УДК 504.064.2.001.18

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ РАЗДЕЛЕНИЯ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ПРИ МОДЕЛИРОВАНИИ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ТЯЖЕЛЫХ МЕТАЛЛОВ В ВЕРХНЕМ СЛОЕ ПОЧВЫ

Е. М. Баглаева, А. П. Сергеев, А. В. Шичкин, А. Г. Буевич,
И. Е. Субботина, А. С. Буторова

Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия

Представлен сравнительный анализ пяти алгоритмов выбора обучающего подмножества для искусственных нейронных сетей, интерполирующих пространственное распределение признака по данным экологических скринингов компонентов окружающей природной среды. В качестве исходных данных использовали содержание хрома в верхнем слое почвы на селитебных территориях г. Ноябрьска (ЯНАО, Россия). Пространственные распределения содержания элементов в верхнем слое почвы интерполировались многослойным персептроном (МЛП). Статистики среднеквадратической ошибки (RMSE) рассчитывались как показатели точности для каждого алгоритма. Меньшую ошибку демонстрирует основанный на пространственном квотировании исходных данных алгоритм Space&quartile quote. Принимая во внимание информацию о статистике ошибок алгоритмов разделения исходного набора, можно повысить точность модели.

Ключевые слова: МЛП; разделение исходных данных; пространственное распределение; моделирование среды; репрезентативность.

1. Введение

Оценка и прогнозирование состояния компонентов окружающей природной среды является важнейшей составляющей оценки антропогенного воздействия и влияния на здоровье человека [1, 2]. Верхний слой почвы – наиболее объективный и устойчивый индикатор техногенного загрязнения почвы и экосистемы [3, 4]. Почвенная съемка обычно используется для изучения характера и характеристик окружающей среды [5, 6]. Каждый образец почвы содержит информацию о локальных пространственно-временных свойствах участка отбора проб [7], которые зависят от местных климатических, гидрогеологических и геологических условий, нерегулярных источников загрязнения и других постоянно меняющихся факторов. Для объективной оценки почвенных характеристик необходимо выбрать места отбора проб и взять как можно больше проб [8]. Предварительная оценка репрезентативности

точек отбора проб может снизить количество проб и получить большой объем информации об изучаемой территории [9, 10].

Почвенная съемка используется для картографирования пространственного распределения различных характеристик почвы, например, гранулометрического состава, объемной плотности, содержания химических реагентов и других [11]. Данные съемки часто сочетаются с методологией искусственной нейронной сети (ИНС) для реконструкции пространственного распределения признака [12, 13]. Почвенные данные представляют собой результаты обследования в нерегулярных точках отбора проб и используются в качестве исходных для обучения и тестирования модели ИНС. Существуют различные методы построения наилучшего обучающего набора [14–16]. Обычно ошибка ИНС сводится к минимуму. Исследование деления данных на обучающее и тестовое подмножество может привести к созданию наилучшего набора обучающих данных для оценки эффективности различных точек выборки для решения конкретной задачи [17, 18]. Разделение данных является важным фактором при построении ИНС, метод перекрестной проверки часто используется для оценки производительности моделей ИНС [19, 20]. Методология разделения выборочных данных может оказать значительное влияние на качество обучающих, тестовых и проверочных подмножеств для ИНС. Неправильное разделение данных может привести к высокой дисперсии и систематической ошибке в полученных результатах, а также к ложному отчету о прогностической эффективности модели [21, 22]. Способ разделения данных не зависит от типа модели ИНС, для проверки возможностей метода деления удобно использовать наиболее простую для понимания модель, например, многослойный персептрон (МЛП). Нейронная сеть МЛП обучена прогнозировать закономерности пространственно-временной изменчивости для коротких периодов [23, 24]. Таким образом, оценка репрезентативности точек отбора проб имеет большое практическое значение для построения оптимальной схемы пробоотбора при экологическом мониторинге, с другой стороны – для построения наилучшего обучающего множества для ИНС.

Методы отбора наилучшей обучающей выборки позволяют оценить репрезентативность расположения наиболее информативных точек выборки. В нашей статье мы сравниваем пять полученных (с использованием различных подходов к разделению данных) способов построения обучающего подмножества для моделирования пространственного распределения тяжелых металлов в верхнем слое почвы.

2. Материалы и методы

2.1. География отбора проб

Почвенная съемка была организована в селитебной зоне г. Ноябрьска (N63.2°, E75.5°), расположенного у южных границ Ямало-Ненецкого автономного округа, Россия (рис. 1). Основные предприятия города связаны с добычей углеводородных ресурсов. Климат резко континентальный. Город находится в природной зоне тайги. Преобладающий тип почвы – таежные подзолы [25].

2.2. Отбор проб почвы, подготовка образцов и химический анализ

Получение исходных данных состояло из трех этапов: отбор проб почвы, подготовка образцов и химический анализ. Процедура отбора проб почвы подробно описана [10]. На рис. 1 показана подробная схема пространственного отбора проб и отбор проб почвы на месте. Каждый образец был упакован в двойной полиэтилено-

вый пакет с идентификатором образца. Идентификатор был отмечен районом, географическими координатами (N6X.XXX°, E7X.XXX°) и количеством проб почвы. Упакованные образцы были доставлены в лабораторию.

Подготовка почвенных образцов включала четыре этапа (рис. 2): подсушивание на воздухе, просеивание, квартование и измельчение в соответствии с действующими стандартными требованиями, подробно описанными в [10]. Первый этап представлял собой сушку на воздухе при температуре 20 °С и давлении 101 МПа. Вторым этапом было просеивание через сито диаметром 1 мм для удаления крупных неорганических и органических остатков. Квартирование представляло собой перемешивание и гомогенизацию для получения 20 г однородного субобразца. Четвертый – измельчение на зерно диаметром 0,74 мм для химического анализа. Химический анализ включает (HNO_3 + HF + $HClO_4$) кислотное разложение и масс-спектрометрию с индуктивно связанной плазмой (Perkin Elmer, ELAN 9000). Измерено содержание в почве Si, K, Ca, V, Cr, Mn, Ni, Cu, Zn. В дальнейшем моделировали содержание хрома.

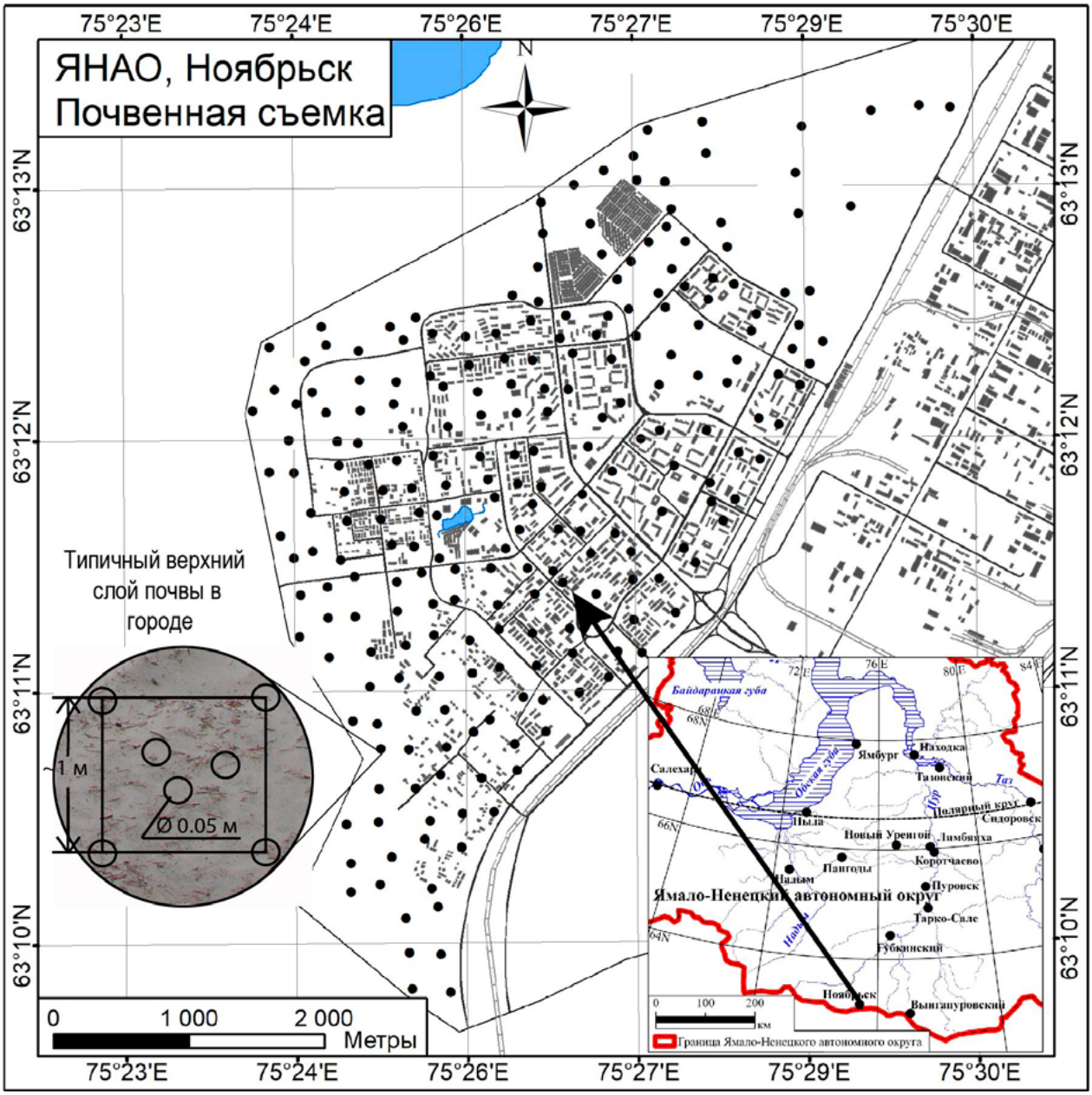

Рис. 1. Схема отбора проб

2.3. Подготовка данных для моделирования

Географические координаты были входными данными для моделирования, а содержание элемента – выходными. Для очистки исходных данных от выбросов использовался метод линейной интерполяции соседних значений [26]. Пространственные координаты и содержание Cr масштабировали от 0 до 1.

Рис. 2. Этапы пробоподготовки и химического анализа

2.4. МЛП

В качестве эталонной модели искусственной нейронной сети использовался многослойный персептрон. Построение модели МЛП заключалось в выборе количества нейронов внутри скрытого слоя. В процессе обучения МЛП связи между нейронами устанавливались путем присвоения весов, обновлений веса и значений смещения в соответствии с функцией потерь как наименьшей суммы квадратов ошибок в обучающей выборке (алгоритм обучения Левенберга – Марквардта). Мы использовали логистическую функцию активации.

Структура МЛП была выбрана с помощью компьютерного моделирования на основе минимального RMSE для всех данных (237 точек). Входной слой МЛП состоял из двух нейронов (пространственные координаты x и y). МЛП имел один скрытый слой с числом нейронов от 1 до 10. Выходной слой МЛП включал один нейрон (элемент содержимого).

2.5. Оценка точности модели

Была использована среднеквадратическая ошибка RMSE (1) для проверки точности прогнозирования между прогнозом и набором необработанных данных.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p(i)-o(i))^2}{n}}, \quad (1)$$

где $p(i)$ – прогнозируемые данные,

$o(i)$ – измеренные данные,

n – количество точек.

2.6. Выбор обучающего подмножества для построения МЛП

2.6.1. Случайный алгоритм

Полностью случайный (Random) алгоритм, при котором набор необработанных данных случайным образом разбивали на обучающее и тестовое подмножества в соотношении 70 %/30 % соответственно.

Каждую структуру сети (10 сетей) с соответствующим количеством нейронов в скрытом слое обучали десять раз и выбирали лучшую (с минимальным RMSE). Для каждого разделения получали 100 сетей: от 1 до 10 нейронов и десятикратное обучение. Среди всех этих сетей выбирали лучшую с минимальным RMSE.

2.6.2. Трехшаговый (3step) алгоритм разделения данных

Исходные данные многократно делили случайным образом на обучающее и тестовое подмножества в соотношении 70 %/30 % соответственно. Количество способов разделения равно числу сочетаний без повторения. Алгоритм разделения исходных данных состоит из трех шагов.

1. Набор исходных данных 1 000 раз случайным образом разделили на два непересекающихся множества, обучающую и тестовую выборки в соотношении 70 %/30 % соответственно. Таким образом, получили 1 000 обучающих и 1 000 тестовых подмножеств.

2. Для каждого случайного обучающего подмножества построили 1 000 сетей. Для каждой обученной сети определялась среднеквадратическая ошибка (RMSE) прогноза тестового набора.

3. Отобрали 100 (10 % от общего числа или 0,1-квантиль) сетей с наименьшим RMSE. Частоты попаданий в обучающую выборку вычислялись путем суммирования попаданий каждой точки пространства по выбранным сетям. Для включения в обучающее подмножество выбирались точки, частота попаданий которых в обучающее подмножество превышала 75 %.

2.6.3. Четырехшаговый (4step) алгоритм построения обучающего подмножества

Алгоритм построения обучающего подмножества 4step, подобно трехшаговому алгоритму, основан на делении по числу сочетаний без повторения, но отличается построением индивидуальной для каждой точки оценки репрезентативности по числу попаданий в обучающий набор.

1. Исходные данные 100 раз случайным образом разделили на обучающее и тестовое подмножества с соотношением 70 %/30 % соответственно. Обучающее подмножество включало 166 точек, тестовое 71 точку. Получили 100 разбиений на два непересекающихся подмножества.

2. Для каждого разделения построили десять моделей МЛП. Всего обучили 1 000 МЛП. Для каждого разделения выбирали один из десяти MLP с минимальной RMSE для оценки репрезентативности точки выборки. Таким образом, на 100 разделений было получено 100 МЛП, дли которых рассчитывали среднеквадратичные ошибки обучающих и тестовых подмножеств.

3. Каждой точке выборки присвоили набор наилучших сетей, в которых она участвовала в обучении, и построили гистограммы распределения RMSE для обучающего, тестового и общего множеств.

4. Для каждой точки выборки рассчитали базовые статистики RMSE обучающего, тестового и общего подмножества для сетей, в которых точка участвовала в обучении. Репрезентативность точки оценивалась путем сравнения средних значений и стандартных отклонений.

2.6.4. Алгоритм пространственного квотирования

Алгоритм пространственного квотирования (Space&max&min quote) основан на идее, что каждый участок изучаемой территории может быть представлен квотой отобранных на ней проб. Алгоритм включал следующие шаги.

1. Район съемки оконтуривали выпуклым многоугольником так, что геодезическая линия, проведенная между любыми двумя точками опробования, находилась внутри этого многоугольника. Оконтуривание выполнялось путем соединения граничных точек. Такая процедура удовлетворяет условию интерполяции, поскольку любой прогноз находится внутри этого многоугольника.

2. Полученный многоугольник разбивали на участки, включающие одинаковое количество наблюдений – точек опробования. В обучающее подмножество обязательно включали граничные точки полигона, максимальные и минимальные значения из каждой пространственной квоты. Количество пространственных квот полагали не менее четырех. Это согласуется с выбранными географическими направлениями. При наличии ярко выраженных географических особенностей (например, резкое изменение растительности, наличие ярко выраженного рельефа и т. п.) эту зону выделяли в отдельный пространственный участок. Объем пространственной квоты ограничен снизу необходимым количеством точек тестового подмножества. Из общих статистических соображений желательно, чтобы это подмножество содержало не менее 30 точек для надежной оценки точности интерполятора. Если объем данных позволяет, то количество пространственных участков можно увеличить.

3. Из каждого участка случайным образом добирали для обучающего подмножества точки так, чтобы их доля составляла 70 %.

Окончательное обучающее подмножество состояло из граничных точек, максимальных и минимальных значений из каждого участка и случайных добавлений из каждого участка до 70 % от общего числа точек. Оставшиеся 30 % точек составляли тестовое подмножество.

2.6.5. Алгоритм пространственно-вероятностного квотирования

Алгоритм пространственно-вероятностного квотирования (Space&quartile quote) повторял пространственное квотирование исходных данных, но с учетом возможного изменения значений моделируемой переменной.

1 и 2 шаги одинаковы из алгоритма Space&max&min.

3. После второго шага пространственного квотирования исходных данных оставшиеся значения моделируемой переменной из каждого участка были разбиты на вероятностные квоты (в обучающую выборку попали граничные точки, максимальное и минимальное значения). Желательно, чтобы количество вероятностных квот было не менее четырех для минимально необходимой репрезентативности функции плотности вероятности (левый, правый хвост и левый, правый центр).

4. 70 % значений из каждой квоты вероятностей были случайным образом выбраны для обучающего подмножества. Возможны варианты, в которых граничные точки и максимальное и минимальное значения могут совпадать. В этих случаях в обучающее подмножество попадает большее количество «случайных» значений. В тестовую выборку вошли оставшиеся 30 % точек. В общем случае количество вероятностных квот не может превышать количество тестовых случаев, т. е. каждая вероятностная квота должна быть представлена хотя бы одной точкой.

3. Результаты

Описательная статистика содержания хрома представлена в табл. 1. Гистограмма распределения содержаний хрома на выбранной территории приведена на рис. 3. Общее содержание Cr на городском фоне находится в пределах от 17 до 140 мг/кг, что не превышает кларк (по Уралу) [2, 27]. Известно, что общее содержание Cr в подзолах в почвах мира находится в пределах от 2,6 до 200 мг/кг [28]. Распределение хрома скошено вправо, с правым хвостом. Средние значения больше, чем (справа) медианы. Тест Шапиро – Уилка показал, что распределение отличается от нормального.

Таблица 1. Характеристика распределения хрома в верхнем слое почвы г. Ноябрьска

Содержание, мг/кг					Коэффициент вариации, %	Асимметрия	Эксцесс	p-значение Шапиро – Уилка
Минимум	Максимум	Среднее	СО*)	Медиана				
17	140	63	23	60	37	0,8	0,6	0,00

СО*) – стандартное отклонение.

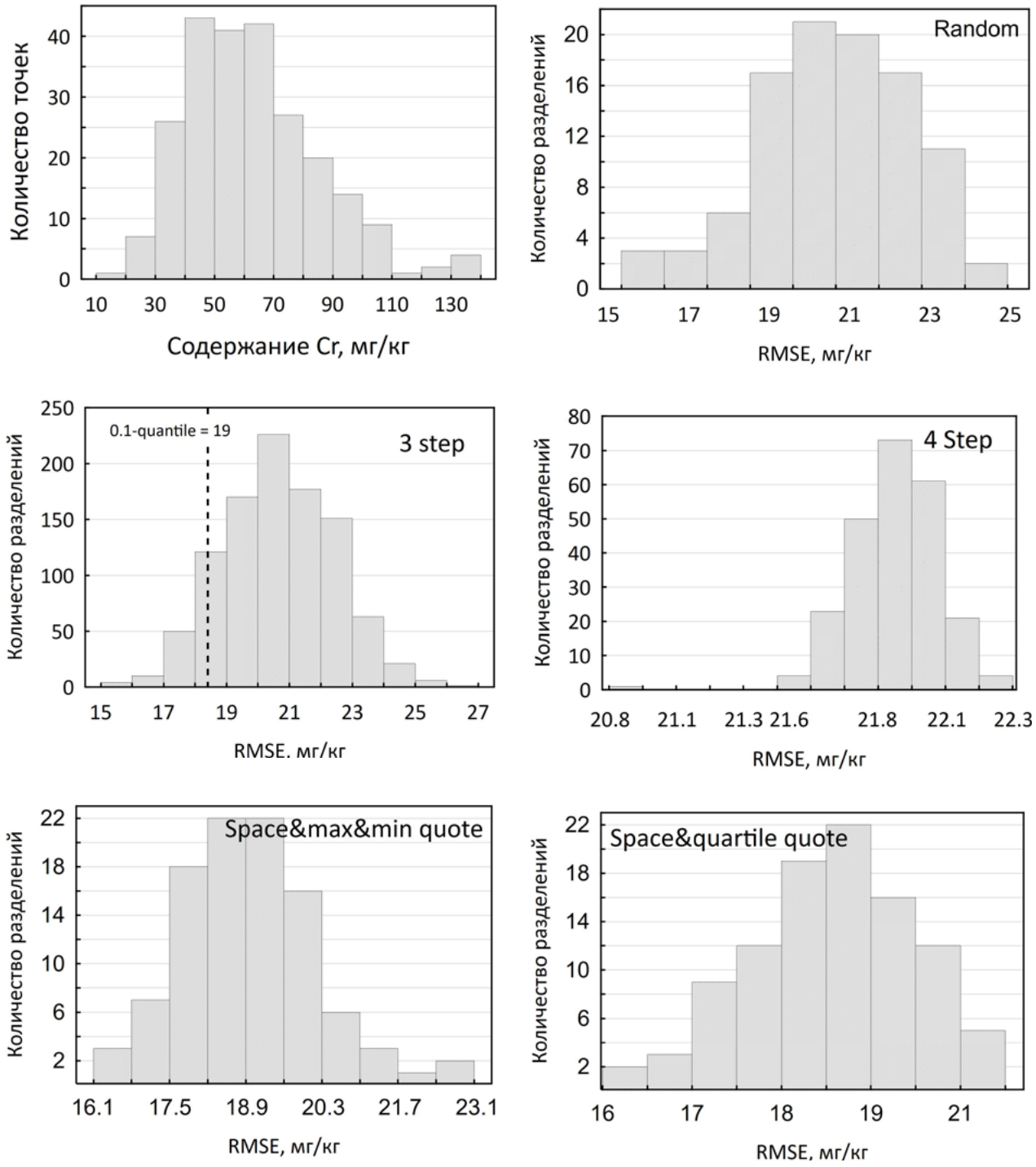

Рис. 3. Распределение исходных данных распределения признака и RMSE для разных алгоритмов разделения данных на обучающее и тестовое подмножества

В табл. 2 представлены статистические характеристики RMSE для разных способов разделения данных. На рис. 2 приведены гистограммы распределений RMSE для разных способов выбора обучающего подмножества. На рис. 3 приведено сравнение разных способов выбора обучающего множества по средним значениям точности модели RMSE. Наилучший результат по среднему значению RMSE показывает алгоритм пространственно-вероятностного квотирования. Интерес представляет четырехшаговый алгоритм построения обучающего подмножества, при самой высокой средней ошибке из представленных, распределение RMSE отличается от нормального и имеет самые низкие стандартное отклонение и коэффициент вариации.

Таблица 2. Характеристика распределения RMSE для разных алгоритмов

Алгоритм	RMSE, мг/кг							Коэффициент вариации, %	Асимметрия	Эксцесс
	Среднее	Медиана	Минимум	Максимум	СО	Среднее – СО	Среднее + СО			
Random	20,6	20,6	15,3	24,9	1,9	18,7	22,5	9,2	– 0,36	– 0,05
3step	20,6	20,6	15,4	26,5	1,8	18,8	22,4	8,5	0,04	– 0,21
4step	**21,9**	**21,9**	**20,8**	**22,3**	**0,2**	**21,7**	**22,1**	**0,8**	**– 1,01**	**6,26**
Space& max& min quote	18,9	18,9	16,1	23,0	1,3	17,6	20,2	6,7	0,59	0,95
Space& quartile quote	**18,8**	**18,9**	**15,8**	**21,1**	**1,1**	**17,7**	**19,9**	**5,8**	**– 0,34**	**– 0,19**

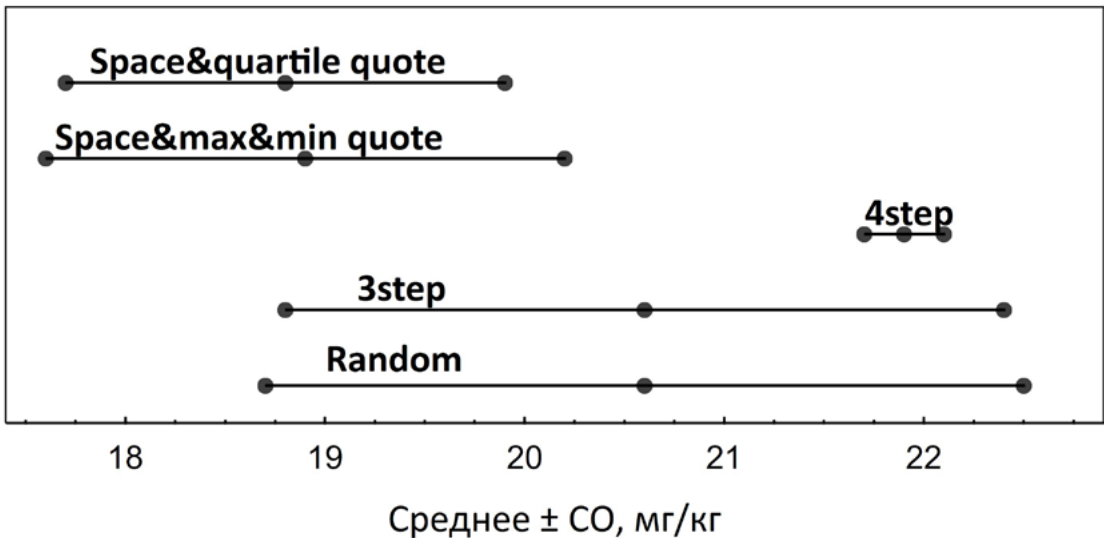

Рис. 4. Вариативность среднего для разных алгоритмов

Алгоритмы Space&quartile quote и Space&max&min quote относятся к детерминированным с элементом случайности, поскольку часть точек обучающего подмножества зафиксирована. 3step и 4step алгоритмы основаны на разделении исходных данных только случайным образом.

На рис. 5 представлена схема расположения точек, включение которых в обучающее подмножество обеспечивает для разных алгоритмов наименьшую RMSE. На рис. 5 представлены три списка точек для обучения: для детерминированного (Space quote), 3step и 4step алгоритмов. Для детерминированных Space&quartile

quote и Space&max&min quote способов обязательные для включения в обучающее множество точки совпадают по условиям построения алгоритмов. Наилучшие для обучения точки для 3step и 4step алгоритмов отличаются. Точки обучающего подмножества для разных алгоритмов не являются общими.

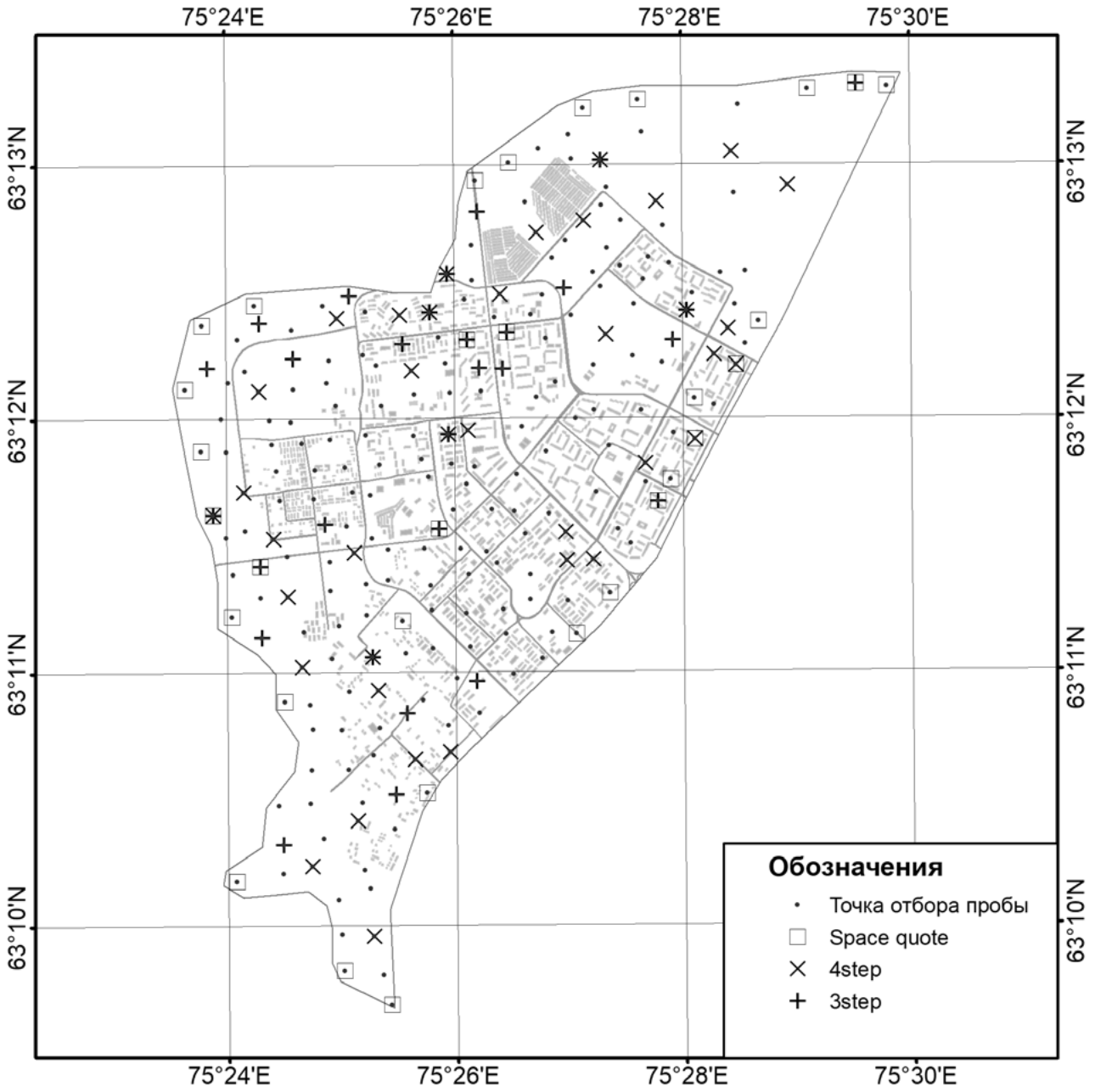

Рис. 5. Расположение репрезентативных точек обучающего подмножества для разных алгоритмов разделения данных

4. Обсуждение

Модель МЛП, использующая контролируемые типы разбиения, оказалась более точной, чем МЛП, обученная на случайном разбиении. Алгоритм Space&quartile quote показал наиболее точные результаты. Относительно небольшое количество разных точек в этих обучающих подмножествах может объяснить небольшую разницу между двумя детерминированными типами разделения. Модели Space&quartile и Space&max&min с контролируемым или детерминированным разделением более

точны. При этом распределение ошибок RMSE этих моделей у́же, чем в моделях, основанных на случайном разбиении. МЛП с пространственно-вероятностным квотированием исходных данных с учетом разброса значений Space&quartile оказался несколько более точным, чем МЛП только с пространственным квотированием Space&max&min. Разница между алгоритмами Space&quartile и Space&max&min была незначительной (менее 5 %). Для относительно небольших выборок процедура разделения на обучающее и тестовое подмножества может быть практически полностью определена. Это упрощает расчет модели при неуменьшении точности модели. Для таких детерминированных способов разделения одна реализация будет более точной (или сравнимой по точности), чем лучшая из сотен сетей, обучавшихся на случайно разделенной выборке. Алгоритмы, построенные на многократном случайном разделении, обеспечивают более высокую среднюю RMSE при меньшем отклонении.

5. Выводы

1. Точность модели ИНС для предсказания пространственного распределения признака в экологических задачах зависит от способа разделения исходных данных на обучающее и тестовое подмножества.
2. Детерминированные алгоритмы, основанные на квотировании исходных данных при разделении на обучающее и тестовое подмножества, обеспечивают меньшую среднюю ошибку модели, но разброс этой ошибки будет составлять 5–7 %.

6. Список литературы

1. *Demyanov*, *V*. A special issue on data science for geosciences / V. Demyanov, E. Gloaguen, M. Kanevski // Math. Geosci. – 2020. – Vol. 52. – P. 1–3. https://doi.org/10.1007/s11004-019-09846-0
2. Геохимия окружающей среды / Ю. Е. Сает, Б. А. Ревич, Е. П. Янин [и др.]. – М. : Недра, 1990. – 335 с. – ISBN 5-247-01127-9. – EDN XDXBQN.
3. A geochemical survey of heavy metals in agricultural and background soils of the Isfahan industrial zone Iran / A. Esmaeili, F. Moore, B. Keshavarzi [et al.] // Catena. – 2017. – Vol. 121. – P. 88–98.
4. Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China / G. H. Guo, F. Wu, F. Xie [et al.] // J. Env. Sci. – 2012. – Vol. 24(3). – P. 410–418.
5. High variation topsoil pollution forecasting in the Russian Subarctic: Using artificial neural networks combined with residual kriging / D. A. Tarasov, A. G. Buevich, A. P. Sergeev [et al.] // Applied Geochemistry. – 2018. – Vol. 88. – P. 188–197. – DOI 10.1016/j.apgeochem.2017.07.007. – EDN XXSNEL.
6. *Zhong*, *L*. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks / X. Guo, Zh. Xu, M. Ding // Geoderma. – 2021. – Vol. 402. – P. 115366. https://doi.org/10.1016/j.geoderma.2021.115366
7. Nonlinear variable selection for artificial neural networks using particle mutual information / R. J. May, H. R. Maier, G. D. Dandy [et al.] // Environmental Modelling & Software. – 2008. – Vol. 23 (10–11). – P. 1312–1326.
8. *Shaker*, *R*. *R*. Exploring non-linear relationships between landscape and aquatic ecological condition in southern Wisconsin: A GWR and ANN approach / R. R. Shaker, T. J. Ehlinger // International J. of Applied Geospatial Research. – 2014. – Vol. 5(4). – P. 1–20.

9. *Kanevski*, *M*. Mapping of soil contamination by using artificial neural networks and multivariate geostatistics / M. Kanevski, V. Demyanov, M. Maignan // Lecture Notes in Computer Science. – 1997. – Vol. 1327. – P. 1125–1130. – DOI 10.1007/bfb0020304. – EDN WQHJZF.

10. The Effect of Splitting of Raw Data into Training and Test Subsets on the Accuracy of Predicting Spatial Distribution by a Multilayer Perceptron / E. M. Baglaeva, A. P. Sergeev, A. V. Shichkin [et al.] // Mathematical Geosciences. – 2020. – Vol. 52. – No 1. – P. 111–121. – DOI 10.1007/s11004-019-09813-9. – EDN HBCZYX.

11. Почвы и ноосфера: Монография. Научное электронное издание / Е. Р. Абрамова, И. А. Аднан, А. В. Базаров [и др.]; Министерство науки и высшего образования РФ, Дальневосточный федеральный университет, Российская академия наук, Дальневосточное отделение, ФНЦ биоразнообразия наземной биоты Восточной Азии, Общество почвоведов им. В. В. Докучаева, Дальневосточное отделение общества почвоведов им. В. В. Докучаева, Far Eastern Climate Smart Lab. – Владивосток : Дальневосточный федеральный университет, 2019. – 409 с. – ISBN 978-5-7444-4707-6. – EDN LMCWKD.

12. A back propagation neural network model optimized by mind evolutionary algorithm for estimating Cd, Cr, and Pb concentrations in soils using Vis-NIR diffuse reflectance spectroscopy / X. Wang, Sh. An, Y. Xu [et al.] // Appl. Sci. – 2020. – Vol. 10(51). – P. 1–17. https:// doi:10.3390/app10010051.

13. Coordinate Transformation between Global and Local Data Based on Artificial Neural Network with K-Fold Cross-Validation in Ghana / Y. Y. Ziggah, H. Youjian, A. R. Tierra [et al.] // Earth Sciences Research J. – 2019. – Vol. 23(1). – PP. 67–77. https://doi.org/10.15446/esrj.v23n1.63860.

14. *Nath*, *A*. 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 / A. Nath, K. Subbiah // Neurocomputing. – 2018. – Vol. 272. – P. 294–305. http://dx.doi.org/10.1016/j.neucom.2017.07.004.

15. *Челахов*, *В*. *М*. Метод кортежирования обучающей выборки для повышения точности нейросетевой аппроксимации данных / В. М. Челахов, К. В. Деркачев // Современные проблемы науки и образования. – 2012. – № 6. – С. 107. – EDN TODLRF.

16. *Бурый*, *Я*. *А*. Экстраполирующее обучение нейронных сетей / Я. А. Бурый, Д. И. Самаль // Информатика. – 2019. – Т. 16, № 1. – С. 86–92. – EDN PQWKJT.

17. *Донской*, *В*. *И*. Извлечение оптимизационных моделей из данных: применение нейронных сетей / В. И. Донской // Таврический вестник информатики и математики. – 2018. – № 2(39). – С. 71–89. – EDN YPPSLR.

18. *Пастухов*, *А*. *А*. Применение самоорганизующихся карт Кохонена для формирования представительской выборки при обучении многослойного персептрона / А. А. Пастухов, А. А. Прокофьев // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Физико-математические науки. – 2016. – № 2(242). – С. 95–107. – DOI 10.5862/JPM.242.11. – EDN WAEDIH.

19. *Malof*, *J*. *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 / J. M. Malof, D. Reichman, L. M. Collins // Proceedings Volume 10628, Detection and Sensing of Mines, Explosive Objects, and Obscured Targets. – 2018. – Vol. XXIII. – 106280C. https://doi.org/10.1117/12.2305793.

20. *Kramm*, *T*. Assessing the influence of environmental factors and datasets on soil type prediction with two machine learning algorithms in a heterogeneous area in the Rur catch-

ment, Germany / T. Kramm, D. Hoffmeister // Geoderma Regional. – 2020. – Vol. 22. – e00316. https://doi.org/10.1016/j.geodrs.2020.e00316.

21. *Fernandez Jaramillo*, *J. M*. Sample selection via angular distance in the space of the arguments of an artificial neural network / J. M. Fernandez Jaramillo, R. Mayerle // Computers and Geosciences. – 2018. – Vol. 114. – P. 98–106. https://doi.org/10.1016/j.cageo.2018.02.003.
22. *Sakizadeh*, *M*. Support vector machine and artificial neural network to model soil pollution: a case study in Semnan Province, Iran / M. Sakizadeh, R. Mirzaei, H. Ghorbani // Neural. Comput. & Applic. – 2017. – Vol. 28. – P. 3229–3238. https://ezproxy.urfu.ru:3055/10.1007/s00521-016-2231-x.
23. Forecasting the spatiotemporal variability of soil CO2 emissions in sugarcane areas in southeastern Brazil using artificial neural networks / L. P. S. Freitas, M. L. M. Lopes, L. B. Carvalho [et al.] // Environ. Monit. Assess. – 2018. – Vol. 190. – P. 741. https://doi.org/10.1007/s10661-018-7118-0
24. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals / A. P. Sergeev, A. G. Buevich, E. M. Baglaeva [et al.] // Catena. – 2019. – Vol. 174. – P. 425–435. – DOI 10.1016/j.catena.2018.11.037. – EDN AZBEVO.
25. *Добровольский*, *Г. В*. География почв / Г. В. Добровольский, И. С. Урусевская. – М. : Изд-во Московского университета, 1984. – 416 с. – EDN TSHOUB.
26. Clean Outlier Data of Matlab. www.mathworks.com/help/matlab/ref/cleanoutlierdata.html
27. Краткий справочник по геохимии / Г. В. Войткевич, А. Е. Мирошников, А. С. Поваренных [и др.]. – М. : Недра, 1977. – 180 с.
28. *Kabata-Pendias*, *A*. Trace elements in soils and plants / A. Kabata-Pendias. – Taylor and Francis Group CRC Press, 2011. https://doi.org/10.1201/b10158.

Сведения об авторах:

Баглаева Елена Михайловна, канд. ф.-м. н., старший научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия. Эл. почта: elenbaglaeva@gmail.com.

Сергеев Александр Петрович, канд. ф.-м. н., заведующий Лаборатории физики и экологии Института промышленной экологии УрО РАН, г. Екатеринбург, Россия.

Шичкин Андрей Васильевич, научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия

Буевич Александр Геннадьевич, научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия

Субботина Ирина Евгеньевна, канд. ф.-м. н., научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия

Буторова Анастасия Сергеевна, инженер-исследователь Института промышленной экологии УрО РАН, г. Екатеринбург, Россия

COMPARATIVE ANALYSIS OF DATA SPLITTING ALGORITHMS FOR TRAINING ARTIFICIAL NEURAL NETWORKS IN MODELING THE SPATIAL DISTRIBUTION OF HEAVY METALS IN THE TOPSOIL

E. M. Baglaeva, A. P. Sergeev, A. V. Shichkin, A. G. Buevich, I. E. Subbotina, A. S. Butorova

Institute of Industrial Ecology, Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia

A comparison of five algorithms for selecting a training subset for artificial neural networks to interpolate the spatial distribution of a feature according to environmental components screenings is presented. The initial data were the contents of chromium in the topsoil in the residential areas of the city of Noyabrsk (YNAO, Russia). The spatial distributions of the element content in the topsoil were interpolated by a multilayer perceptron (MLP). Root mean squared error (RMSE) statistics were calculated as accuracy metrics for each algorithm. A smaller error is shown by the Space&quartile quote algorithm based on the spatial-probabilistic quoting of the initial data. Taking into account the information about the error statistics of the algorithms for splitting the original set, it is possible to improve the accuracy of the artificial neural network model in solving environmental problems.

Key words: MLP; splitting data; spatial distribution; environment modelling; representativeness.

References

1. *Demyanov*, *V*. A special issue on data science for geosciences / V. Demyanov, E. Gloaguen, M. Kanevski // Math. Geosci. – 2020. – Vol. 52. – P. 1–3. https://doi.org/10.1007/s11004-019-09846-0
2. Geochemistry of the envirenment / Yu. E. Saet, B. A. Revich, E. P. Yanin [et al.]. – Moscow : Nedra Publishing House, 1990. – 335 p. – ISBN 5-247-01127-9. – EDN XDXBQN.
3. *Esmaeili*, *A*. A geochemical survey of heavy metals in agricultural and background soils of the Isfahan industrial zone Iran / A. Esmaeili, F. Moore, B. Keshavarzi [et al.] // Catena. – 2017. – Vol. 121. – P. 88–98.
4. Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China / G. H. Guo, F. Wu, F. Xie [et al.] // J. Env. Sci. – 2012. – Vol. 24(3). – P. 410–418.
5. High variation topsoil pollution forecasting in the Russian Subarctic: Using artificial neural networks combined with residual kriging / D. A. Tarasov, A. G. Buevich, A. P. Sergeev [et al.] // Applied Geochemistry. – 2018. – Vol. 88. – P. 188–197. – DOI 10.1016/j.apgeochem.2017.07.007. – EDN XXSNEL.

6. *Zhong*, *L*. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks / X. Guo, Zh. Xu, M. Ding // Geoderma. – 2021. – Vol. 402. – P. 115366. https://doi.org/10.1016/j.geoderma.2021.115366
7. Nonlinear variable selection for artificial neural networks using particle mutual information / R. J. May, H. R. Maier, G. D. Dandy [et al.] // Environmental Modelling & Software. – 2008. – . Vol. 23 (10–11). – P. 1312–1326.
8. *Shaker*, *R*. *R*. Exploring non-linear relationships between landscape and aquatic ecological condition in southern Wisconsin: A GWR and ANN approach / R. R. Shaker, T. J. Ehlinger // International J. of Applied Geospatial Research. – 2014. – Vol. 5(4). – P. 1–20.
9. *Kanevski*, *M*. Mapping of soil contamination by using artificial neural networks and multivariate geostatistics / M. Kanevski, V. Demyanov, M. Maignan // Lecture Notes in Computer Science. – 1997. – Vol. 1327. – P. 1125–1130. – DOI 10.1007/bfb0020304. – EDN WQHJZF.
10. The Effect of Splitting of Raw Data into Training and Test Subsets on the Accuracy of Predicting Spatial Distribution by a Multilayer Perceptron / E. M. Baglaeva, A. P. Sergeev, A. V. Shichkin [et al.] // Mathematical Geosciences. – 2020. – Vol. 52. – No 1. – P. 111–121. – DOI 10.1007/s11004-019-09813-9. – EDN HBCZYX.
11. Soils and noosphere: Monograph. Scientific electronic edition / E. R. Abramova, I. A. Adnan, A. V. Bazarov [et. al.]; Ministry of Science and Higher Education of the Russian Federation, Far Eastern Federal University, Russian Academy of Sciences, Far Eastern Branch, Federal Scientific Center for Biodiversity of Terrestrial Biota of East Asia, Society of Soil Scientists named after. V. V. Dokuchaev, Far Eastern Branch of the Society of Soil Scientists named after A. I. V. V. Dokuchaeva, Far Eastern Climate Smart Lab. – Vladivostok: Far Eastern Federal University, 2019. – 409 p. – ISBN 978-5-7444-4707-6. – EDN LMCWKD.
12. A back propagation neural network model optimized by mind evolutionary algorithm for estimating Cd, Cr, and Pb concentrations in soils using Vis-NIR diffuse reflectance spectroscopy / X. Wang, Sh. An, Y. Xu [et al.] // Appl. Sci. – 2020. – Vol. 10(51). – P. 1–17. https:// doi:10.3390/app10010051.
13. Coordinate Transformation between Global and Local Data Based on Artificial Neural Network with K-Fold Cross-Validation in Ghana / Y. Y. Ziggah, H. Youjian, A. R. Tierra [et al.] // Earth Sciences Research J. – 2019. – Vol. 23(1). – PP. 67–77. https://doi.org/10.15446/esrj.v23n1.63860.
14. *Nath*, *A*. 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 / A. Nath, K. Subbiah // Neurocomputing. – 2018. – Vol. 272. – P. 294–305. http://dx.doi.org/10.1016/j.neucom.2017.07.004.
15. *Chelakhov*, *V*. *M*. Method of training sample tuple to improve the accuracy of neural network approximation of data / V. M. Chelakhov, K. V. Derkachev // Modern problems of science and education. – 2012. – No. 6. – P. 107. – EDN TODLRF.
16. *Bury*, *Ya*. *A*. Extrapolating training of neural networks / Ya. A. Bury, D. I. Samal // Informatics. – 2019. – T. 16, No. 1. – S. 86–92. – EDN PQWKJT.
17. *Donskoy*, *V*. *I*. Extraction of optimization models from data: the use of neural networks / V. I. Donskoy // Tauride Bulletin of Informatics and Mathematics. – 2018. – No. 2 (39). – S. 71–89. – EDN YPPSLR.
18. *Pastukhov*, *A*. *A*. The use of self-organizing Kohonen maps for the formation of a representative sample when training a multilayer perceptron / A. A. Pastukhov, A. A. Prokofiev // Scientific and technical statements of the St. Petersburg State Polytechnic University.

Physical and mathematical sciences. – 2016, No. 2 (242). – P. 95–107. – DOI 10.5862/JPM.242.11. – EDN WAEDIH.

19. *Malof*, *J. 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 / J. M. Malof, D. Reichman, L. M. Collins // Proceedings Volume 10628, Detection and Sensing of Mines, Explosive Objects, and Obscured Targets. – 2018. – Vol. XXIII. – 106280C. https://doi.org/10.1117/12.2305793.
20. *Kramm*, *T*. Assessing the influence of environmental factors and datasets on soil type prediction with two machine learning algorithms in a heterogeneous area in the Rur catchment, Germany / T. Kramm, D. Hoffmeister // Geoderma Regional. – 2020. – Vol. 22. – e00316. https://doi.org/10.1016/j.geodrs.2020.e00316.
21. *Fernandez Jaramillo*, *J. M*. Sample selection via angular distance in the space of the arguments of an artificial neural network / J. M. Fernandez Jaramillo, R. Mayerle // Computers and Geosciences. – 2018. – Vol. 114. – P. 98–106. https://doi.org/10.1016/j.cageo.2018.02.003.
22. *Sakizadeh*, *M*. Support vector machine and artificial neural network to model soil pollution: a case study in Semnan Province, Iran / M. Sakizadeh, R. Mirzaei, H. Ghorbani // Neural. Comput. & Applic. – 2017. – Vol. 28. – P. 3229–3238. https://ezproxy.urfu.ru:3055/10.1007/s00521-016-2231-x.
23. Forecasting the spatiotemporal variability of soil CO2 emissions in sugarcane areas in southeastern Brazil using artificial neural networks / L. P. S. Freitas, M. L. M. Lopes, L. B. Carvalho [et al.] // Environ. Monit. Assess. – 2018. – Vol. 190. – P. 741. https://doi.org/10.1007/s10661-018-7118-0
24. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals / A. P. Sergeev, A. G. Buevich, E. M. Baglaeva [et al.] // Catena. – 2019. – Vol. 174. – P. 425–435. – DOI 10.1016/j.catena.2018.11.037. – EDN AZBEVO.
25. *Dobrovolsky*, *G. V*. Geography of soils / G. V. Dobrovolsky, I. S. Urusevskaya. – M. : Publishing House of the Moscow State University, 1984. – 416 p. – EDN TSHOUB.
26. Clean Outlier Data of Matlab. www.mathworks.com/help/matlab/ref/cleanoutlierdata.html
27. Brief reference book on geochemistry / G. V. Voitkevich, A. E. Miroshnikov, A. S. Cookery [et al.]. – M. : Nedra, 1977. – 180 p.
28. *Kabata-Pendias*, *A*. Trace elements in soils and plants / A. Kabata-Pendias. – Taylor and Francis Group CRC Press, 2011. https://doi.org/10.1201/b10158.