Сравнительная оценка эффективности методов отбора предикторов при картографировании кислотности почв алгоритмом машинного обучения Random Forest

к.б.н. Гопп Н.В., natalia.gopp@gmail.com
ФГБУН Институт почвоведения и агрохимии СО РАН
Новосибирск

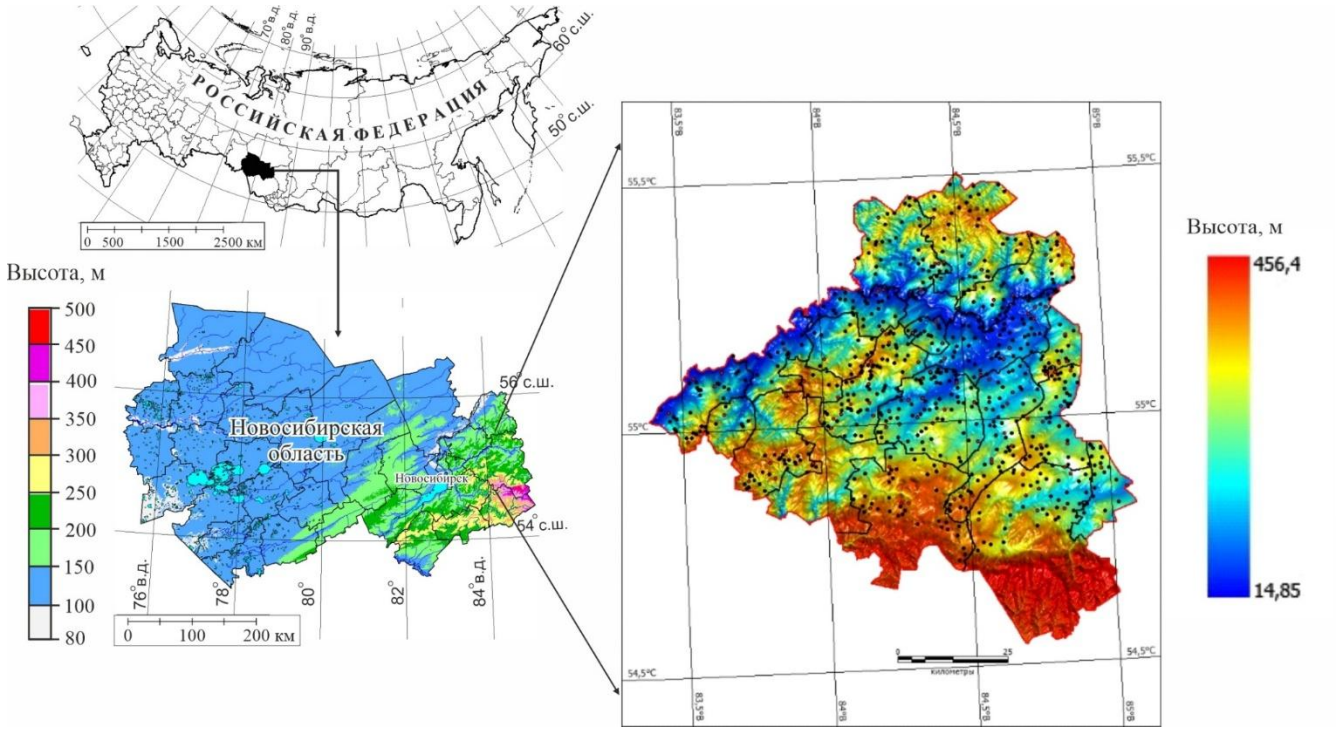
Актуальность

Сведения о кислотности в картографическом представлении необходимы для оценки пригодности почв для возделывании сельскохозяйственных культур, а также для планирования мероприятий по проведению химической мелиорации. Значительное пространственное варьирование pH обусловливает необходимость её картографирования с использованием современных методов, реализация которых успешно осуществляется с помощью базы данных лабораторно-полевых обследований почв, данных дистанционного зондирования Земли, и алгоритмов машинного обучения.

Цель исследования

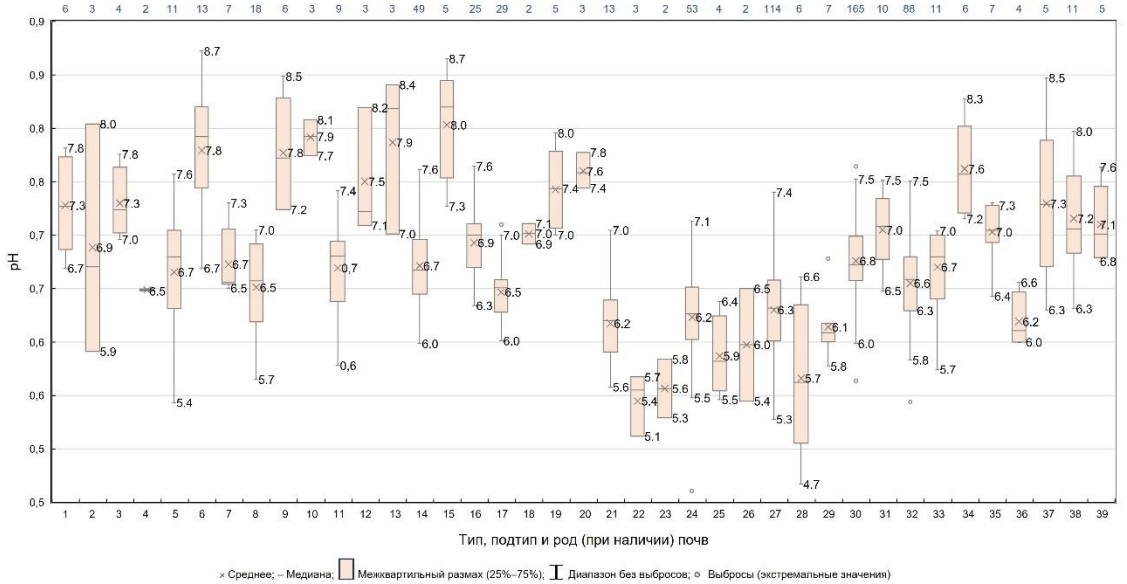
Цель исследования – сравнительная оценка эффективности методов отбора предикторов (Boruta, RFE) при картографировании кислотности почв алгоритмом машинного обучения Random Forest, реализованном на онлайн-платформе Google Earth Engine.

Территория исследования: Предсалаирье (в пределах Тогучинского района, Новосибирской области), площадь 3925 км².



Район исследования по геоморфологическому районированию относится к Кузнецко-Салаирской геоморфологической провинции Алтае-Саянской горной области.

Данные по pH водной суспензии почв собраны по материалам ЗапСибНИИгипрозем (1984-1994 гг.)



Варьирование pH водной суспензии в 0–30 см слое почв. Типы, подтипы и роды (при наличии) почв: 1 – аллювиальные дерновые обычные; 2 – аллювиальные луговые; 3 – аллювиальные лугово-болотные; 4 – дерново-подзолистые; 5 – луговые выщелоченные; 6 – луговые карбонатные; 7 – луговые обычные; 8 – луговые оподзоленные; 9 – луговые солончаковатые; 10 – луговые солончаковые; 11 – лугово-болотные перегнойные; 12 – лугово-болотные перегнойные карбонатные; 13 – лугово-болотные перегнойные солончаковатые; 14 – лугово-черноземные выщелоченные; 15 – лугово-черноземные карбонатные; 16 – лугово-черноземные обычные; 17 – лугово-черноземные оподзоленные; 18 – лугово-черноземные солонцеватые; 19 – лугово-черноземные солончаковатые; 20 – лугово-черноземная солончаковая; 21 – светло-серые лесные; 22 – светло-серые лесные глеевые осолоделые; 23 – светло-серые лесные остаточно-карбонатные; 24 – серые лесные; 25 – серые лесные глеевые; 26 – серые лесные скелетные; 27 – темно-серые лесные; 28 – темно-серые лесные глеевые; 29 – темно-серые лесные остаточно-карбонатные; 30 – черноземы выщелоченные; 31 – черноземы обыкновенные; 32 – черноземы оподзоленные; 33 – черноземно-луговые выщелоченные; 34 – черноземно-луговые карбонатные; 35 – черноземно-луговые обычные; 36 – черноземно-луговые оподзоленные; 37 – черноземно-луговые солончаковатые; 38 – солонцы лугово-черноземные глубокие; 39 – солонцы черноземно-луговые глубокие. Объем выборки указан цифрами сверху над диаграммами размаха

Описательная статистика для pH водной суспензии в 0–30 см слое почв для общего, обучающего и валидационного наборов данных

Показатель	Набор данных		
	Общий (*n*=722)	Обучающий (*n*=612)	Валидационный (*n*=110)
Среднее	6.6	6.7	6.6
Стандартное отклонение	0.6	0.6	0.5
Минимум	4.6	4.7	4.6
Медиана	6.6	6.6	6.6
Максимум	8.8	8.7	7.9
Коэффициент вариации, %	8.5	8.6	7.8
Коэффициент асимметрии	0.33	0.41	-0.38
Коэффициент эксцесса	1.25	1.18	1.68

Feature Selection with the Boruta Package

Miron B. Kursa, Witold R. Rudnicki

Abstract

This article describes a R package **Boruta**, implementing a novel feature selection algorithm for finding emph{all relevant variables}. The algorithm is designed as a wrapper around a Random Forest classification algorithm. It iteratively removes the features which are proved by a statistical test to be less relevant than random probes. The **Boruta** package provides a convenient interface to the algorithm. The short description of the algorithm and examples of its application are presented.

Files: Paper | R package (Boruta) | R example code from the paper

Published: Sep 16, 2010

DOI: 10.18637/jss.v036.i11

Home > Machine Learning > Article

Gene Selection for Cancer Classification using Support Vector Machines

Published: January 2002

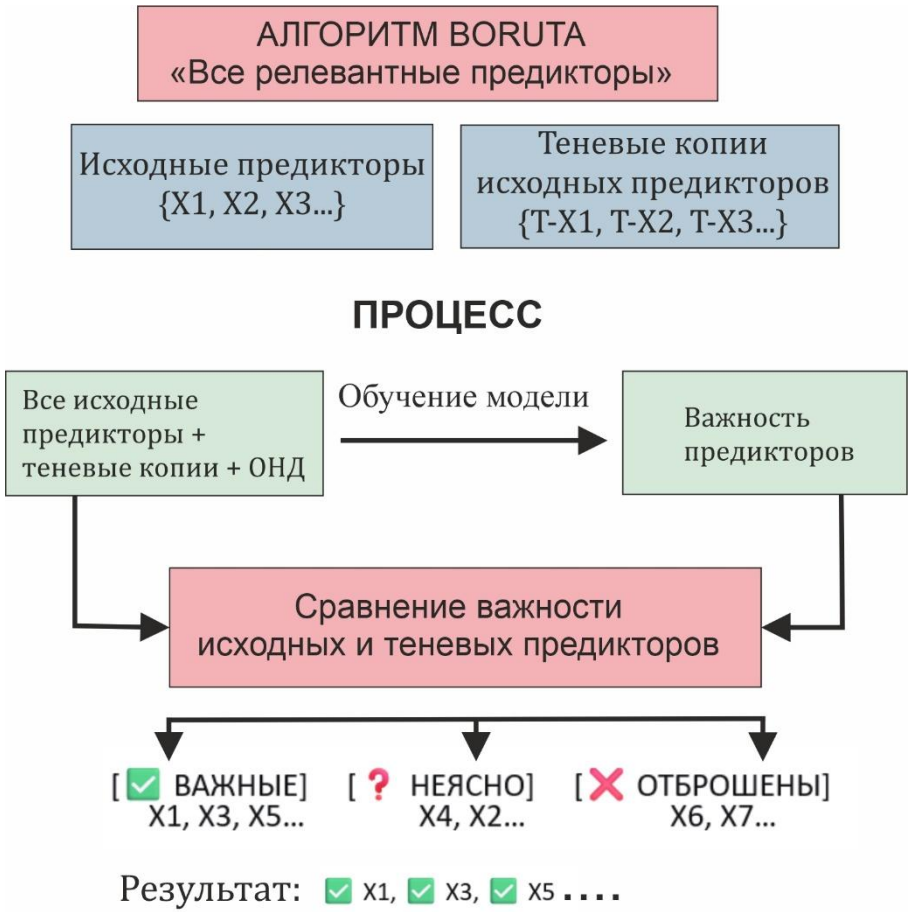
Volume 46, pages 389–422, (2002) Cite this article

Download PDF

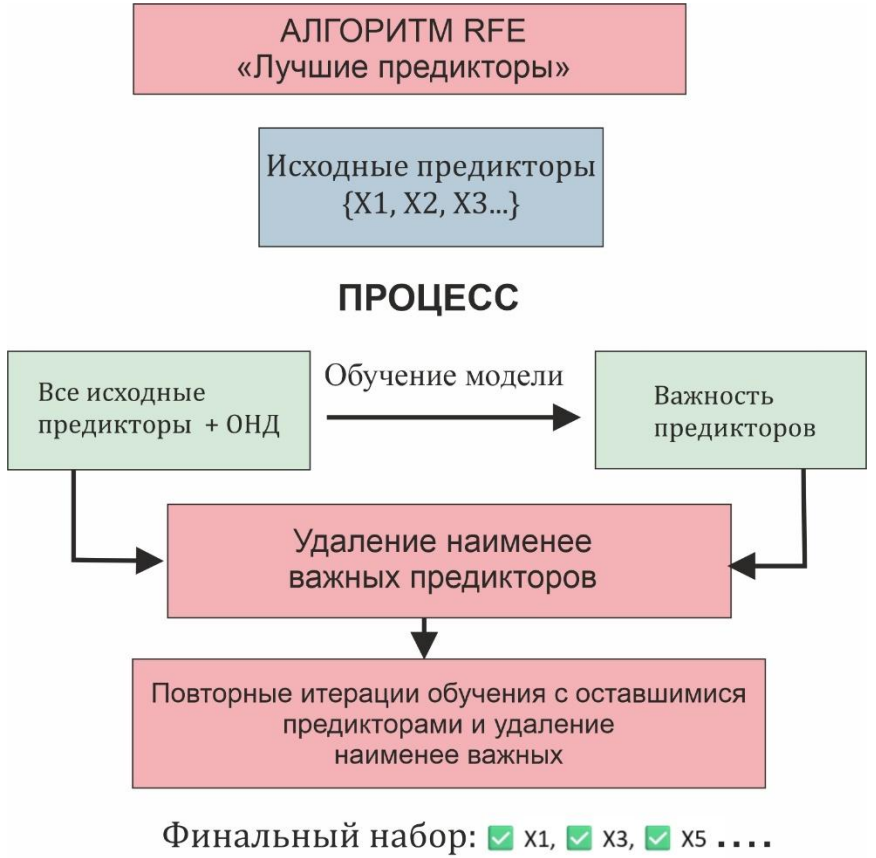
Isabelle Guyon, Jason Weston, Stephen Barnhill & Vladimir Vapnik

60k Accesses 8216 Citations 69 Altmetric 4 Mentions Explore all metrics

СХЕМЫ АЛГОРИТМОВ ОТБОРА ПРЕДИКТОРОВ



RFE (Recursive Feature Elimination)



Авторы (2010 г.): Miron B. Kursa, Witold R. Rudnicki

ОНД – обучающий набор данных

Автор рисунков Гопп Н.В.

Авторы (2002 г.): Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik

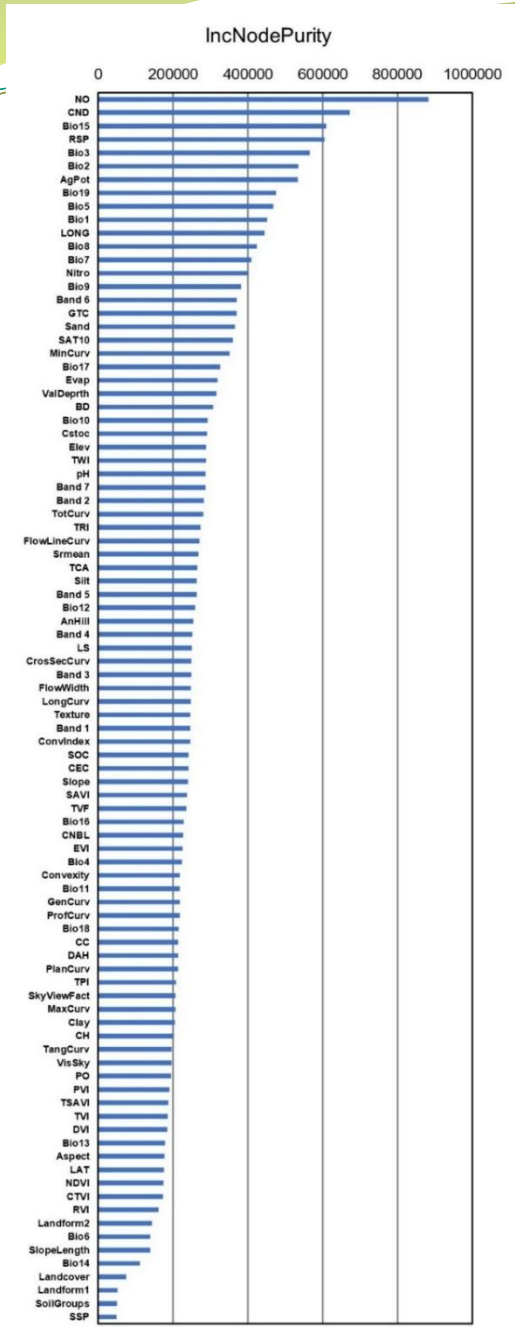
Модель SCORPAN

$$Sc = f(s, c, o, r, p, a, n) \quad \text{и} \quad Sa = f(s, c, o, r, p, a, n),$$

где ***Sc*** – почвенные таксономические единицы; ***Sa*** – количественная характеристика почвы; s – почва (другие характеристики почвы); **c** – климат (климатические характеристики); **o** – организмы, растительность, фауна, человек; **r** – рельеф (ЦМР и морфометрические величины); **p** – материнская порода, литология; **a** – возраст, время, повторность при мониторинге; **n** – пространственное положение.

Модель SCORPAN предназначена для эмпирического количественного описания взаимосвязей между почвенными свойствами и пространственно распределенными предикторами (McBratney et al., 2003).

Ранжирование по степени важности

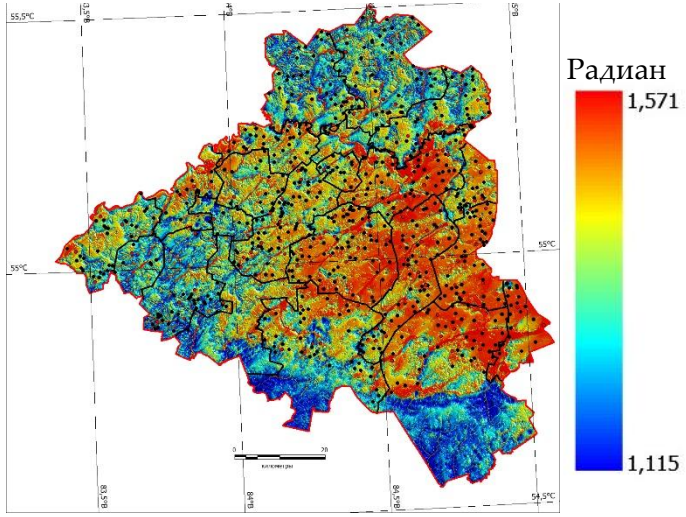


Перечень предикторов, использованных в картографировании кислотности почв

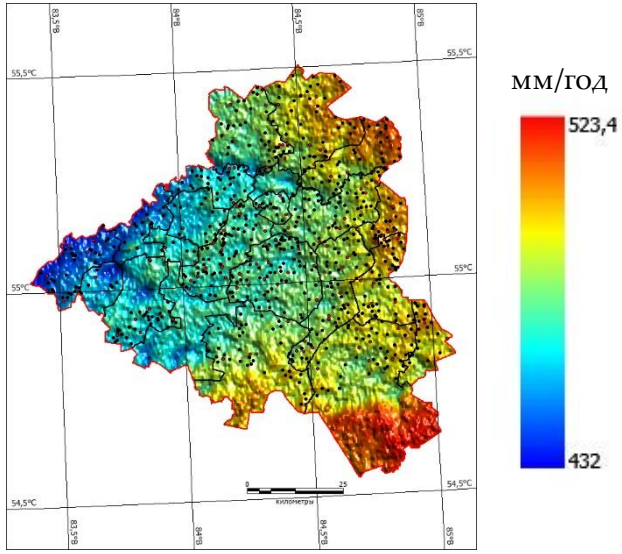
Предикторы, характеризующие климат (WorldClim и архивные карты) [18, 19]. Исходное разрешение 1 × 1 км, преобразованное – 30 × 30 м	BIO1 – среднегодовая температура; BIO2 – среднесуточная разность (среднее значение за месяц (Tmax – Tmin)); BIO3 – изотермичность; BIO4 – сезонность температуры; BIO5 – максимальная температура самого теплого месяца; BIO6 – минимальная температура самого холодного месяца; BIO7 – годовой диапазон температур; BIO8 – средняя температура самого влажного квартала; BIO9 – средняя температура самого сухого квартала; BIO10 – средняя температура самого теплого квартала; BIO11 – средняя температура самого холодного квартала; BIO12 – годовое количество осадков; BIO13 – количество осадков в самый влажный месяц; BIO14 – количество осадков в самый засушливый месяц; BIO15 – сезонность осадков (коэффициент вариации); BIO16 – количество осадков в самом влажном квартале; BIO17 – количество осадков в самом сухом квартале; BIO18 – количество осадков в самом теплом квартале; BIO19 – количество осадков в самом холодном квартале; SRmean – среднее солнечное излучение (сумма средних значений (с 1970 по 2000 г.г.) по месяцам деленная на 12); SAT10 – сумма активных температур выше 10°C; CC – коэффициент континентальности климата; EVAP – испаряемость; CH – коэффициент увлажнения; AgPot – агроэкологический потенциал; HTC – гидротермический коэффициент Селянинова
Предикторы, характеризующие рельеф (на основе FABDEM V1-2, разрешение 30 × 30 м)	ELEV – высота над уровнем моря; Slope – крутизна склонов; LS – коэффициент соотношения длины и крутизны склона; Aspect – экспозиция склонов; CrosSecCurv – кривизна поперечного сечения; FlowLineCurv – кривизна линии потока; GenCurv – главная кривизна; LongCurv – продольная кривизна; MinCurv – минимальная кривизна; MaxCurv – максимальная кривизна; ProfCurv – профильная кривизна; PlanCurv – плановая кривизна; TanCurv – тангенциальная кривизна; TotCurv – общая кривизна; ConvIndex – индекс конвергенции; Texture – рельефная текстура поверхности; Convexity – индекс выпуклости; AnalitHill – аналитическая затененность холмов; VallDepth – глубина долин; TWI – топографический индекс влажности; TCA – общая площадь водосбора; RSP – относительное положение на склоне; TPI – топографический индекс положения; TRI – индекс расчлененности рельефа; CND – расстояние до водотоков (дренажной сети); CNBL – базовый уровень водотоков; VisSky – видимость неба; SkyViewFact – фактор видимости неба; PO – положительная открытость ландшафта; NO – отрицательная открытость (замкнутость ландшафта); SSP – специфичные точки поверхности; SlopeLength – длина склона; FlowWidth – ширина потока; TVF – коэффициент обзора местности; DAH – суточный анизотропный нагрев; Landform 1 – 6 форм рельефа; Landform 2 – 16 форм рельефа
Предикторы, характеризующие пространственное положение (разрешение 30 × 30 м)	LONG – географическая долгота; LAT – географическая широта
Предикторы, характеризующие растительность (Landsat 5 TM, разрешение 30 × 30 м)	NDVI – нормализованный разностный вегетационный индекс; CTVI – скорректированный трансформированный вегетационный индекс; DVI – разностный вегетационный индекс; NRVI – нормализованный относительный вегетационный индекс; RVI – относительный вегетационный индекс; SAVI – вегетационный индекс с коррекцией по почве; TSAVI – трансформированный вегетационный индекс с коррекцией по почве; TVI – трансформированный вегетационный индекс; TTVI – трансформированный вегетационный индекс Тиама; Band 1 – видимый синий диапазон; Band 2 – видимый зеленый диапазон; Band 3 – видимый красный диапазон; Band 4 – ближний инфракрасный диапазон; Band 5 – ближний инфракрасный диапазон; Band 6 – тепловой диапазон; Band 7 – средний инфракрасный диапазон; LandCov – наземный покров
Предикторы, характеризующие почву (SoilGrids, усредненные данные для слоя 0–30 см) [20]. Исходное разрешение 250 × 250 м, преобразованное – 30 × 30 м	pH – кислотность; NITRO – содержание общего азота; SOC – содержание органического углерода; SOSC – запасы органического углерода; CEC – емкость катионного обмена; CLAY – содержание ила (глины); SAND – содержание песка; SILT – содержание пыли; BD – плотность сложения почв; Soil Groups – почвенные группы по WRB 2006

РАСТРОВЫЕ КАРТЫ ПРЕДИКТОРОВ

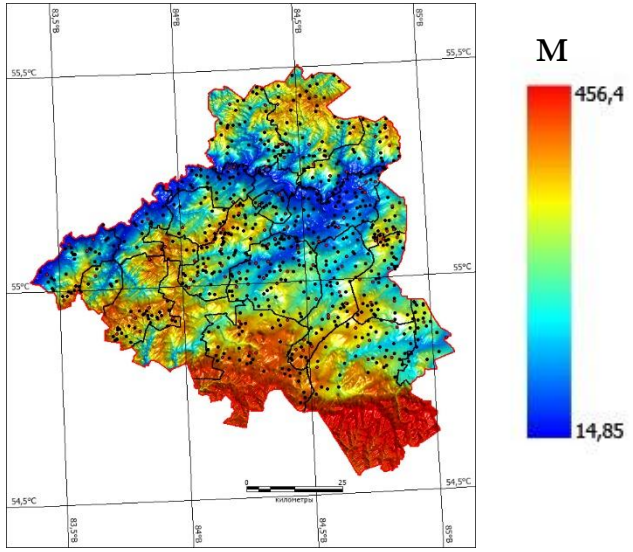
NO – отрицательная открытость (замкнутость ландшафта)



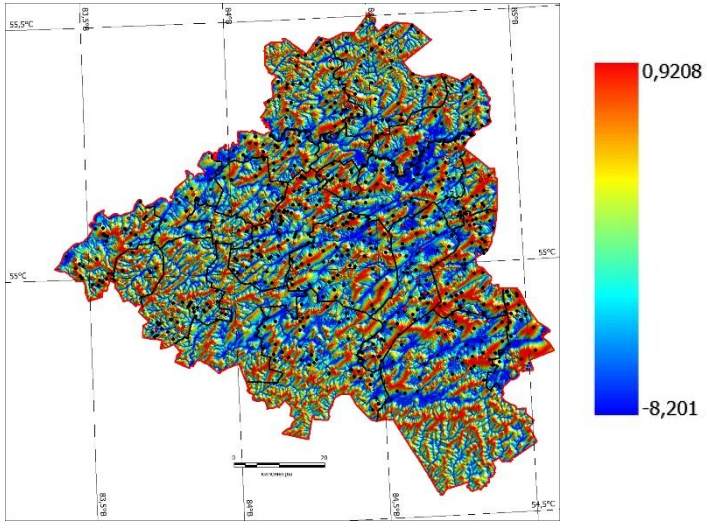
BIO12 – годовое количество осадков (WorldClim 2.0)



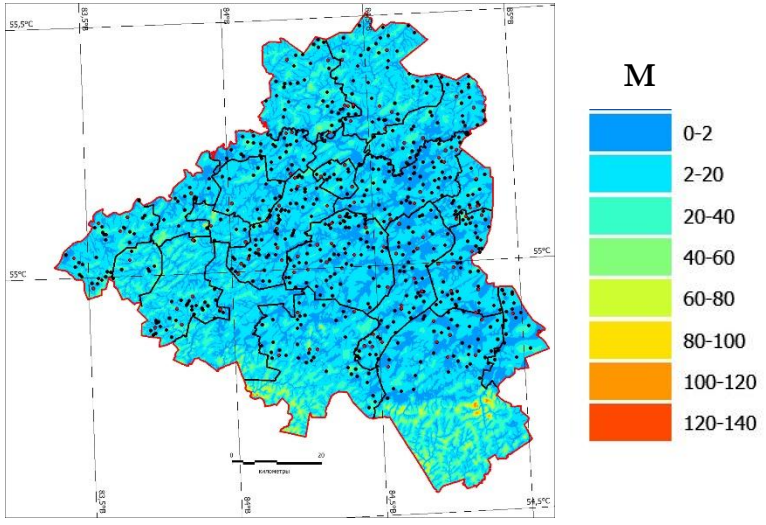
ELEV – высота над уровнем моря (FABDEM V1-2)



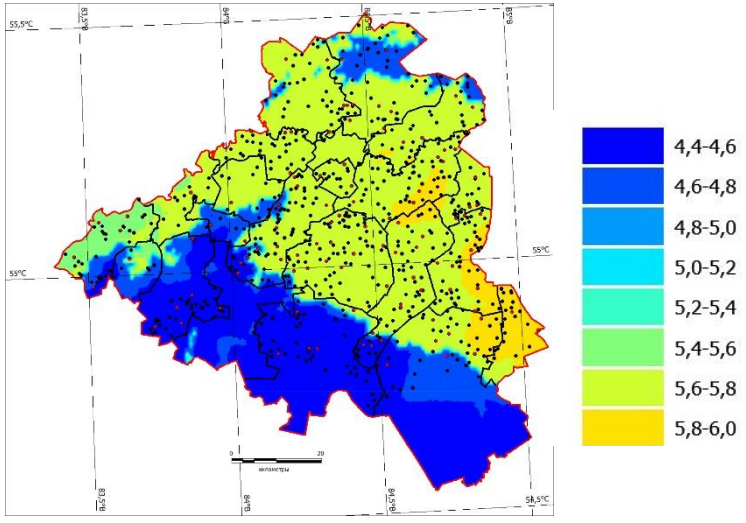
RSP – относительное положение на склоне



CND – расстояние до водотоков (дренажной сети)

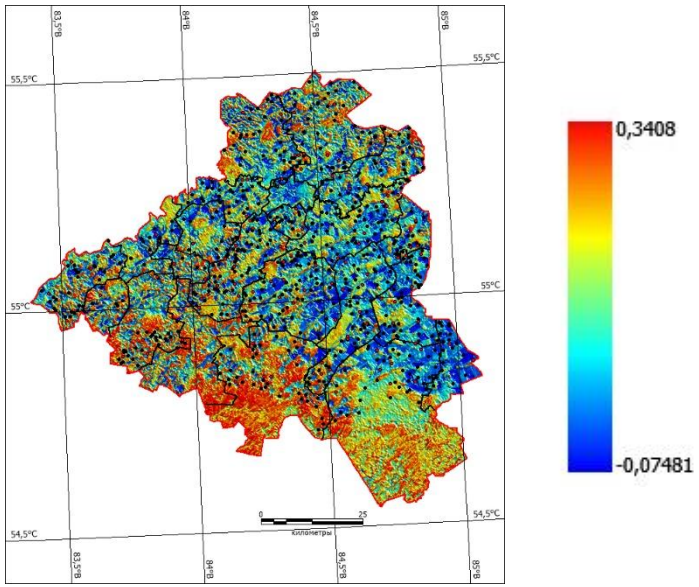


AgPot – агроэкологический потенциал

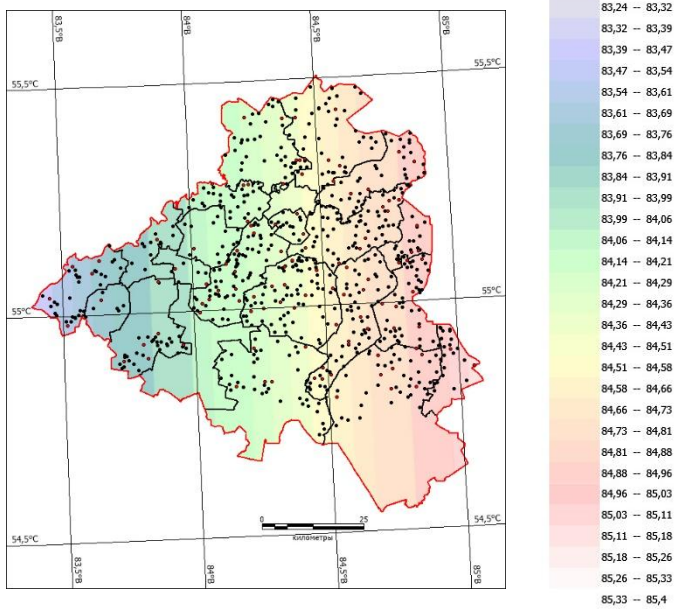


РАСТРОВЫЕ КАРТЫ ПРЕДИКТОРОВ

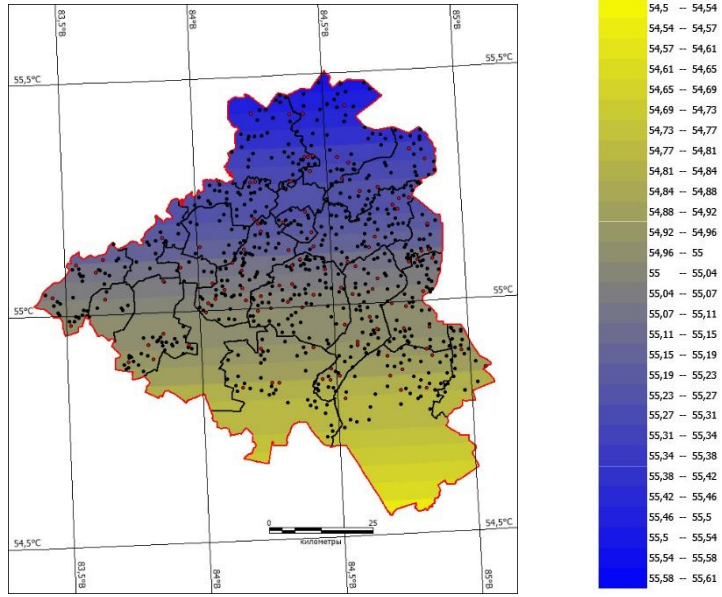
NDVI – нормализованный разностный вегетационный индекс (Landsat-5)



LONG – географическая долгота



LAT – географическая широта



Random Forest (Случайный лес)

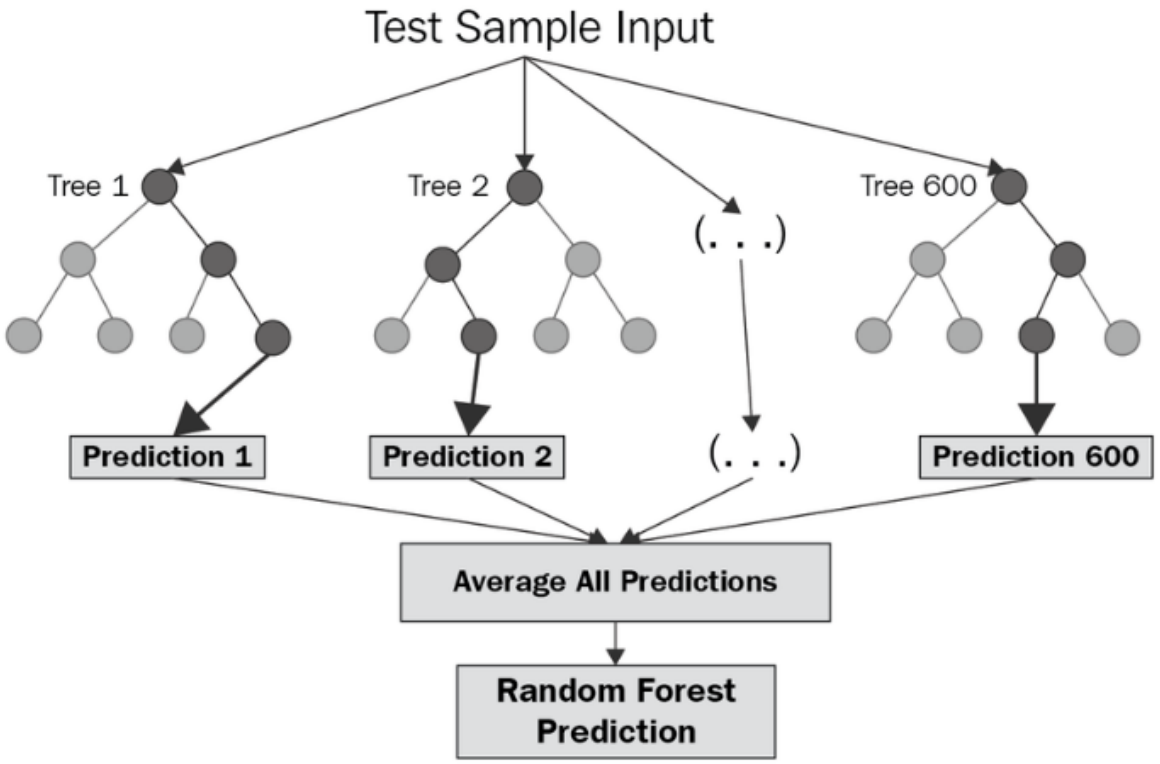
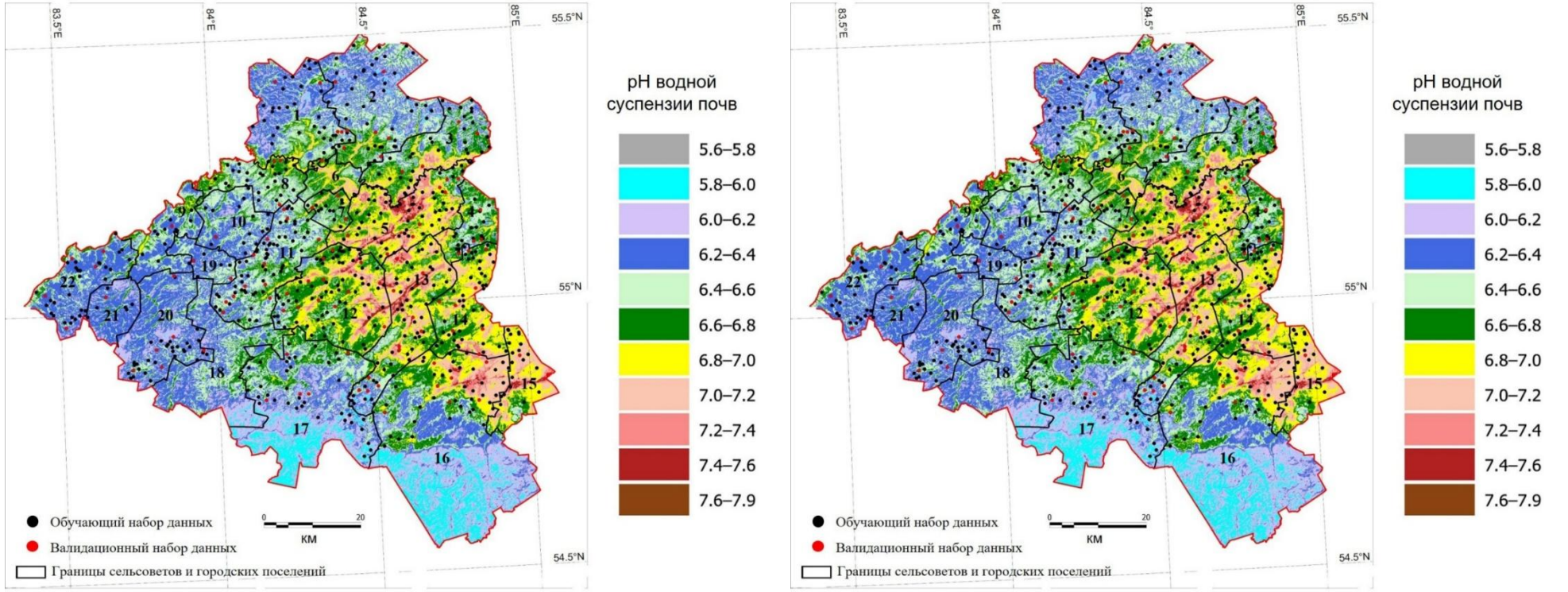


Рисунок из работы: Jain A., Fandango A., Kapoor A.: TensorFlow Machine Learning Projects: Build 13 realworld projects with advanced numerical computations using the Python ecosystem. 2018

Термин «дерево» в концепции алгоритма Random Forest означает независимый алгоритм, обучающийся на случайной подвыборке данных и подмножестве признаков. Алгоритм Random Forest позволяет создавать множество «деревьев» и дальнейшее объединение результатов работы разных независимых «деревьев» или алгоритмов даёт более точные и устойчивые предсказания для зависимой/целевой переменной. Конечный результат – это среднее значение предсказаний по всем созданным «деревьям».

Brieman L. Random Forests // Mach. Learn. 2001. V. 45. P. 5–32. https://doi.org/10.1023/A:1010933404324

BORUTA **RFE**



pH водной суспензии:

сильнокислая (5.0–6.0)

слабокислая (6.0–6.5)

нейтральная (6.5–7.5)

слабощелочная (7.5–8.5)

Карта pH водной суспензии в 0–30 см слое почв (по данным 1984-1994 г.г.)

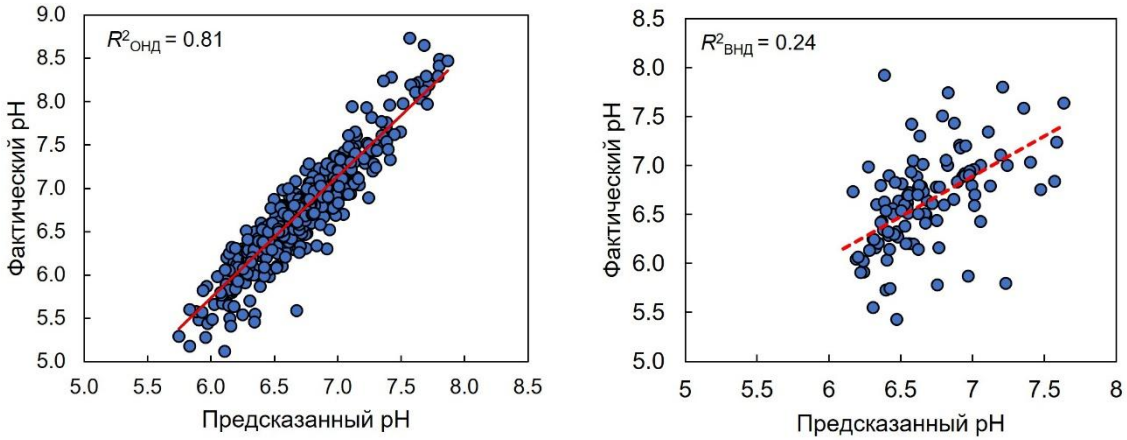
Названия сельсоветов и городских поселений Тогучинского района Новосибирской области:

1 – Гутовский; 2 – Сурковский; 3 – Киикский; 4 – Шахтинский; 5 – Заречный; 6 – городское поселение Тогучин; 7 – Нечаевский; 8 – Кудринский; 9 – Буготакский; 10 – Борцовский; 11 – Кудельно-Ключевской; 12 – Вассинский; 13 – Завьяловский; 14 – Кировский; 15 – Степногутовский; 16 – Коуракский; 17 – Лебедевский; 18 – Чемской; 19 – городское поселение Горный; 20 – Усть-Каменский; 21 – Мирновский; 22 – Репьёвский.

Оценка эффективности моделирования

Зависимость между фактическим и предсказанным pH в обучающем ($R^2_{ОНД}$) и валидационном ($R^2_{ВНД}$) наборах данных

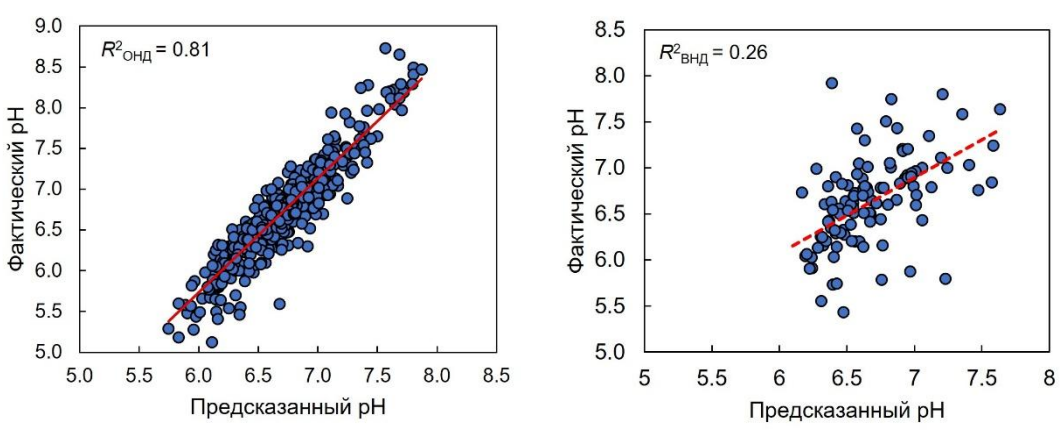
BORUTA



$RMSE_{ВНД}$=0,46; $MAPE_{ВНД}$=5,0; $MAE_{ВНД}$=0,30

Предикторы: **CND** – расстояние до водотоков (дренажной сети); **NO** – отрицательная открытость (замкнутость ландшафта); **BIO15** – сезонность осадков (коэффициент вариации); **AgPot** – агроэкологический потенциал; **BIO2** – среднесуточная разность; **BIO7** – годовой диапазон температур; **BIO1** – среднегодовая температура; **BIO3** – изотермичность; **BIO5** – максимальная температура самого теплого месяца; **RSP** – относительное положение на склоне.

RFE

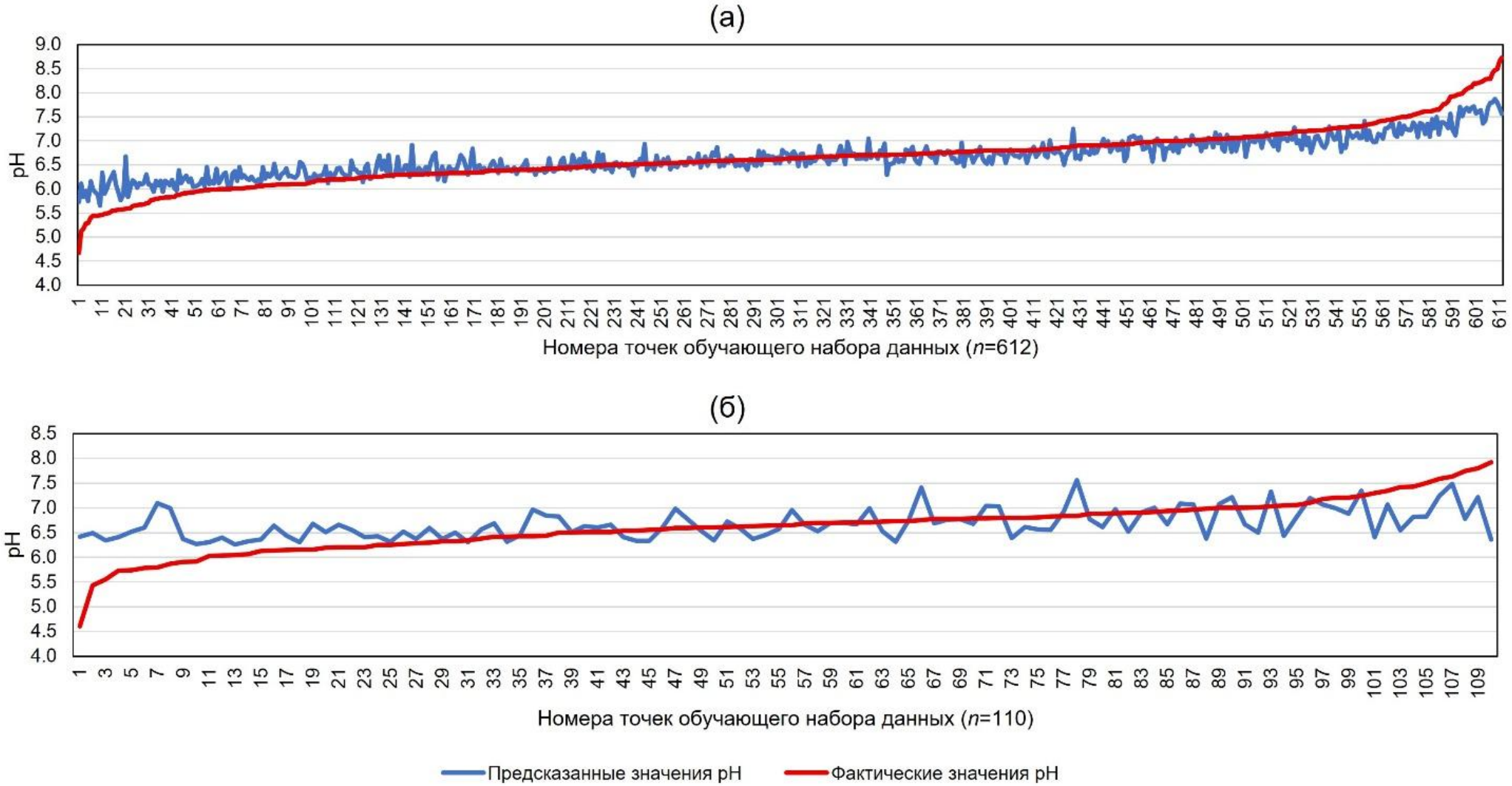


$RMSE_{ВНД}$=0,45; $MAPE_{ВНД}$=4,9; $MAE_{ВНД}$=0,29

Предикторы: **NO** – отрицательная открытость (замкнутость ландшафта); **CND** – расстояние до водотоков (дренажной сети); **BIO15** – сезонность осадков (коэффициент вариации); **RSP** – относительное положение на склоне; **BIO3** – изотермичность; **BIO2** – среднесуточная разность; **AgPot** – агроэкологический потенциал; **BIO19** – количество осадков в самом холодном квартале; **BIO5** – максимальная температура самого теплого месяца; **BIO1** – среднегодовая температура.

В общем случае $R^2 = 1$ означает, что модель идеально эффективна для моделирования, а $R^2 = 0$ – модель объясняет вариацию изучаемого показателя не лучше, чем среднее значение. $R^2_{ОНД}$ показывает, насколько качественно модель аппроксимирует фактические данные, на которых она обучалась. $R^2_{ВНД}$ отражает реальную предсказательную способность модели для новых данных, т.е. тех данных, которые не использовали в обучающем наборе данных.

Сравнение фактических и предсказанных значений pH в обучающем (а) и валидационном (б) наборах данных



ВЫВОДЫ

Методы отбора предикторов (RFE и Boruta) позволили выявить наиболее важные предикторы для картографирования pH; к числу наиболее важных относились предикторы, характеризующие рельеф и климат.

Сравнительный анализ показал, что у модели с предикторами, отобранными с использованием метода RFE, лучшие показатели эффективности моделирования: $R^2_{ОНД}$=0,81; $R^2_{ВНД}$=0,26; корень из среднеквадратической ошибки $RMSE_{ВНД}$=0,45; средняя абсолютная процентная ошибка $MAPE_{ВНД}$=4,9; средняя абсолютная ошибка $MAE_{ВНД}$=0,29.

Согласно составленной карте, изучаемые почвы характеризовались сильнокислой (5,0–6,0), слабокислой (6,0–6,5), нейтральной (6,5–7,5) и слабощелочной (7,5–8,5) реакцией среды. Сильнокислые и слабокислые почвы выявлены в почвах северной, западной и южной частях Тогучинского района, а почвы восточной и частично центральной части района характеризовались нейтральной и слабощелочной реакцией среды.

Спасибо за внимание!