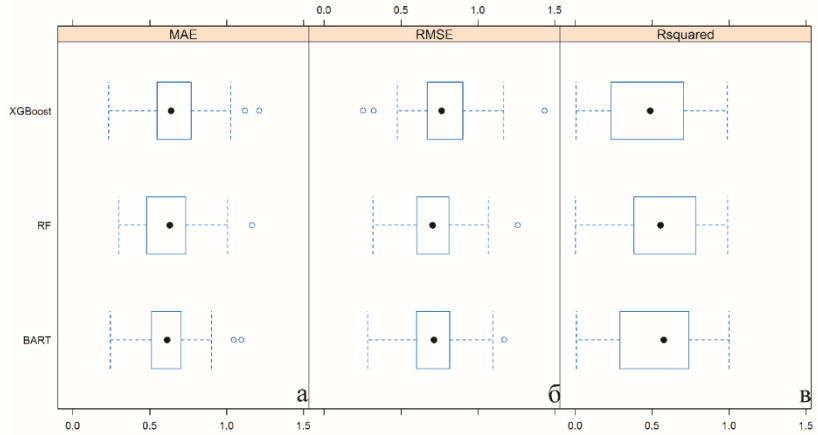

Рис. 4. Статистика результатов перекрестной проверки по второму набору предикторов (а – средняя абсолютная ошибка, б – корень среднеквадратичной ошибки, в – коэффициент детерминации).

решений (наиболее значимыми переменными являются (в порядке убывания) – 1-я главная компонента, 2-я главная компонента, 5-я главная компонента и 4-я главная компонента). Наименьшие значения интерквартильного размаха для значений $RMSE_{cv}$ и R^2_{cv} также характерны для модели на основе ансамблей деревьев решений, что также видно на рис. 4.

Так же, как и в варианте с первым набором данных, полученные модели по второму набору данных характеризуются схожими

результатами, поэтому была получена модель объединенного прогноза (рис. 3г) для уменьшения эффекта пере- или недообучения отдельных моделей.

В целом, пространственное варьирование содержания органического углерода пахотных горизонтов почв тестовых участков по полученным моделям (по двум наборам данных переменных-предикторов) согласуется с нашими представлениями о почвенно-ландшафтных связях. Наиболее низкие значения исследуемого свойства характерны для: а) почв с близким (≈30 см) подстиланием элюво-делювием коренных меловых пород и близким (40–80 см) подстиланием неогеновыми песчаными отложениями; б) нижних частей склонов различных экспозиций с почвами, подверженными процессам эрозии. Высокие значения характерны для почв, формирующихся на плоских поверхностях водоразделов, выположенных шлейфах склонов.

Хочется отметить, что полученные результаты картографирования содержания органического углерода пахотного горизонта почв нельзя считать окончательными. В качестве факторов дифференциации исследуемого свойства рассмотрены только рельеф и данные дистанционного зондирования, хотя и представленные набором информативных показателей. В дальнейшем для уменьшения ошибки и неопределенности прогноза можно и нужно задействовать и другую информацию (карты почвообразующих пород), чтобы в достаточной мере описать каждую из переменных в уравнении модели *scorpan*. Также можно улучшать качество полученных моделей путем отбора новых образцов, например, используя стратегию латинского гиперкуба (Minasny, McBratney, 2006), а также задействовать другие статистические методы и методы машинного обучения.

ВЫВОДЫ

В результате проведенных исследований установлено, что

– методы машинного обучения позволили создать надежные пространственные модели содержания органического углерода пахотных горизонтов почв тестовых участков Среднерусской возвышенности Воронежской области;

– по результатам моделирования выявлено, что для данных первого набора переменных-предикторов (ЦМР и ее производные,

данные Landsat 8) удалось получить более устойчивые модели, причем коэффициент детерминации изменяется от 0.6 до 0.7, тогда как по данным второго набора переменных (ЦМР и ее производные, данные Sentinel 2) R^2_{cv} изменяется от 0.47 до 0.55;

– в различных моделях по различным наборам данных наиболее значимыми оказывались различные переменные-предикторы. Так, по второму набору данных, для модели на основе ансамблей деревьев решений наиболее значимыми оказались следующие (в порядке убывания): 1-я главная компонента, 2-я главная компонента, 5-я главная компонента и 4-я главная компонента (3-я главная компонента не вносила вклад в модель), для модели на основе байесовских регрессионных деревьев наиболее значимыми оказались следующие (также в порядке убывания): 2-я главная компонента, 1-я главная компонента, 5-я главная компонента и 3-я главная компонента (4-я не вносила вклад в модель). При этом следует сказать, что эти модели характеризуются практически аналогичными показателями точности;

– полученные результаты могут служить отправной точкой для дальнейшей работы по созданию крупномасштабных цифровых карт почв/свойств почв региона исследования.

СПИСОК ЛИТЕРАТУРЫ

1. *Добровольский Г.В., Урусевская И.С.* География почв. М.: Изд-во Моск. ун-та, 2015. 458 с.
2. *Жоголев А.В.* Актуализация региональных почвенных карт на основе спутниковых и геоинформационных технологий (на примере Московской области): Автореф. дис. ... к. с.-х. н. М., 2016. 22 с.
3. *Савин И.Ю., Прудникова Е.Ю.* Об оптимальном сроке спутниковой съемки для картографирования пахотных почв // Бюл. Почв. ин-та им. В.В. Докучаева. 2014. № 74. С. 66-77.
4. *Флоринский И.В.* Гипотеза Докучаева — центральная идея цифрового прогнозного почвенного картографирования (к 125-летию публикации). Почвоведение. 2012. № 4. С. 500-506.
5. *Arrouays D., Savin I., Leenaars J., McBratney A.B. (eds.)* GlobalSoilMap - Digital Soil Mapping from Country to Globe. Balkem: CRC Press, 2018. 174 p.
6. *Arrouays D., McKenzie N., Hempel J., Richer de Forges A., McBratney A.* GlobalSoilMap: basis of the global spatial soil information system. Balkem: CRC Press, 2014. 494 p.
7. *Breiman L.* Random Forests // Machine Learning. 2001. № 1 (45). С. 5–32. doi: 10.1023/A:1010933404324

8. *Bui E.N., Henderson B.L., Viergever K.* Knowledge discovery from models of soil properties developed through data mining // Ecological Modelling. 2006. № 3 (191). С. 431–446. doi: 10.1016/j.ecolmodel.2005.05.021
9. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 785–794 с. doi: 10.1145/2939672.2939785
10. *Chipman H.A., George E.I., McCulloch R.E.* BART: Bayesian additive regression trees // The Annals of Applied Statistics. 2010. № 1 (4). С. 266–298. doi: 10.1214/09-AOAS285
11. *Conrad O., Bechtel M., Bock M., Dietrich H., Fischer E.* System for Automated Geoscientific Analyses (SAGA) v. 2.1.4 // Geoscientific Model Development. 2015. № 7 (8). С. 1991–2007. doi: 10.5194/gmd-8-1991-2015
12. *Gobin A.* Participatory and spatial-modeling methods for land resources analysis. PhD thesis. Katholik Universiteit, Leuven, 2000. 282 с.
13. *Grinand C., Arrouays D., Laroche D., Martin M.P.* Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context // Geoderma. 2008. № 1 (143). С. 180–190. doi: 10.1016/j.geoderma.2007.11.004
14. *Hengl T., Heuvelink G.B.M., Kempen B., Leenaars J.G.B., Walsh M.* Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions // PLOS ONE. 2015. № 6 (10). С. e0125814. doi: 10.1371/journal.pone.0125814
15. *Hengl T., Mendes de Jesus J., Heuvelink G.B.M., Ruiperez Gonzalez M., Kilibarda M.* SoilGrids250m: Global gridded soil information based on machine learning // PLOS ONE. 2017. № 2 (12). С. e0169748. doi: 10.1371/journal.pone.0169748
16. *Hengl T., Leenaars K., Shepherd K.D., Walsh M., Heuvelink G.B.M.* Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning // Nutrient Cycling in Agroecosystems. 2017. № 1 (109). С. 77–102. doi: 10.1007/s10705-017-9870-x
17. *Jenny H.* Factors of Soil Formation // Soil Science. 1941. № 5 (52). С. 415. doi: 10.1097/00010694-194111000-00009
18. *Kuhn M.* Building Predictive Models in R Using the caret Package // Journal of Statistical Software. 2008. № 5 (28). doi: 10.18637/jss.v028.i05
19. *Lagacherie P., Holmes S.* Addressing geographical data errors in a classification tree for soil unit prediction // International Journal of Geographical Information Science. 1997. № 2 (11). С. 183–198. doi: 10.1080/136588197242455
20. *McBratney A., Mendonça Santos M., Minasny B.* On digital soil mapping // Geoderma. 2003. № 1–2 (117). С. 3–52. doi: 10.1016/S0016-7061(03)00223-4
21. *Minasny B., McBratney A.B.* A conditioned Latin hypercube method for sampling in the presence of ancillary information // Computers & Geosciences. 2006. № 9 (32). С. 1378–1388. doi: 10.1016/j.cageo.2005.12.009

22. *Core R., Team R.* A language and environment for statistical computing // 2016.
23. *Sollich P., Krogh A.* Learning with ensembles: How overfitting can be useful, Proceedings of the 1995 Conference, Vol. 8, 1996. 190–196 c.
24. *Taghizadeh-Mehrjardi R., Minasny B., McBratney A.B., Triantafilis J.* Digital soil mapping of soil classes using decision trees in central Iran // Proceedings of the 5th Global Workshop on Digital Soil Mapping, 2012. C. 197–202. doi: 10.1201/b12728-40
25. *Vermote E., Justice C., Claverie M., Franch B.* Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product // Remote Sensing of Environment. 2016. № 185. C. 46–56. doi: 10.1016/j.rse.2016.04.008

THE LARGE SCALE DIGITAL MAPPING OF SOIL ORGANIC CARBON USING MACHINE LEARNING ALGORITHMS

© 2018 A. V. Chinilin[1], I. Yu. Savin[2]

[1]RSAU-MTAA, 127550, Russian Federation, Moscow, Timiryazevskaya st., 49
e-mail: andreychinilin@gmail.com

[2]V.V. Dokuchaev Soil Science Institute, Russia, 119017, Moscow, Pyzhevskii per. 7-2
e-mail: savin_iyu@esoil.ru

The results of digital mapping of organic carbon content within the arable horizons of soils and the assessment of obtained models accuracy with the use of machine learning methods for the area of Central Russian Upland in Voronezh Oblast are presented. The digital mapping was based on 22 points of soil samplings, applied for the learning and verification of models, and also on several sets of predictor variables. We took also digital elevation model, its derivatives and also remote sensing data of different spatial resolution as predictor variables. Several methods were used to create the spatial variability models for the investigated property based on the decision trees methods: random forest, boosting regression trees and Bayessian regression trees. The assessment of the models obtained accuracy was conducted by a method of cross-validation. As the accuracy indices we used the determination coefficient, mean absolute error and the root mean square error. The modelling results showed that the use of predictor variables presented by digital elevation model, its derivatives and Landsat 8 data we were able to obtain more sustainable models. The determination coefficient varied from 0.6 to 0.7, RMSEcv, i.e., the prognosing error varied from 0.5791 to 0.6520. Whereas, the best model was obtained with the method of Bayessian regression trees; whereas the predictor variables presented

by the digital elevation model, its derivatives and Sentinel 2 data determination coefficient varied from 0.47 to 0.55, and the prognosing error varied from 0.7031 to 0.7909. It was revealed that in the described models according to different data sets the most significant were the various predictor variables.

Key words: spatial prediction, digital elevation model, random forest, boosting

REFERENCES

1. Dobrovol'skii G.V., Urusevskaya I.S. *Soil geography*, Moscow, MGU Publ., 2015, 458 p. (in Russian)
2. Zhogolev A.V. *Regional soil maps actualization based on geographical information systems and remote sensing data (the case of Moscow region*), Extended abstract of candidate's thesis, 2016, 22 p. (in Russian)
3. Savin I.Yu., Prudnikova E.Yu. About optimal dates of satellite images acquisition for arable soil mapping, *Dokuchaev Soil Bulletin.*2014, V. 74, pp. e52-e61.
4. Florinsky I.V. The Dokuchaev Hypothesis as a Basis for Predictive Digital Soil Mapping (On the 125th Anniversary of Its Publication), *Eurasian Soil Science*, 2012, V. 45 (4), pp. 445-451. doi: 10.1134/S1064229312040047
5. Arrouays D., Savin I., Leenaars J., McBratney A.B. (eds.) *GlobalSoilMap - Digital Soil Mapping from Country to Globe*, Balkem, CRC Press, 2018, 174p.
6. Arrouays D., McKenzie N., Hempel J., Richer de Forges A., McBratney A. *GlobalSoilMap: basis of the global spatial soil information system*, Balkem, CRC Press, 2014, 494 p.
7. Breiman L. Random Forests, *Machine Learning*, 2001, No. 1 (45), pp. 5–32. doi: 10.1023/A:1010933404324
8. Bui E.N., Henderson B.L., Viergever K. Knowledge discovery from models of soil properties developed through data mining, *Ecological Modelling*, 2006, No. 3 (191), pp. 431–446. doi: 10.1016/j.ecolmodel.2005.05.021
9. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp.785–794. doi: 10.1145/2939672.2939785
10. Chipman H.A., George E.I., McCulloch R.E. BART: Bayesian additive regression trees, *The Annals of Applied Statistics*, 2010, No. 1 (4), pp. 266–298. doi: 10.1214/09-AOAS285
11. Conrad O., Bechtel M., Bock M., Dietrich H., Fischer E. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, *Geoscientific Model Development*, 2015, No. 7 (8), pp. 1991–2007. doi: 10.5194/gmd-8-1991-2015
12. Gobin A. *Participatory and spatial-modeling methods for land resources analysis*, PhD thesis, Katholik Universiteit, Leuven, 2000, 282 p.
13. Grinand C., Arrouays D., Laroche D., Martin M.P. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context, *Geoderma*, 2008, No. 1 (143), pp. 180–190. doi: 10.1016/j.geoderma.2007.11.004

14. Hengl T., Heuvelink G.B.M., Kempen B., Leenaars J.G.B., Walsh M. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions, *PLOS ONE*, 2015, No. 6 (10), pp. e0125814. doi: 10.1371/journal.pone.0125814
15. Hengl T., Mendes de Jesus J., Heuvelink G.B.M., Ruiperez Gonzalez M., Kilibarda M. SoilGrids250m: Global gridded soil information based on machine learning, *PLOS ONE*, 2017, No. 2 (12), pp. e0169748. doi: 10.1371/journal.pone.0169748
16. Hengl T., Leenaars K., Shepherd K.D., Walsh M., Heuvelink G.B.M. Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning, *Nutrient Cycling in Agroecosystems*, 2017, No. 1 (109), pp. 77–102. doi: 10.1007/s10705-017-9870-x
17. Jenny H. Factors of Soil Formation, *Soil Science*, 1941, No. 5 (52), pp. 415. doi: 10.1097/00010694-194111000-00009
18. Kuhn M. Building Predictive Models in R Using the caret Package, *Journal of Statistical Software*, 2008, No. 5 (28), pp. 1-26. doi: 10.18637/jss.v028.i05
19. Lagacherie P., Holmes S. Addressing geographical data errors in a classification tree for soil unit prediction, *International Journal of Geographical Information Science*, 1997, No. 2 (11), pp. 183–198. doi: 10.1080/136588197242455
20. McBratney A., Mendonça Santos M., Minasny B. On digital soil mapping, *Geoderma*, 2003, No. 1–2 (117), pp. 3–52. doi: 10.1016/S0016-7061(03)00223-4
21. Minasny B., McBratney A.B. A conditioned Latin hypercube method for sampling in the presence of ancillary information, *Computers & Geosciences*, 2006, No. 9 (32), pp. 1378–1388. doi: 10.1016/j.cageo.2005.12.009
22. R Core Team R: *A language and environment for statistical computing*, 2016.
23. Sollich P., Krogh A. Learning with ensembles: How overfitting can be useful, Proceedings of the 1995 Conference, V. 8, 1996, pp. 190–196.
24. Taghizadeh-Mehrjardi R., Minasny B., McBratney A.B., Triantafilis J. Digital soil mapping of soil classes using decision trees in central Iran, *Proceedings of the 5th Global Workshop on Digital Soil Mapping*, 2012, pp. 197–202. doi: 10.1201/b12728-40
25. Vermote E., Justice C., Claverie M., Franch B. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product, *Remote Sensing of Environment*, 2016, No. 185, pp. 46–56. doi: 10.1016/j.rse.2016.04.008

Ссылки для цитирования

Чинилин А.В., Савин И.Ю. Крупномасштабное цифровое картографирование содержания органического углерода почв с помощью методов машинного обучения // Бюл. Почв. ин-та им. В.В. Докучаева. 2018. Вып. 91. С. 46-62. doi: 10.19047/0136-1694-2018-91-46-62
Chinilin A.V., Savin I. Yu. The large scale digital mapping of soil organic carbon using machine learning algorithms, *Dokuchaev Soil Bulletin*, 2018, Vol. 91, pp. 46-62. doi: 10.19047/0136-1694-2018-91-46-62