ГЕОЭКОЛОГИЯ. ИНЖЕНЕРНАЯ ГЕОЛОГИЯ. ГИДРОГЕОЛОГИЯ. ГЕОКРИОЛОГИЯ, 2023, № 5, с. 88–96

МЕТОДОЛОГИЯ И МЕТОДИКА ИССЛЕДОВАНИЙ

УДК 504.064.2.001.18

МОДЕЛИРОВАНИЕ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ХРОМА И МАРГАНЦА В ПОЧВЕ: ПОДБОР ОБУЧАЮЩЕГО ПОДМНОЖЕСТВА

© 2023 г. А. С. Буторова[1,2,*], А. В. Шичкин[1,], А. П. Сергеев[1,***], Е. М. Баглаева[1,****], А. Г. Буевич[1,*****]**

[1]*Институт промышленной экологии Уральского отделения Российской академии наук (ИПЭ УрО РАН), ул. С. Ковалевской 20, Екатеринбург, 620990 Россия*

[2]*Уральский федеральный университет имени первого Президента России Б.Н. Ельцина (ФГАОУ ВО "УрФУ им. Б.Н. Ельцина), ул. Мира 19, Екатеринбург, 620002 Россия*

**E-mail: a.s.butorova@urfu.ru*

***E-mail: and@ecko.uran.ru*

****E-mail: sergeev@ecko.uran.ru*

*****E-mail: e.m.baglaeva@urfu.ru*

******E-mail: bag@ecko.uran.ru*

Поступила в редакцию 05.06.2023 г.
После доработки 28.07.2023 г.
Принята к публикации 08.09.2023 г.

Выбор метода разбиения исходных данных на обучающее и тестовое подмножества в моделях на основе искусственных нейронных сетей (ИНС) – недостаточно изученная проблема непрерывной интерполяции пространственно-временного поля. В частности, выбор наилучшего обучающего подмножества для моделирования пространственного распределения элементов в верхнем слое почвы – нетривиальная задача, поскольку точки отбора проб не эквивалентны. Они содержат разное количество "информации" в каждой конкретной модели, поэтому при моделировании целесообразно задействовать большинство точек, содержащих "полезную" для этой модели информацию. Неправильное разбиение данных может привести к неточным и чрезвычайно изменчивым характеристикам модели, высокой дисперсии и систематической ошибке в сгенерированных результатах. В качестве исходных данных были взяты данные о содержании хрома (Cr) и марганца (Mn) в верхнем слое почвы жилых районов в г. Ноябрьск (субарктическая зона России). Разработан трехэтапный алгоритм извлечения исходных данных с разбиением на обучающее и тестовое подмножества для моделирования пространственного распределения этих тяжелых металлов (ТМ) Для построения модели пространственного распределения содержания ТМ в верхнем слое почвы использовался многослойный персептрон (MLP), который учитывал пространственную неоднородность и правила обучения. Структура MLP была выбрана путем минимизации среднеквадратичной ошибки. Все точки разделились на три класса: "полезные", "обычные" и "бесполезные", по количеству попаданий в обучающее подмножество. Учет этой информации на этапе разбиения исходных данных позволяет повысить точность прогностической модели.

Ключевые слова: *моделирование*, *искусственные нейронные сети*, *обучающее подмножество*, *почва*, *тяжелые металлы*

DOI: 10.31857/S0869780923050028, **EDN:** ZZEIKY

ВВЕДЕНИЕ

Искусственные нейронные сети (ИНС) становятся все более востребованными для изучения пространственного распределения какого-либо признака. Модели на основе ИНС обеспечивают приемлемую точность при моделировании сложных экологических задач [2, 14, 17, 20, 21]. Прогностическая точность, достигаемая с помощью ИНС, часто выше, чем у других методов [1, 8, 18, 25].

Особенно важно прогнозирование пространственного распределения признака в свете растущих темпов изменения климата, которые проявляются в арктических и субарктических регионах планеты [6, 24]. Точные прогнозы позволят обществу более эффективно реагировать на вероятное негативное воздействие грядущих изменений.

Методы повышения точности прогнозов могут касаться как улучшения самих моделей ИНС (выбора типа и архитектуры, обучающих алгоритмов, создания гибридных моделей и т.д.), так и работы с исходными данными. Для моделирования необходимо разбить пространственно расположенную выборку на обучающее и тестовое подмножества с учетом вида исходных данных [10, 16, 26].

Выбор способа разбиения исходных данных на обучающее и тестовое подмножества в моделях ИНС – это проблема непрерывной интерполяции пространственно-временного поля, которая изучена недостаточно [9]. Данные, полученные в ходе мониторинга (скрининга) для оценки уровня загрязнения окружающей среды в неконтролируемых местах, зачастую не могут быть отобраны по равномерной сетке. Это связано со значительными различиями в характеристиках этих мест (перепадом высот, различными типами почв, городскими воздействиями и др.). Кроме того, точки отбора проб не эквивалентны для определения закономерностей распределения содержания элементов в верхнем слое почвы. Ошибки и выбросы, присутствующие в распределении, могут привести к погрешностям.

Методология разбиения данных оказывает значительное влияние на качество подмножеств, которые используются для обучения и тестирования ИНС. Некорректное разбиение данных может привести к неточным и чрезвычайно изменчивым характеристикам модели, высокой дисперсии и систематической ошибке в сгенерированных результатах. Статистический анализ результатов показывает, что такое разбиение данных может привести к снижению прогностической эффективности модели [23, 26]. Однако разработчики моделей ИНС редко уделяют должное внимание подбору методологии выборки [7].

В этой статье обсуждается разбиение исходных данных на основе статистического подсчета попаданий точек, включенных в обучающее подмножество. Такой подход позволяет исследователю учитывать структуру исходных данных. Разделение входных данных на N неперекрывающихся подмножеств, многократное обучение на N-1 подмножествах и тестирование на исключенном подмножестве, так называемая перекрестная проверка, часто используются для оценки производительности алгоритма обучения [9, 15]. Повторяющееся разбиение позволит получить статистическое распределение попаданий каждой точки в обучающее подмножество.

Цель работы – создание алгоритма определения наиболее полезных точек для включения в обучающее подмножество для задач интерполяции с помощью многослойного персептрона для моделирования содержания Cr и Mn в верхнем слое почвы.

МАТЕРИАЛЫ И МЕТОДЫ

Место отбора проб

Обследование почв было проведено в жилых зонах субарктического г. Ноябрьск (Ямало-Ненецкий автономный округ, Россия (рис. 1)). Это территория, расположенная севернее 60-й северной параллели. Ноябрьск находится на водоразделе двух крупнейших рек Сибири (Обь и Пур) в природной зоне тайги, в окружении множества небольших озер, рек и болот. Это один из самых молодых городов России, основанный в 1976 г.

Основная промышленность города – добыча углеводородов. Этот регион является субарктическим климатическим районом (Dfc по климатической классификации Кеппена). Район расположен в зоне распространения многолетнемерз-

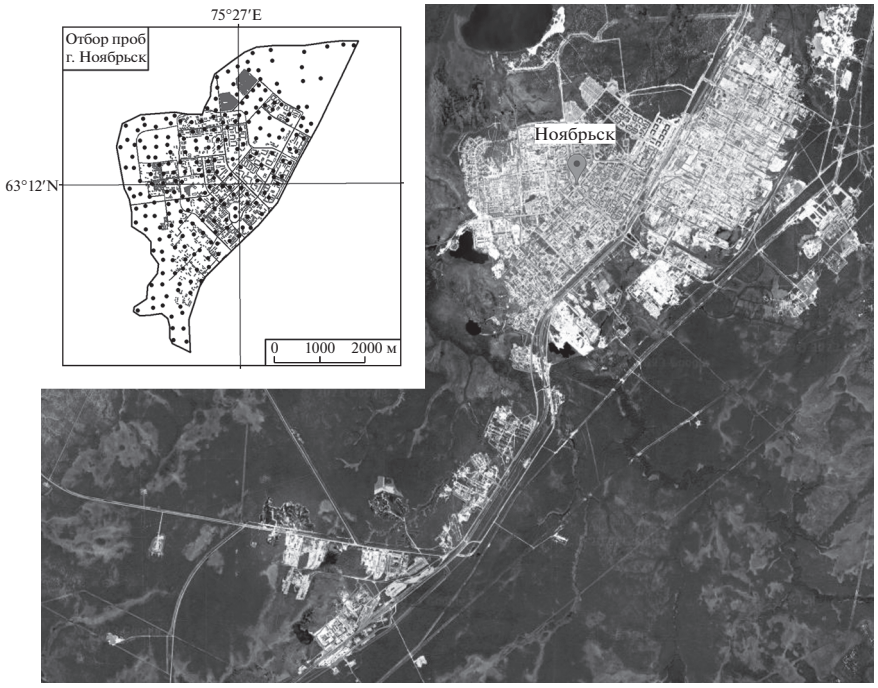

Рис. 1. Место отбора проб.

Таблица 1. Характеристики места отбора проб (жилая зона)

Место	Координаты	Количество образцов почвы	Тип почвы [4]	Текстура почвы	Тип почвы по FAO/UNESCO
Ноябрьск	63.2° N, 75.5° E	237	Глеевая таежная	100% песок	Gd 23-1ab

лых пород. Характеристики места отбора проб приведены в табл. 1.

Отбор проб почвы и химический анализ

Для исследований отбирался верхний слой урбанизированной почвы на глубине 0.05 м. Отбор проб почвы производился на нетронутых участках в узлах квадратной сетки с шагом 250 м. Их фактические географические координаты определились при отборе проб непосредственно на местности. Семь кернов были взяты на площади 1×1 м^2 с помощью пробоотборника из нержавеющей стали с внутренним диаметром 0.05 м и упакованы в двойные полиэтиленовые мешки. Внутренний пакет был промаркирован идентификатором образца. Масса каждого высушенного образца составляла примерно 1 кг [7]. Образцы почвы были доставлены в сертифицированную лабораторию в соответствии со стандартом ISO/IEC 17025:2005.

Подготовка и химический анализ проводились в соответствии с действующими нормативными требованиями Федеральной системы сертификации РФ. Подготовка образцов почвы заключалась в сушке на воздухе при стандартных условиях, просеивании через сито 1 мм, разделении на четвертины и гомогенизации до 20-граммовых дополнительных образцов и измельчении до зерен диаметром 0.074 мм.

Общее содержание Si, K, Ca, V, Cr, Mn, Ni, Cu и Zn в образцах почвы было проанализировано с помощью масс-спектрометрии с индуктивно связанной плазмой (ICP-MS); прибор представлял собой Perkin Elmer ELAN 9000 с пределом обнаружения 0.1 мг/кг для каждого элемента. Перед анализом образцы почвы растворялись концентрированной азотной и плавиковой кислотой. После перемешивания и нагревания до 95°C раствор обрабатывался концентрированной хлорной кислотой и после охлаждения обрабатывался соляной кислотой при медленном нагревании в течение 30 мин. После охлаждения раствор был разбавлен до 50 мл деионизированной водой, тщательно перемешан и помещен в полиэтиленовый флакон. Содержание Cr и Mn использовалось в качестве исходных данных для моделирования.

Алгоритм разбиения

Алгоритм разбиения исходных данных состоял из трех шагов (рис. 2):

1. Набор исходных данных был 1000 раз случайным образом разбит на обучающее и тестовое подмножества в соотношении 70% к 30% соответственно. Таким образом, были получены 1000 разбиений на два непересекающихся множества.

2. Было построено 1000 сетей для каждого случайного обучающего подмножества. Для каждой обученной сети определялась среднеквадратичная ошибка (RMSE) предсказания тестового подмножества.

3. Для этого этапа было отобрано 100 сетей (10% от общего числа 0.1-квантилей) с наименьшим значением RMSE. Частоты попадания в обучающее подмножество рассчитывались суммированием попаданий каждой точки пространства по выбранным сетям. Точки, для которых частота совпадений в обучающем подмножестве превышала 75%, выбирались для включения в обучающее подмножество.

Построение MLP

Для тестирования производительности нового метода потребовалась простая в использовании настраиваемая модель искусственной нейронной сети. Многослойный персептрон (MLP) с алгоритмом обучения Левенберга-Марквардта [12] был выбран в качестве эталонной модели ИНС. Эта простая и легкая в обучении сеть отлично зарекомендовала себя в прогнозировании пространственного распределения. Использование предложенной методики извлечения обучающего подмножества актуально для всех моделей, прогнозирующих пространственное распределение признака.

Построение модели MLP заключалось в подборе параметров: количества скрытых слоев и количества нейронов внутри каждого скрытого слоя. В процессе обучения MLP связи между нейронами, устанавливаемые путем присвоения весов, обновляют значения веса и смещения в соответствии с функцией потерь как наименьшей суммы квадратов ошибок в обучающем подмножестве (алгоритм обучения Левенберга-Марквардта).

1. Структура MLP была выбрана компьютерным моделированием на основе минимизации RMSE. Входной слой MLP состоял из двух нейронов (пространственные координаты точек на территории Ноябрьска x и y). MLP имел один скрытый слой с числом нейронов от 2 до 20. Выходной слой MLP включал один нейрон (содержание химического элемента).

Рис. 2. Блок-схема алгоритма разбиения.

2. Каждая сеть (1000 сетей, как показано на рис. 2) имела структуру с соответствующим количеством нейронов в скрытом слое и была обучена 500 раз. Затем выбиралась лучшая из них (с минимальным среднеквадратичным отклонением для каждого числа нейронов).

3. Структура сети MLP с определением оптимального количества нейронов в скрытом слое для каждой области.

Оценка точности модели

Для оценки точности прогноза между прогнозируемым и исходным наборами данных используются MAE (1) и RMSE (2):

$$MAE = \frac{\sum_{i=1}^{n} \left| p_i(x) - o_i(x) \right|}{n}, \qquad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(p_i(x) - o_i(x) \right)^2}{n}}, \qquad (2)$$

где $p_i(x)$ и $o_i(x)$ – прогнозируемая и наблюдаемая концентрации соответственно, n – количество точек.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Описательные статистики содержания Cr и Mn обследованной территории представлены в табл. 2. Сравнивалось содержание фоновых элементов в почвах в Уральском регионе (Ural Clarke) и в мире (World Clarke). Суммарное содержание Cr на городском фоне не превышает контрольных значений, а общее содержание Cr в зонах аномалий в несколько раз превышает Ural Clarke [3, 5]. Общее содержание Cr в подзолах,

Таблица 2. Описательные статистики исходных данных

Место	Элемент	Содержание, мг/кг			Коэффициент вариации, %	Асимметрия	Эксцесс	p-уровень Шапиро–Уилка
		Min–Max / Среднее	Стандартное отклонение	Медиана				
Ноябрьск	Cr	17 – 140 / 63	23	60	37	0.8	0.6	<0.05
	Mn	62 – 529 / 141	56	130	40	3.8	20.8	<0.05

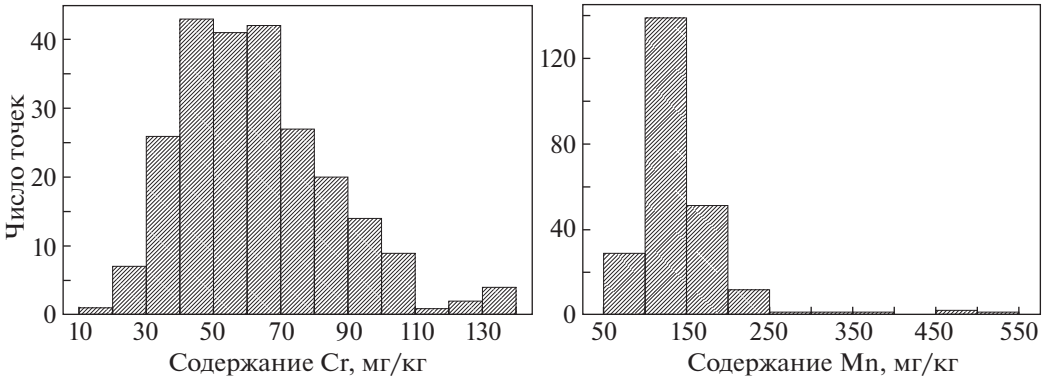

Рис. 3. Гистограмма содержания элементов.

как известно, находится в диапазоне от 2.6 до 34 мг/кг в Канаде [11], от 3 до 200 мг/кг в США [19], от 18 до 25 мг/кг в России [13]. Известно, что общее содержание Mn в подзолах находится в диапазоне от 7 до 2000 мг/кг в США и от 135 до 310 мг/кг в России [13]. Общее содержание Mn не превышает контрольных средних значений 545 мг/кг [13] в мире или 580 мг/кг в России [22].

Согласно данным, представленным в табл. 2, распределение вероятностей содержаний элементов имеет тяжелый правый хвост. Высокий коэффициент эксцесса для Mn указывает на выбросы (рис. 3).

Содержание Cr имеет одномодальное распределение (см. рис. 3). Содержание Mn имеет правый хвост. Тест Шапиро–Уилка показал, что распределение содержания Cr и Mn статистически значимо отличается от нормального ($p < 0.05$).

Был определен лучший набор образцов почвы для обучающего подмножества для прогнозирования содержания элементов в тестовом подмножестве с минимальной среднеквадратической ошибкой. Скрытые слои состояли из восьми нейронов. Окончательная структура модели MLP составила 2-8-1 для каждого элемента.

В табл. 3 приведены основные результаты распределения ошибок полученных моделей для обследованной территории. Распределения RMSE и MAE унимодальные и симметричные с низкими коэффициентами вариации. Тест Шапиро–Уилка показал, что полученные распределения RMSE статистически значимо отличаются от нормального ($p = 0.0005$) для содержания Mn и не отличаются от нормального для содержания Cr с $p > 0.52$. Жирным шрифтом в табл. 3 выделены лучшие значения ошибок: это RMSE для моделей Cr. Распределения соответствующих элементов симметричны и унимодальны, хотя и отличаются от нормальных. Минимум (Min) и максимум (Max) для RMSE и MAE в табл. 3 показывают точность между прогнозом и исходными данными с единичным разбиением данных. Включение наиболее многообещающих точек с точки зрения частоты попаданий в обучающий набор для задач интерполяции многослойным перцептроном содержания элементов в верхнем слое почвы повышает точность прогноза.

Гистограмма RMSE моделирования содержания элемента показана на рис. 4. Пунктирные 0.1-квантильные значения на рис. 4 меньше средних значений для каждого элемента.

На рис. 5 показано пространственное расположение точек обучающего подмножества для визуализации полученных результатов. Совпадающие для двух элементов точки отбора проб составляют около 9% (22 точки) от общего числа точек. Чем больше наблюдается особенностей содержа-

Таблица 3. Оценка точности

Точность оценки	Элемент	Содержание, мг/кг			Коэффициент вариации, %	Асимметрия	Эксцесс	p-уровень Шапиро–Уилка
		min–max Среднее	Медиана	Стандартное отклонение				
MAE	Cr	12 – 20 16	16	1	8	0.14	–0.07	0.20
	Mn	27 – 80 58	58	7	12	–0.61	0.73	<0.05
RMSE	Cr	15 – 27 21	**21**	**2**	**9**	**0.04**	**–0.21**	**0.52**
	Mn	33 – 84 64	64	6	10	–0.80	1.48	<0.05

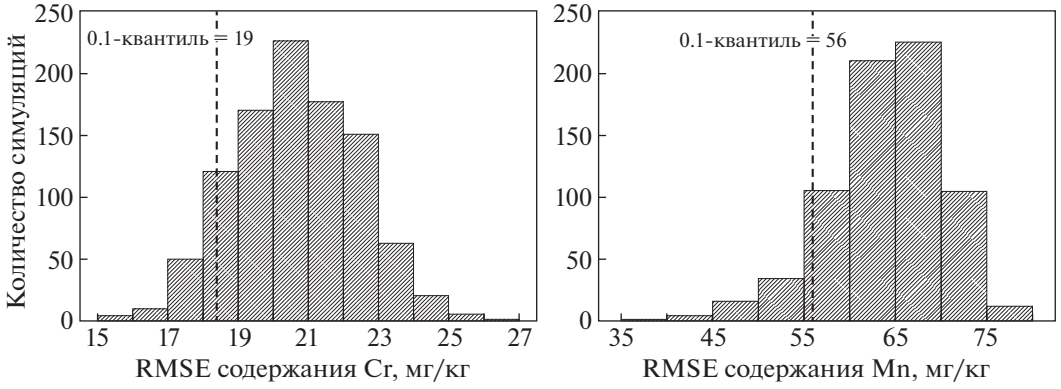

Рис. 4. Гистограмма RMSE моделирования содержания элемента.

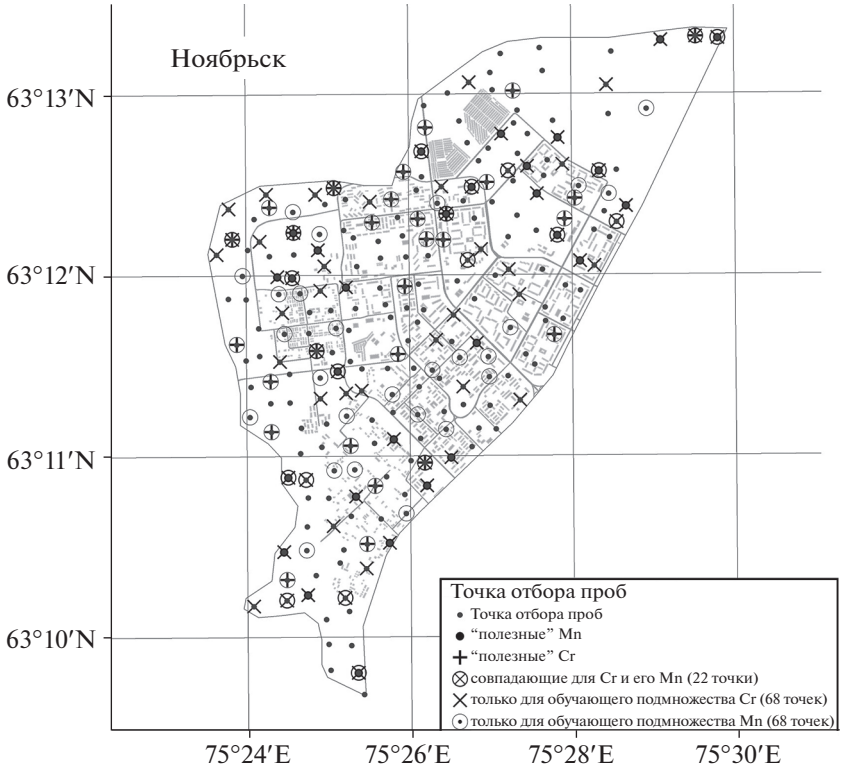

Рис. 5. Пространственное расположение обучающего подмножества для г. Ноябрьск.

ния территориального элемента, тем больше совпадающих точек выборки можно выделить при построении нейронной сети.

В финальное обучающее подмножество чаще всего входят граничные точки и точки, учитывающие территориальные особенности, например, морфологию места съемки, расположение улиц и т.д. В тестовое подмножество входят оставшиеся 30% точек.

Элементный состав однородный, без аномалий. Репрезентативность точек отбора проб практически такая же. Все точки имеют равные шансы попасть в обучающее подмножество.

Если существует правило пространственного распределения по области для содержимого элемента, то оценка количества попаданий точки в обучающее подмножество может определить полезность каждой точки для обучения. На рис. 6 показано распределение попаданий точек выборки в обучающее подмножество для оценки репрезентативности каждой из них. Для каждого элемента по количеству попаданий в обучающее подмножество точки делятся на три класса: “полезные”, “обычные” и “бесполезные”.

“Полезные” точки — это точки, сформировавшие правило пространственного распределения. “Обычный” класс – это равные точки без особенностей. Равномерное распределение элемента в почве означает практически полное отсутствие “полезных” и “бесполезных” точек. “Бесполез-

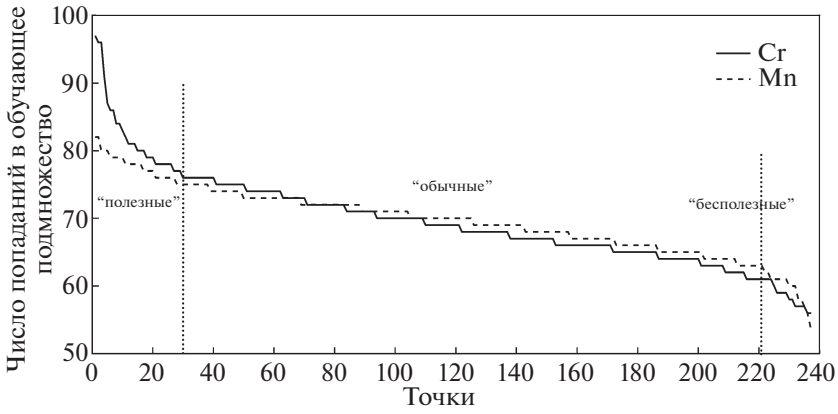

Рис. 6. Оценка репрезентативности точек выборки для обучающего подмножества.

ные" точки содержат существенную ошибку или недостаток некоторой информации об изучаемом явлении или процессе, что сводит на нет ее ценность для модели интерполяции. Они являются своего рода репрезентативными выбросами для данной территории.

Модель MLP, основанная на методе контролируемого разбиения, оказалась более точной, чем случайное разбиение [1]. Однако контролируемое разбиение требует предварительного знания распределения признака. Использование алгоритма, представленного в данной работе, не требует априорной информации о репрезентативности точек. Точность между наборами прогнозируемых и исходных данных этого метода оказалась не ниже, чем у метода контролируемого разбиения на все элементы. Недостатком метода является большой объем вычислений, необходимых для построения модели.

ЗАКЛЮЧЕНИЕ

Точность интерполяции экологических данных преимущественно связана с предварительной подготовкой исходных данных для моделирования. В этом исследовании выполняется разбиение исходных данных для обучающего подмножества так, чтобы учесть частоту попадания каждой точки в те модели, для которых ошибки были минимальными. Искусственная нейронная сеть показала лучшую точность на данных с унимодальными характеристиками.

По количеству попаданий в обучающее подмножество для каждой территории выделяются "полезные", "обычные" и "бесполезные" точки. "Полезные" точки должны быть включены в обучающее подмножество для повышения точности модели ИНС, а "бесполезные" – нет.

В последующих работах авторы планируют усовершенствовать предложенный подход и намерены подтвердить преимущества описанной методики извлечения обучающей выборки при моделировании пространственного распределения химических элементов в верхнем слое почвы на примере других урбанизированных территорий.

СПИСОК ЛИТЕРАТУРЫ

1. *Буевич А.Г., Субботина И.Е., Шичкин А.В. и др.* Оценка пространственного распределения хрома в субарктическом Ноябрьске с использованием кокригинга, генерализованной регрессионной нейронной сети, многослойного персептрона и гибридной техники // Геоэкология. 2019. № 2. С. 77–86.
2. *Буторова А.С., Сергеев А.П., Шичкин А.В. и др.* Метод встречного прогнозирования пространственного ряда на примере содержания пыли в снеговом покрове // Геоинформатика. 2022. № 1. С. 32–39.
3. *Войткевич Г.В., Мирошников А.Е., Поваренных А.С., Прохоров В.Г.* Краткий справочник по геохимии. М.: Недра, 1977. 184 с.
4. *Добровольский Г.В., Урусевская И.С.* География почв. 2-е изд., перераб. и доп. М.: Изд-во МГУ, Изд-во "КолосС", 2004. 460 с.
5. *Сает Ю.Е.* Геохимия окружающей среды [Кол. авт.: Ю.Е. Сает, Б.А. Ревич, Е.П. Янин и др.]. М.: Недра, 1990. С. 84–108.
6. AMAP. Snow, Water, Ice and Permafrost. Summary for Policy-makers / Arctic Monitoring and Assessment Programme (AMAP). Oslo, Norway. 2017. 20 p.
7. *Baglaeva E.M., Sergeev A.P., Shichkin A.V., Buevich A.G.* The Effect of Splitting of Raw Data into Training and Test Subsets on the Accuracy of Predicting Spatial Distribution by a Multilayer Perceptron // Mathematical Geosciences. 2020. V. 52. P. 111–121.
8. *Dai F., Zhoua O., Lva Z., Wang X., Liu G.* Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau // Ecological Indicators. 2014. V. 45. P. 184–194.
9. *Demyanov V., Gloaguen E., Kanevski M.* A special issue on data science for geosciences // Mathematical Geosciences. 2020. V. 52. P. 1–3.
10. *Fernandez J.M., Mayerle R.* Sample selection via angular distance in the space of the arguments of an artificial neural network // Computers and Geosciences. 2018. V. 114. P. 98–106.

11. *Frank R., Ishida K., Suda P.* Metals in agricultural soils of Ontario // Canadian Journal of Soil Science. 1976. V. 56. P. 181–196.
12. *Goovaerts P.* Geostatistics in soil science: State of the art and perspectives // Geoderma. 1999. V. 89. P. 1–45.
13. *Kabata-Pendias A.* Trace elements in soils and plants / Taylor and Francis Group CRC Press. 2011. P. 201–260.
14. *Liodakis S., Kyriakidis P., Gaganis P.* Conditional Latin Hypercube Simulation of (Log)Gaussian Random Fields // Mathematical Geosciences. 2018. V. 50. P. 127–146.
15. *Malof J.M., Reichman D., Collins L.M.* How do we choose the best model? The impact of cross-validation design on model evaluation for buried threat detection in ground penetrating radar / Материалы конференции Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIII. 2018. V. 10628. 106280C.
16. *Nath A., Subbiah K.* The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins // Neurocomputing. 2018. V. 272. P. 294–305.
17. *Sakizadeh M., Mirzaei R., Ghorbani H.* Support vector machine and artificial neural network to model soil pollution: a case study in Semnan Province, Iran // Neural Computing & Applications. 2017. V. 28. P. 3229–3238.
18. *Sergeev A.P., Buevich A.G., Baglaeva E.M., Shichkin A.V.* Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals // Catena. 2019. V. 174. P. 425–435.
19. *Shacklette H.T., Boerngen J.G.* Element concentrations in soils and other surficial materials of the conterminous United States / U.S. Geological Survey professional paper // United states government printing office, Washington. 1984. 105 p.
20. *Shaker R.R., Ehlinger T.J.* Exploring non-linear relationships between landscape and aquatic ecological condition in southern Wisconsin: A GWR and ANN approach // International Journal of Applied Geospatial Research. 2014. V. 5(4). P. 1–20.
21. *Sun C., Liu J., Wang Y., Sun L., Yu H.* Multivariate and geostatistical analyses of the spatial distribution and sources of heavy metals in agricultural soil in Dehui, Northeast China // Chemosphere. 2013. V. 92 (5). P. 517–523.
22. *Timofeeva Y.O., Kosheleva Y., Semal V., Burdukovskii M.* Origin, baseline contents, and vertical distribution of selected trace lithophile elements in soils from nature reserves, Russian Far East // Journal of Soils and Sediments. 2018. V. 18 (3). P. 968–982.
23. *Wieland R., Mirschel W., Zbell B., et al.* A new library to combine artificial neural networks and support vector machines with statistics and a database engine for application in environmental modeling // Environmental Modelling & Software. 2012. V. 25. P. 412–420.
24. WMO. The Global Climate in 2015–2019 // World Meteorological Organization (WMO-№ 1249), Geneva, Switzerland. 2020. 24 p.
25. *Worsham L., Markewitz D., Nibbelink N.* Incorporating spatial dependence into estimates of soil carbon contents under different land covers // Soil Science Society of America Journal. 2010. V. 74. P. 635–646.
26. *Ziggah Y.Y., Youjian H., Tierra A.R., Laari P.B.* Coordinate Transformation between Global and Local Data Based on Artificial Neural Network with K-Fold Cross-Validation in Ghana // Earth Sciences Research Journal. 2019. V. 23 (1). P. 67–77.

MODELING OF THE SPATIAL DISTRIBUTION OF CHROME AND MANGANESE IN SOIL: SELECTION OF A TRAINING SUBSET

A. S. Butorova[a,b,#], A. V. Shichkin[a,##], A. P. Sergeev[a,###], E. M. Baglaeva[a,####], and A. G. Buevich[a,#####]

[a] *Institute of Industrial Ecology, Ural Branch, Russian Academy of Sciences, ul. S.Kovalevskoi 20, Yekaterinburg, 620990 Russia*

[b] *Ural Federal University, ul. Mira 19, Yekaterinburg, 620002 Russia*

[#] *E-mail: a.s.butorova@urfu.ru*

[##] *E-mail: and@ecko.uran.ru*

[###] *E-mail: sergeev@ecko.uran.ru*

[####] *E-mail: e.m.baglaeva@urfu.ru*

[#####] *E-mail: bag@ecko.uran.ru*

The selection of a method for dividing the raw data into training and test subsets in models based on artificial neural networks (ANN) is an insufficiently studied problem of continuous space-time field interpolation. In particular, selecting the best training subset for modeling the spatial distribution of elements in the topsoil is not a trivial task, since the sampling points are not equivalent. They contain a different amount of "information" in point of each specific model, therefore, when modeling, it is advisable to use most of the points containing information which is "useful" for this model. Incorrect data division may lead to inaccurate and highly variable model characteristics, high variance and bias in the generated results. The raw data included contents of chromium (Cr) and manganese (Mn) in the topsoil in residential areas of Noyabrsk (a city in Russian subarctic zone). A three-stage algorithm for extracting raw data with a division into training and test subsets has been developed for modeling the spatial distribution of heavy metals. According to the algorithm, the ini-

tial data set was randomly divided into training and test subsets. For each training subset, an ANN based on multilayer perceptron (MLP) was built and trained. MLP was used to model the spatial distribution of heavy metals in the upper soil layer, which took into account spatial heterogeneity and learning rules. The MLP structure was chosen by minimizing the root mean square error (RMSE). The networks with the lowest RMSE were selected, and the number of hits into the training subset of each point in space was calculated. By the number of hits in the training subset, all points were divided into three classes: "useful", "ordinary" and "useless". Taking this information into account, at the stage of the raw data division it possible to increase the accuracy of the predictive model.

Keywords: *modeling*, *artificial neural networks*, *training subset*, *soil*, *heavy metals*

REFERENCES

1. Buevich, A.G., Subbotina, I.E., Shichkin, A.V., et al. [Assessment of chrome distribution in subarctic Noyabrsk using co-kriging, generalized regression neural network, multilayer perceptron, and hybrid technics]. *Geoekologiya*, 2019, no. 2, pp. 77–86. (in Russian)
2. Butorova, A.S., Sergeev, A.P., Shichkin, A.V., et al. [Counter-prediction method of the spatial series on the example of the dust content in the snow cover]. *Geoinformatika*, 2022, no. 1, pp. 32–39. (in Russian)
3. Voitkevich, G.V., Miroshnikov, A.E., Povarennykh, A.S., Prokhorov, V.G. [The short manual in geochemistry]. Moscow, Nedra Publ., 1977, 184 p. (in Russian)
4. Dobrovolskii, G.V., Urusevskaya, I.S. [Soil geography]. Moscow, MSU Publ., KolosS Publ., 2004, 460 p. (in Russian)
5. Saet, Yu.E., Revich, B.A., Yanin, E.P. [Environment geochemistry]. Moscow, Nedra Publ., 1990, pp. 84–108. (in Russian)
6. AMAP. Snow, water, ice and permafrost. Summary for policy-makers. In: Arctic Monitoring and Assessment Programme (AMAP), Oslo, Norway, 2017, 20 p.
7. Baglaeva, E.M., Sergeev, A.P., Shichkin, A.V., Buevich, A.G. The Effect of splitting of raw data into training and test subsets on the accuracy of predicting spatial distribution by a multilayer perceptron. *Mathematical Geosciences*, 2020, vol. 52, pp. 111–121.
8. Dai, F., Zhoua, O., Lva, Z., Wang, X., Liu, G. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. *Ecological Indicators*, 2014, vol. 45, pp. 184–194.
9. Demyanov, V., Gloaguen, E., Kanevski, M. A special issue on data science for geosciences. *Mathematical Geosciences*, 2020, vol. 52, pp. 1–3.
10. Fernandez, J.M., Mayerle, R. Sample selection via angular distance in the space of the arguments of an artificial neural network. *Computers and Geosciences*, 2018, vol. 114, pp. 98–106.
11. Frank, R., Ishida, K., Suda, P. Metals in agricultural soils of Ontario. *Canadian Journal of Soil Science*, 1976, vol. 56, pp. 181–196.
12. Goovaerts, P. Geostatistics in soil science: State of the art and perspectives. *Geoderma*, 1999, vol. 89, pp. 1–45.
13. Kabata-Pendias, A. Trace elements in soils and plants. Taylor and Francis Group CRC Press, 2011, pp. 201–260.
14. Liodakis, S., Kyriakidis, P., Gaganis, P. Conditional Latin hypercube simulation of (log)Gaussian random fields. *Mathematical Geosciences*, 2018, vol. 50, pp. 127–146.
15. Malof, J.M., Reichman, D., Collins, L.M. How do we choose the best model? The impact of cross-validation design on model evaluation for buried threat detection in ground penetrating radar. In: Proc. of Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIII, 2018, vol. 10628, 106280C.
16. Nath, A., Subbiah, K. The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins. *Neurocomputing*, 2018, vol. 272, pp. 294–305.
17. Sakizadeh, M., Mirzaei, R., Ghorbani, H. Support vector machine and artificial neural network to model soil pollution: a case study in Semnan Province, Iran. *Neural Computing & Applications*, 2017, vol. 28, pp. 3229–3238.
18. Sergeev, A.P., Buevich, A.G., Baglaeva, E.M., Shichkin, A.V. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. *Catena*, 2019, vol. 174, pp. 425–435.
19. Shacklette, H.T., Boerngen, J.G. Element concentrations in soils and other surficial materials of the conterminous United States. U.S. Geological Survey professional paper, US government printing office, Washington, 1984, 105 p.
20. Shaker, R.R., Ehlinger, T.J. Exploring non-linear relationships between landscape and aquatic ecological condition in southern Wisconsin: A GWR and ANN approach. *International Journal of Applied Geospatial Research*, 2014, vol. 5 (4), pp. 1–20.
21. Sun, C., Liu, J., Wang, Y., Sun, L., Yu, H. Multivariate and geostatistical analyses of the spatial distribution and sources of heavy metals in agricultural soil in Dehui, Northeast China. *Chemosphere*, 2013, vol. 92 (5), pp. 517–523.
22. Timofeeva, Y.O., Kosheleva Y., Semal V., Burdukovskii M. Origin, baseline contents, and vertical distribution of selected trace lithophile elements in soils from nature reserves, Russian Far East. *Journal of Soils and Sediments*, 2018, vol. 18 (3), pp. 968–982.
23. Wieland, R., Mirschel, W., Zbell, B., et al. A new library to combine artificial neural networks and support vector machines with statistics and a database engine for application in environmental modeling. *Environmental Modelling & Software*, 2012, vol. 25, pp. 412–420.
24. WMO. The Global Climate in 2015–2019. In: World Meteorological Organization (WMO-№ 1249), Geneva, Switzerland, 2020, 24 p.
25. Worsham, L., Markewitz, D., Nibbelink, N. Incorporating spatial dependence into estimates of soil carbon contents under different land covers. *Soil Science Society of America Journal*, 2010, vol. 74, pp. 635–646.
26. Ziggah, Y.Y., Youjian, H., Tierra, A.R., Laari, P.B. Coordinate transformation between global and local data based on artificial neural network with K-fold cross-validation in Ghana. *Earth Sciences Research Journal*, 2019, vol. 23 (1), pp. 67–77.