

Геоинформатика. 2023. № 3. С. 63–70.
Geoinformatika. 2023;(3):63–70.

Геоэкология

Научная статья
 УДК 504.064.2.001.18
<https://doi.org/10.47148/1609-364X-2023-3-63-70>

Индивидуальная и парная репрезентативность точек пробоотбора в задачах интерполяции распределения тяжелых металлов в верхнем слое почвы

© 2023 г. — Е.М. Баглаева^{1, 2, а)}, А.П. Сергеев^{1, 2, б)}, А.В. Шичкин^{1, с)}, А.Г. Бувич^{1, 2, д)}, А.С. Буторова^{1, 2, е)}

¹Федеральное государственное бюджетное учреждение науки Институт промышленной экологии Уральского отделения Российской академии наук (ИПЭ УрО РАН); Россия, г. Екатеринбург

²Федеральное государственное автономное образовательное учреждение высшего образования «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина» (ФГАОУ ВО «УрФУ им. Б.Н. Ельцина»); Россия, г. Екатеринбург

^{а)}e.m.baglaeva@urfu.ru, ^{б)}sergeev@ecko.uran.ru, ^{с)}and@ecko.uran.ru, ^{д)}bag@ecko.uran.ru, ^{е)}a.s.butorova@urfu.ru

Аннотация: Статья посвящена проблеме выбора репрезентативного тренировочного подмножества для обучения искусственной нейронной сети в задачах интерполяции распределения тяжелых металлов в верхнем слое почвы. Экологические данные, часто используемые для построения моделей искусственных нейронных сетей, представляют наборы данных в нерегулярных точках. Традиционно разделение исходных данных на обучающее и тестовое подмножество происходит случайным образом, что приводит к ряду проблем. Для отбора в обучающее подмножество предложен способ оценки индивидуальной и коллективной репрезентативности точек, обеспечивающих достоверные данные о содержании элемента в почве на заданной территории. Наиболее представительными по индивидуальной репрезентативности оказались точки с максимальным содержанием элемента, их присутствие в обучающем подмножестве ИНС уменьшает ошибку и повышает корреляцию между результатами модельных расчетов и натуральных измерений на тестовом наборе. При оценке парной репрезентативности обнаружены эффекты синергии, при котором включение пары обеспечивает лучшую точность модели) и антисинергии (пара точек представляет меньше информации для описания распределения содержания элементов, чем отдельно точки, входящие в пару). Разные геопозиции сэмплинга несут разную информацию и имеют неодинаковую ценность для интерполяции распределения признака.

Ключевые слова: репрезентативность точек, сэмплинг, тяжелые металлы, искусственные нейронные сети, выбор обучающего подмножества

Для цитирования: Баглаева Е.М., Сергеев А.П., Шичкин А.В., Бувич А.Г., Буторова А.С. Индивидуальная и парная репрезентативность точек пробоотбора в задачах интерполяции распределения тяжелых металлов в верхнем слое почвы // Геоинформатика. — 2023. — № 3. — С. 63–70. <https://doi.org/10.47148/1609-364X-2023-3-63-70>.

Geoinformation system

Original article

Individual and pair representativeness of sampling points selection in interpolation tasks of the heavy metals distribution in the topsoil

© 2023 — E.M. Baglaeva^{1, 2, а)}, A.P. Sergeev^{1, 2, б)}, A.V. Shichkin^{1, с)}, A.G. Buevich^{1, 2, д)}, A.S. Butorova^{1, 2, е)}

¹Institute of Industrial Ecology UB RAS; Ekaterinburg, Russia

²Ural Federal University named after B.N. Yeltsin; Ekaterinburg, Russia

^{а)}e.m.baglaeva@urfu.ru, ^{б)}sergeev@ecko.uran.ru, ^{с)}and@ecko.uran.ru, ^{д)}bag@ecko.uran.ru, ^{е)}a.s.butorova@urfu.ru

Abstract: The article is devoted to the problem of choosing a representative selection of a subset for an artificial neural network in the tasks of interpolation of the distribution of metals in the topsoil. Environmental data, often used to build artificial neural network models, are datasets at irregular points. The traditional division of the input data into training and test subsets occurs randomly, which transfers to a number of problems. For selection in the training subset, the question of individual and collective representativeness of points is asked, sending them a request for data on the content of the element in the soil in a given area. The most representative in terms of individual representativeness arise with the maximum reference points, their presence in the training subset of the ANN measurement of error and an increase in the correlation between the results of model calculations and natural measurements on the test subset. When assessing the pairwise representativeness of the identified synergy effects, which, when included, achieve high model reliability) and anti-synergy (the parameters of using less information to describe the content of the elements than separately the points of view included in the pair). The various sampling locations have different information and unequal meaning for feature interpolation.

Key words: representativeness of points, sampling, heavy metals, artificial neural networks, choice of training subset

For citation: Baglaeva E.M., Sergeev A.P., Shichkin A.V., Buevich A.G., Butorova A.S. Individual and pair representativeness of sampling points selection in interpolation tasks of the heavy metals distribution in the topsoil. *Geoinformatika*. 2023;(3):63–70. <https://doi.org/10.47148/1609-364X-2023-3-63-70>. In Russ.

1. Введение

Вопросам репрезентативного отбора проб компонентов окружающей природной среды посвящено большое число публикаций [1, 4, 6, 14]. Задача оценки репрезентативности и построения репрезентативного множества возникает при организации сэмплинга для оценки качества компонентов окружающей среды, при статистической обработке экологических данных мониторингов и при выборе обучающего множества для искусственных нейронных сетей (ИНС), моделирующих пространственное распределение признака [3, 11, 15].

Существующая в России нормативная документация по сэмплингу и экологическому мониторингу для оценки степени загрязнения компонентов окружающей природной среды, в частности почвы, недостаточна для отражения истинной картины загрязнения, поэтому не может служить для принятия решений по сохранению и защите окружающей среды [2, 5, 8].

Мониторинг параметров окружающей среды в условиях городской застройки не способен обеспечить получение полных пространственно-временных характеристик загрязнения. Для комплексной оценки уровней загрязнения окружающей среды в городах [10, 16, 17, 18] часто сочетают мониторинг с другими методами получения данных, среди которых модели на основе ИНС. В научной литературе опубликованы работы, в которых выбор точек в обучающее подмножество происходит с использованием информации о распределении исследуемого признака, но выигрыш в точности модели оказывается небольшим. На точность прогноза моделей ИНС большое влияние оказывает выбор точек, используемых для обучения ИНС. Как показано в работах [9, 12, 13], разные точки вносят разный вклад в ошибку прогноза ИНС, т.е. имеют разную представительность — репрезентативность для целей прогноза.

Под репрезентативностью будем понимать характеристику отдельных точек изучаемой статистической совокупности адекватно отражать характеристику исследуемого признака. Репрезентативный отбор проб или репрезентативный выбор точек обучающего множества обеспечивает в пределах заданной точности достоверные данные о содержании загрязнителя в компоненте окружающей среды на выбранной территории в заданный момент времени.

Мы предполагаем, что не только точки отличаются разной представительностью для оценки признака, но и наборы точек (пары, тройки, ..., *n*-ки) обладают разной представительностью. В настоящей работе предложено рассматривать сравнение индивидуальной и коллективной репрезентативности при включении точек в тренировочный набор. Под индивидуальной репрезентативностью точки сэмплинга будем понимать частоту ее попаданий в тренировочное множество, обучение на котором обеспечивает наименьшую ошибку модели. Коллективная репрезентативность — это частота попаданий коллектива точек (пар, троек, четверок и т.д.) в тренировочное множество, обучение на котором обеспечивает наименьшую ошибку модели. Для построения репрезентативного обучающего набора необходимо 1) определить, какие *n*-ки наиболее репрезентативны (объем *n*-ки и уровень репрезентативности); 2) выявить связь между индивидуальной и коллективной репрезентативностью. Определение объема всех репрезентативных *n*-ок, требует больших вычислительных мощностей компьютера, поэтому коллективную репрезентативность в настоящей работе оценивали по парной.

2. Материалы и методы

2.1. География отбора проб

Почвенная съемка была организована в жилой зоне города Ноябрьска (N63.2°, E75.5°), Ямало-Ненецкий автономный округ, Россия (рис.1).

Рис. 1. Схема отбора (слева — Ноябрьск, центр — содержание Cu, справа — содержание Ni)

Fig. 1. Sampling scheme (left — Noyabrsk, center — Cu content, right — Ni content)

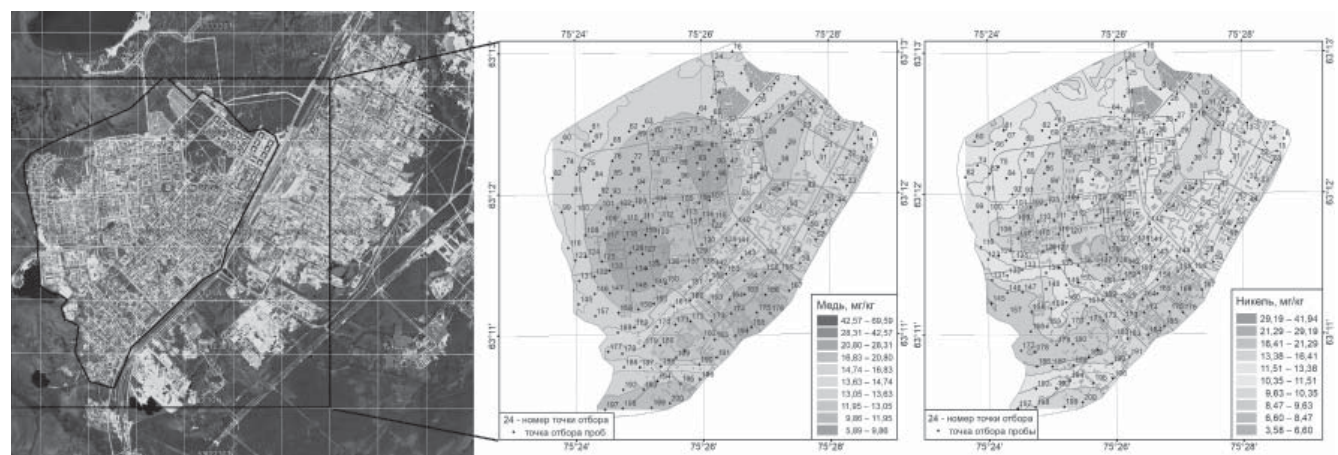


Схема отбора проб представлена на рисунке 1. Для моделирования использовали 200 точек, расположенных в центральной части жилой застройки города. Промышленность города связана с углеводородной энергетикой. Климат резко континентальный. Город находится в природной зоне тайги. Преобладающий тип почвы — таежные подзолы [7].

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 описана ранее [5]. На рисунке 1 показана схема пространственного отбора проб. Каждый образец упаковывали в двойной полиэтиленовый пакет с идентификатором образца. Идентификатор отмечал район, географические координаты и количество 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 [4]. Первый этап представлял собой сушку на воздухе при температуре 20°C и атмосферном давлении. 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 ($\text{HNO}_3 + \text{HF} + \text{HClO}_4$) 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 меди и никеля.

2.3. Подготовка данных для моделирования

Географические координаты были входными данными для моделирования, а содержание элемента — выходными.

2.4. Многослойный перцептрон

Для моделирования пространственного распределения признака использовали многослойный перцептрон как наиболее простую модель для демонстрации возможностей метода. Построение

модели многослойного перцептрона (МЛП) заключалось в выборе количества нейронов внутри скрытого слоя. Структуру МЛП выбирали, предварительно проведя несколько циклов обучения и оценив ошибку для тестового набора данных. 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 обучающем подмножестве (алгоритм обучения Левенберга-Марквардта). Мы использовали тангенциальную функцию активации. Структура представляла однослойный перцептрон с одним входным слоем, состоящим из двух нейронов (пространственные координаты x и y), скрытым слоем из 9 нейронов и одним выходным слоем, включающим один нейрон (содержание элемента)

2.5. Определение репрезентативности

Под репрезентативностью точки пробоотбора будем понимать способность этой точки обеспечивать малую среднеквадратическую ошибку $RMSE$ (1) для оценки точности и высокий коэффициент корреляции $Corr$ (2) для проверки синхронности изменений между прогнозируемыми и наблюдаемыми значениями при участии этой точки в обучении.

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

где $p(i)$ — прогнозируемые данные, $o(i)$ — наблюдаемые данные, n — количество точек. $RMSE$ (1) использовали для проверки точности прогнозирования между прогнозируемыми и наблюдаемыми значениями параметра.

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

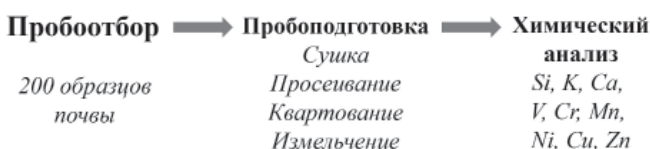
где \bar{p} — прогнозируемое среднее, \bar{o} — наблюдаемое среднее. Коэффициент корреляции показывает статистическую взаимосвязь прогнозируемых значений и наблюдаемых, насколько изменения прогнозируемых значений повторяют систематические изменения наблюдаемых.

Индивидуальная репрезентативность для каждой точки учитывает набор наилучших (малые $RMSE$ (1) и высокий $Corr$ (2)) сетей, в которых точка участвовала в обучении.

Коллективная репрезентативность — это репрезентативность точки, учитывающая связи с окружением, позволяющие обеспечивать малую $RMSE$ и высокий $Corr$ при участии в обучении коллектива точек. Коллективная репрезентативность учитывает комбинации точек, участвующие в обучении, может быть парная, тройная и т.п.

Рис. 2. Этапы пробоподготовки и химического анализа

Fig. 2. Sample preparation and chemical analysis steps



2.5.1. Оценка индивидуальной репрезентативности

Для оценки репрезентативности точек, участвующих в обучении, использовали четырехшаговый (4step) алгоритм разделения данных на тренировочное и тестовое подмножества.

Исходные данные многократно делили случайным образом на обучающее и тестовое подмножества в соотношении 75%/25% соответственно. 200 точек случайным образом делили на 150 тренировочных и 50 тестовых. Количество способов разделения равно числу сочетаний без повторения. Для разделения 200 точек на 150 и 50 число способов разделения составит 19900. Алгоритм разделения исходных данных состоит из четырех шагов.

1. Набор исходных данных 19900 раз случайным образом разделили на два непересекающихся множества, обучающую и тестовую выборки в соотношении 75%/25%, соответственно. Таким образом, получили 19900 обучающих и 19900 тестовых подмножеств.

2. Для каждого случайного разделения построили 5 сетей. Для каждой обученной сети определяли среднеквадратические ошибки (*RMSE*) прогноза тренировочного и тестового наборов и выбирали сеть с минимальной *RMSE*.

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

4. Для каждой точки выборки рассчитали базовые статистики *RMSE* и коэффициента корреляции обучающего, тестового и общего подмножества для сетей, в которых точка участвовала в обучении.

Индивидуальная репрезентативность точки оценивалась путем сравнения средних значений *RMSE* и стандартных отклонений. Наилучшей репрезентативностью обладает точка, включение которой в обучающий набор обеспечивает меньшее среднее *RMSE* и больший средний коэффициент корреляции с наблюдаемыми значениями.

2.5.2. Оценка коллективной репрезентативности

При оценке коллективной репрезентативности точек обучающего подмножества также использовали четырехшаговый (4step) алгоритм. Использовали те же разделения на обучающее и тестовое подмножества в соотношении 75%/25%, соответственно, результаты обучения как набор наилучших сетей для каждой точки, и соответствующие им распределения *RMSE* и коэффициентов корреляции для обучающего, тестового и общего множеств.

Парную репрезентативность оценивали по выборкам двух точек из двухсот. Для каждой пары точек рассчитывали базовые статистики *RMSE* обучающего, тестового и общего подмножества для сетей, в которых обе точки участвовали в обучении. Из этих пар отобрали 1000 точек с наименьшими средними *RMSE* для пары, попавшей в тренировочный набор, что соответствовало квантили 0,051. Репрезентативность точек оценивали по количеству ее связей с другими точками в пределах квантиля 0,051 (1000 точек) по среднему *RMSE* или среднему коэффициенту корреляции. Построили графы для порога отсечки в 10 связей с другими точками.

В данной работе из-за вычислительных сложностей ограничились парами точек. Гипотезу существования эффекта синергии проверяли сравнением индивидуальной и парной репрезентативностей. Для проверки использовали условное распределение средних *RMSE* и коэффициентов корреляции (при условии, что пара точек попала в наилучший обучающий набор).

3. Результаты

3.1. Описательная статистика моделируемых признаков

В таблице 1 представлена описательная статистика содержаний меди и никеля на исследуемой территории. Распределение содержания меди и никеля в верхнем слое почвы г. Ноябрьска приведено на рисунке 1 (в центре — Cu, справа — Ni).

Общее содержание меди на городском фоне находится в пределах от 5,89 до 69,59 мг/кг, никеля

Табл. 1. Распределения меди и никеля в верхнем слое почвы Ноябрьска
Tab. 1. Cuprum and Niccolum distributions in the topsoil of Noyabrsk

Элемент	Содержание, мг/кг					Коэффициент вариации, %	Асимметрия	Экссесс
	Минимум	Максимум	Среднее	CO*)	Медиана			
Cu	5,89	69,59	16,12	7,64	14,67	47	2,69	13
Ni	3,58	41,94	11,67	4,50	11,15	39	2,16	11

*) CO — стандартное отклонение

от 3,58 до 41,94 мг/кг, что не превышает кларковые значения [16, 9]. Из 200 проб не более чем в 8 (4% проб) обнаружены превышения ориентировочно-допустимых концентраций меди (33 мг/кг для песчаных почв) и никеля (20 мг/кг для песчаных почв) [9].

3.2. Определение репрезентативности

Для каждого разделения данных были построены модели MLP и рассчитаны характеристики репрезентативности: *RMSE* и коэффициенты корреляции. В таблице 2 представлены статистические характеристики оценки репрезентативности.

На рисунке 3 приведены полученные для 19 900 моделей оценки средних *RMSE* и коэффициентов корреляции. Как видно из рисунка 2, распределение *RMSE* разбивается на два кластера по меди и никелю. Кластер с меньшими *RMSE* связан с включением в тренировочный набор единственной

точки 129 для никеля (точка с максимальным содержанием никеля) и 134 для меди (точка с максимальным содержанием меди).

4. Обсуждение результатов

Индивидуальной репрезентативности недостаточно для определения наилучшего обучающего набора. Точка, обладающая низкой индивидуальной репрезентативностью, в парах может быть «хороша» для обучения. По аналогии с индивидуальными характеристиками могут быть пары «хорошие» для обучения (это все 199 пар со 129 точкой для никеля, и 134 точкой для меди), могут быть «плохие», которые повышают ошибку модели при включении их в пары в тренировочный набор. Результаты приведены в таблице 3. Выделены по девять точек: 8, 12, 14, 63, 67, 116, 165, 168, 199 для меди и 49, 52, 84, 102, 103, 104, 118, 146, 192 для никеля. Элементное содержание в этих точках ниже или близко к среднему.

Парная репрезентативность не всегда сводится к индивидуальной. Существует эффект синергии, т.е. учет коллективной (*n*-ной) репрезентативности позволяет уменьшить ошибку модели. Парная (коллективная) репрезентативность характеризует взаимодействие пар точек, т.е. способность пар точек при включении в тренировочный набор обеспечить характеристики *RMSE* и коэффициентов корреляции, превосходящие лучшие индивидуальные характеристики. Ценность, добавляемая парой {94; 129} для никеля и парой {134; 135} для меди выше ценности, вносимой отдельными точками {94} и {129} для никеля и {134} и {135} для меди. Этот эффект синергии создается благодаря связям между точками. В таблице 3 приведены наилучшие пары с наименьшим *RMSE* и наибольшим коэффициентом корреляции и наихудшие с наибольшим *RMSE* и наименьшим коэффициентом корреляции для меди и никеля. Наряду с эффектом синергии существуют взаимосвязи, которые снижают ценность модели. В этом случае пары точек представляют меньше информации для описания распределения содержания элементов, чем отдельно точки, входящие в пару — эффект антисинергии. Это может

Рис. 3. Распределения *RMSE* и коэффициентов корреляции

Fig. 3. Distributions of *RMSE* and correlation coefficients

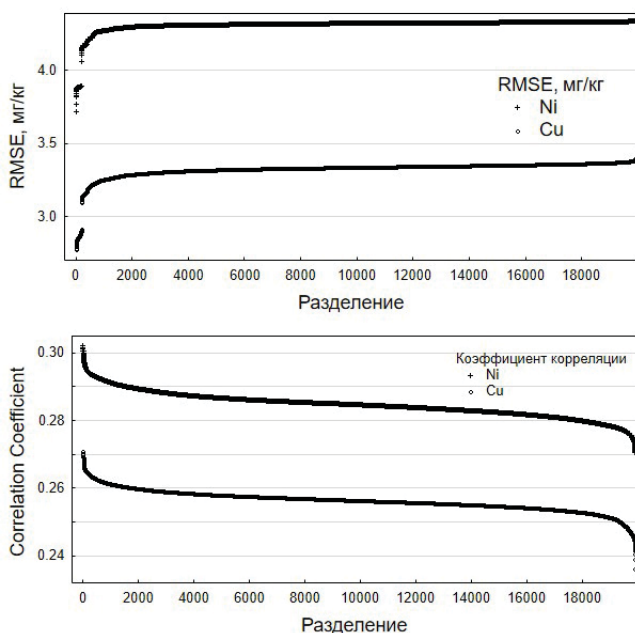


Табл. 2. Характеристики распределения средних *RMSE* и коэффициентов корреляции

Tab. 2. Descriptive statistics of mean *RMSE* and correlation coefficients

Элемент	Характеристика	Среднее	Медиана	Минимум	Максимум	CO	Коэффициент вариации, %	Асимметрия	Эксцесс
Cu	<i>RMSE</i> , mg/kg	3,325	3,336	2,771	3,395	0,059	2	-5,48	39
	<i>Corr</i>	0,256	0,256	0,236	0,271	0,003	1	-0,27	2
Ni	<i>RMSE</i> , mg/kg	4,312	4,322	3,720	4,340	0,049	1	-7,18	58
	<i>Corr</i>	0,285	0,285	0,271	0,302	0,004	1	0,19	1

Табл. 3. Сравнение индивидуальных и парных характеристик репрезентативности

Tab. 3. Comparison of individual and pair representativeness characteristics

Сравнение	«Хорошая» пара по RMSE			«Плохая» пара по RMSE			«Хорошая» пара по коэффициенту корреляции			«Плохая» пара по коэффициенту корреляции		
Элемент	Ni											
<i>n</i>	{94;129}	{94}	{129}	{19;56}	{19}	{56}	{13;25}	{13}	{25}	{88;145}	{88}	{145}
<i>RMSE</i> , мг/кг	3,720	4,167	3,880	4,340	4,325	4,326	4,234	4,266	4,281	4,155	4,166	4,304
<i>Corr</i>	0,281	0,278	0,287	0,284	0,284	0,284	0,302	0,294	0,293	0,271	0,278	0,278
Элемент	Cu											
<i>n</i>	{134;135}	{134}	{135}	{139;165}	{139}	{165}	{22;135}	{22}	{135}	{126;127}	{126}	{127}
<i>RMSE</i> , мг/кг	2,771	2,859	3,319	3,395	3,354	3,358	3,226	3,253	3,319	3,152	3,284	3,148
<i>Corr</i>	0,27	0,259	0,264	0,254	0,258	0,253	0,271	0,263	0,264	0,236	0,248	0,246

быть связано с избыточностью точек для описания распределения признака.

Как видно из таблицы 3, например, избыточны для никеля пара {19; 56} и для меди пара {139; 165}. *RMSE* пары точек {19; 56} для никеля и {139; 165} для меди больше *RMSE* индивидуальных точек {19} и {56} для никеля и {139} и {165} для меди. Наоборот, коэффициенты корреляции пары точек {88; 145} для никеля и {126; 127} для меди меньше коэффициентов корреляции индивидуальных точек {88} и {145} для никеля и {126} и {127} для меди.

Полученные результаты не противоречат гипотезе о том, что разные локации (геопозиции) несут разную информацию и неодинаковую ценность для интерполяции распределения признака. Оценка репрезентативности точек позволит выбирать наиболее представительные точки для территории.

5. Заключение

Сравнение индивидуальной и парной (коллективной) репрезентативности при включении точек в тренировочный набор показало их неравноценность для интерполяции распределения тяжелых металлов в верхнем слое почвы. Наиболее представительными по индивидуальной репрезентативности оказались точки с максимальным содержанием элемента на выбранной территории. Включение этих точек в обучающий набор ИНС уменьшает ошибку и повышает корреляцию между результатами модельных расчетов и натурных измерений на тестовом наборе. В данной работе мы ограничились парной репрезентативностью, определение объема всех репрезентативных *n*-ок требует огромных вычислительных затрат и остается задачей будущего исследования.

Список источников

1. Буслаева О.В., Королев В.А. Неопределенности в эколого-геологических системах и их систематизация // Инженерная геология. – 2013. – № 6. – С. 56–62.
2. ГОСТ 17.4.3.01-2017 Межгосударственный стандарт. Охрана природы. Почвы. Общие требования к отбору проб / Введен в действие Приказом Ростандарта от 01.06.2018 № 302-ст. – М.: Стандартиформ, 2018. – 8 с.
3. Кургузов К.В., Фоменко И.К., Сироткина О.Н. Вероятностно-статистические подходы при оценке неопределенности литотехнических систем // Геоэкология. Инженерная геология, гидрогеология, геокриология. – 2020. – № 2. – С. 80–89. DOI: 10.31857/S0869780920020071.
4. Мокрушин А.А., Тарасов Д.А., Сергеев А.П., Бувич А.Г., Баглаева Е.М. Подбор типа и структуры искусственных нейронных сетей для оценки распределения химических элементов в верхнем слое почвы // Экологические системы и приборы. – 2017. – № 8. – С. 36–48.
5. РД 52.18.156-93. Методические указания. Охрана природы. Почвы. Методы отбора представительных проб почвы и оценка загрязнения сельскохозяйственного угодья остаточными количествами пестицидов [Электронный ресурс] / Утвержден и введен в действие Федеральной службой России по гидрометеорологии и мониторингу окружающей среды 01.09.1999 г. – Режим доступа: <https://docs.cntd.ru/document/1200041909> (дата обращения: 11.09.2023 г.).
6. Радомская В.И., Бородина Н.А. Оценка антропогенного загрязнения почвы урбанизированной территории на примере города Благовещенска // Геоэкология. Инженерная геология, гидрогеология, геокриология. – 2019. – № 6. – С. 79–93. DOI: 10.31857/S0869-78092019679-93.
7. СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания [Электронный ресурс] / Утвержден Постановлением Главного государственного санитарного врача России от 28.01.2021 г. № 2. – Режим доступа: <http://publication.pravo.gov.ru/document/0001202102030022> (дата обращения: 29.06.2023 г.).

8. Тасейко О.В., Сугак Е.В. Репрезентативность пунктов наблюдения при оценке качества воздуха в городской среде [Электронный ресурс] // Современные проблемы науки и образования. – 2014. – № 6. – 11 с. Режим доступа: <https://science-education.ru/ru/article/view?id=15560> (дата обращения: 29.06.2023 г.).
9. Шичкин А.В., Бувич А.Г., Сергеев А.П., Баглаева Е.М., Субботина И.Е. Прогнозирование содержания аномально распределенного в почве хрома гибридными моделями на основе искусственных нейронных сетей // Геоэкология. Инженерная геология, гидрогеология, геоэкология. – 2018. – № 3. – С. 86–96. DOI: 10.7868/S0869780318030109.
10. 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. DOI: 10.1007/s11004-019-09813-9.
11. Demyanov V., Gloaguen E., Kanevski M. A special issue on data science for geosciences // *Mathematical Geosciences*. – 2020. – Vol. 52. – pp. 1–3. DOI: 10.1007/s11004-019-09846-0.
12. Fernandez Jaramillo 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. – P. 98–106. DOI: 10.1016/j.cageo.2018.02.003.
13. Kramm T., Hoffmeister D. 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 // *Geoderma Regional*. – 2020. – Vol. 22. – e00316. DOI: 10.1016/j.geodrs.2020.e00316.
14. 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 // *Proceedings Volume 10628, Detection and sensing of mines, explosive objects, and obscured targets XXIII*. – 2018. – 106280C. DOI: 10.1117/12.2305793.
15. 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. DOI: 10.1016/j.neucom.2017.07.004.
16. 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. – No. 3. – pp. 968–982. DOI: 10.1007/s11368-017-1847-5.
17. Wang X., An Sh., Xu Y., Hou H., Chen F., Yang Y., Zhang Sh., Liu R. 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 // *Applied Sciences*. – 2020. – Vol. 10. – No. 1. – 51. DOI: 10.3390/app10010051.
18. Zhong L., Guo X., Xu Zh., Ding M. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks // *Geoderma*. – 2021. – Vol. 402. – 115366. DOI: 10.1016/j.geoderma.2021.115366.

References

1. Buslaeva O.V., Korolev V.A. Indeterminacies in the environmental-geological systems and their systematization. *Engineering Geology World*. 2013;(6):56–62.
2. GOST 17.4.3.01-2017. Mezhgosudarstvennyi standart. Okhrana prirody. Pochvy. Obshchie trebovaniya k otboru prob [Interstate standard. Protection of nature. Soils. General sampling requirements]. Moscow: Standartinform; 2018. 8 p.
3. Kurguzov K.V., Fomenko I.K., Sirotkina O.N. Probabilistic and statistical approaches to uncertainty assessment in lithotechnogenic systems. *Geoekologiya. Inzheneraya geologiya, gidrogeologiya, geokriologiya*. 2020;(2):80–89. DOI: 10.31857/S0869780920020071.
4. Mokrushin A.A., Tarasov D.A., Sergeev A.P., Buevich A.G., Baglaeva E.M. Selection of type and structure of artificial neural networks for estimation of chemical elements distribution in topsoil. *Ecological Systems and Devices*. 2017;(8):36–48.
5. RD 52.18.156-93. Metodicheskie ukazaniya. Okhrana prirody. Pochvy. Metody otbora predstavitel'nykh prob pochvy i otsenka zagryazneniya sel'skokhozyaistvennogo ugod'ya ostatochnymi kolichestvami pestitsidov [Methodical instructions. Protection of nature. Soils. Methods for taking representative soil samples and assessing contamination of agricultural land with pesticide residues]. Available at: <https://docs.cntd.ru/document/1200041909> (accessed 11.09.2023).
6. Radomskaya V.I., Borodina N.A. Assessment of anthropogenic contamination in an urban territory by the example of Blagoveshchensk city. *Geoekologiya. Inzheneraya geologiya, gidrogeologiya, geokriologiya*. 2019;(6):79–93. DOI: 10.31857/S0869-78092019679-93.
7. SanPin 1.2.3685-21 Gigienicheskie normativy i trebovaniya k obespecheniyu bezopasnosti i (ili) bezvrednosti dlya cheloveka faktorov sredy obitaniya [Hygienic standards and requirements for ensuring the safety and (or) harmlessness of environmental factors for humans]. Available at: <http://publication.pravo.gov.ru/document/0001202102030022> (accessed 29.06.2023).
8. Taseyko O.V., Sugak E.V. Representativeness of urban station for air quality monitoring. *Modern problems of science and education*. 2014;(6). Available at: <https://science-education.ru/ru/article/view?id=15560> (accessed 29.06.2023).
9. Shichkin A.V., Buevich A.G., Sergeev A.P., Baglaeva E.M., Subbotina I.E. Forecasting the content of abnormally distributed chrome in soil by hybrid models based on artificial neural networks. *Geoekologiya. Inzheneraya geologiya, gidrogeologiya, geokriologiya*. 2018;(3):86–96. DOI: 10.7868/S0869780318030109.
10. 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;52:111–121. DOI: 10.1007/s11004-019-09813-9.
11. Demyanov V., Gloaguen E., Kanevski M. A special issue on data science for geosciences. *Mathematical Geosciences*. 2020;52:1–3. DOI: 10.1007/s11004-019-09846-0.
12. Fernandez Jaramillo J. M., Mayerle R. Sample selection via angular distance in the space of the arguments of an artificial neural network. *Computers and Geosciences*. 2018;114:98–106. DOI: 10.1016/j.cageo.2018.02.003.
13. Kramm T., Hoffmeister D. 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. *Geoderma Regional*. 2020;22:e00316. DOI: 10.1016/j.geodrs.2020.e00316.
14. 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. *Proceedings Volume 10628, Detection and sensing of mines, explosive objects, and obscured targets XXIII*. 2018;10628:106280C. DOI: 10.1117/12.2305793.

15. 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;272:294–305. DOI: 10.1016/j.neucom.2017.07.004.
16. 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;18(3):968–982. DOI: 10.1007/s11368-017-1847-5.
17. Wang X., An Sh., Xu Y., Hou H., Chen F., Yang Y., Zhang Sh., Liu R. 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. *Applied Sciences*. 2020;10(1):51. DOI: 10.3390/app10010051.
18. Zhong L., Guo X., Xu Zh., Ding M. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. *Geoderma*. 2021;402:115366. DOI: 10.1016/j.geoderma.2021.115366.

Статья поступила в редакцию 27.12.2022, одобрена после рецензирования 20.05.2023, принята к публикации 01.07.2023.
The article was submitted 27.12.2022; approved after reviewing 20.05.2023; accepted for publication 01.07.2023.

Информация об авторах

Баглаева Елена Михайловна

Кандидат физико-математических наук
Старший научный сотрудник
Институт промышленной экологии УрО РАН
620990 Екатеринбург, ул. С. Ковалевской, д. 20
e-mail: e.m.baglaeva@urfu.ru

Сергеев Александр Петрович

Кандидат физико-математических наук
Ведущий научный сотрудник, и.о. заведующего лабораторией
Институт промышленной экологии УрО РАН
620990 Екатеринбург, ул. С. Ковалевской, д. 20
e-mail: sergeev@ecko.uran.ru

Шичкин Андрей Васильевич

Научный сотрудник
Институт промышленной экологии УрО РАН
620990 Екатеринбург, ул. С. Ковалевской, д. 20
e-mail: and@ecko.uran.ru

Бувевич Александр Геннадьевич

Научный сотрудник
Институт промышленной экологии УрО РАН
620990 Екатеринбург, ул. С. Ковалевской, д. 20
e-mail: bag@ecko.uran.ru

Буторова Анастасия Сергеевна

Инженер-исследователь
Институт промышленной экологии УрО РАН
620990 Екатеринбург, ул. С. Ковалевской, д. 20
Аспирант 1 курса
Институт радиоэлектроники и информационных технологий — РтФ Уральского федерального университета им. Б.Н. Ельцина
620002 Екатеринбург, ул. Мира, д. 19
e-mail: a.s.butorova@urfu.ru

Information about authors

Elena M. Baglaeva

Candidate of Physical and Mathematical Sciences,
Senior Researcher
Institute of Industrial Ecology UB RAS
20, S. Kovalevskoy str., Ekaterinburg, 620990, Russia
e-mail: e.m.baglaeva@urfu.ru

Aleksandr P. Sergeev

Candidate of Physical and Mathematical Sciences
Leading Researcher, Acting Head of the Laboratory
Institute of Industrial Ecology UB RAS
20, S. Kovalevskoy str., Ekaterinburg, 620990, Russia
e-mail: sergeev@ecko.uran.ru

Andrey V. Shichkin

Researcher
Institute of Industrial Ecology UB RAS
20, S. Kovalevskoy str., Ekaterinburg, 620990, Russia
e-mail: and@ecko.uran.ru

Alexander G. Buevich

Researcher
Institute of Industrial Ecology, UB RAS
20, S. Kovalevskoy str., Ekaterinburg, 620990, Russia
e-mail: bag@ecko.uran.ru

Anastasia S. Butorova

Research Engineer
Institute of Industrial Ecology UB RAS
20, S. Kovalevskoy str., Ekaterinburg, 620990, Russia
1st year Postgraduate Student
Institute of Radio Electronics and Information Technologies —
RTF of the Ural Federal University named after B.N. Yeltsin
19, Mira str., Ekaterinburg, 620002, Russia
e-mail: a.s.butorova@urfu.ru