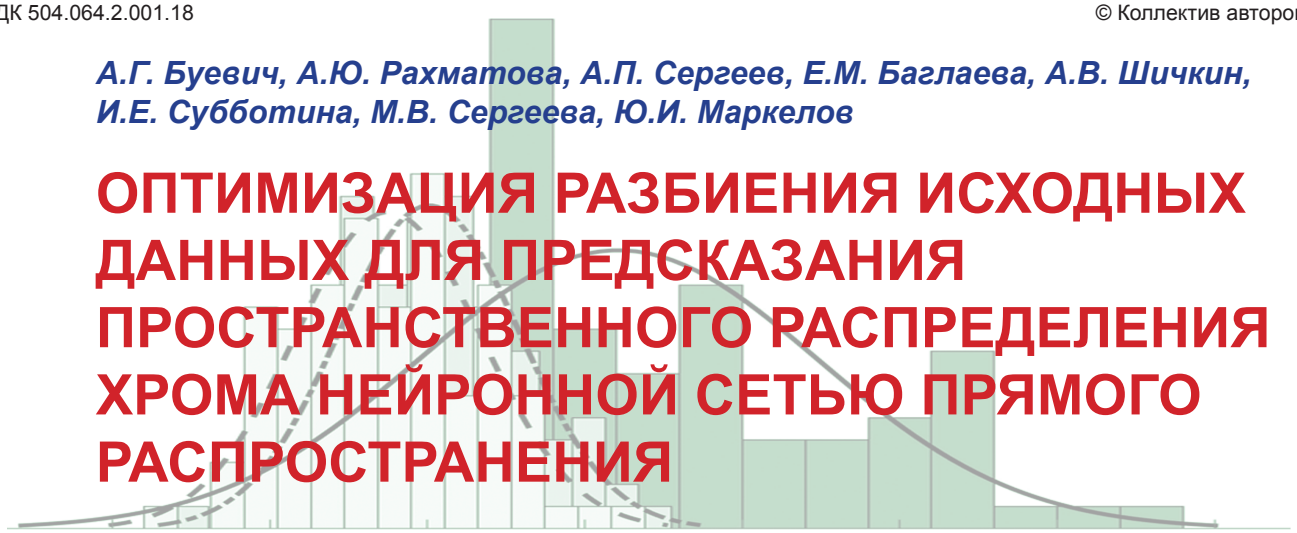
УДК 504.064.2.001.18

А.Г. Буевич, А.Ю. Рахматова, А.П. Сергеев, Е.М. Баглаева, А.В. Шичкин, И.Е. Субботина, М.В. Сергеева, Ю.И. Маркелов

ОПТИМИЗАЦИЯ РАЗБИЕНИЯ ИСХОДНЫХ ДАННЫХ ДЛЯ ПРЕДСКАЗАНИЯ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ХРОМА НЕЙРОННОЙ СЕТЬЮ ПРЯМОГО РАСПРОСТРАНЕНИЯ



Введение

При восстановлении пространственного распределения признака можно выделить две основные задачи. Первая – выбор или создание модели, которая способна с приемлемой точностью воспроизводить картину пространственного распределения признака. Одними из самых продуктивных методов моделирования являются методы на основе искусственных нейронных сетей (ИНС) прямого распространения. Точность восстановления, полученная с помощью ИНС, часто выше, чем у других методов или прогнозов [1, 2]. Модели на основе ИНС могут применяться к измеренным данным, полученным в ходе мониторинга (скрининга) для оценки содержания загрязняющих веществ в неконтролируемых местах [3-5].

Вторая задача состоит в том, чтобы выбранная модель смогла полностью реализовать все свои возможности. Точность модели во многом зависит от процедуры обучения, поэтому одним из путей улучшения возможностей метода является оптимизация процедуры разделения выборочной совокупности (выборки) на обучающее и тестовое подмножества. Процедура разделения выборки традиционно используется при моделировании пространственного распределения. При таком подходе обучающее подмножество используется для обучения моделей, а тестовое подмножество только один раз – для проверки точности модели. Обычно используется метод случайного разбиения, который показывает приемлемую точность, прост и эффективен. Тем не менее у такого подхода есть ряд недостатков.

Исследования пространственного распределения параметров окружающей среды (например, концентрации химических элементов в депонирующих средах) обычно строятся на исходных данных, собранных по некоторой предварительно выбранной схеме. Чаще всего точки отбора находятся в узлах решетки с соответствующим масштабу исследования шагом. Область исследования может представлять собой пересеченную местность со значительно отличающимися характеристиками. Присутствие выраженного рельефа, разных типов почв, наличие водоемов и антропогенного воздействия и ряд других факторов делают обследуемую территорию пространственно гетерогенной. Особенно это относится к урбанизированным территориям. Требование проведения пробоотбора на естественных, ненарушенных участках также значительно влияет на итоговую картину реальных мест опробования. Таким образом, при разбиении пространственно-размещенной выборочной совокупности на обучающее и тестовое подмножества было бы разумно учесть представительство исходных данных.

С другой стороны, как указывается в ряде исследований [6], частотное (вероятностное) распределение самого моделируемого признака часто оказывается далеким от нормального или даже мономодального. Возможен довольно значительный разброс величины моделируемого показателя, наличие пятен или областей с аномально высокими (или низкими) значениями. Этот факт также необходимо учитывать при разбиении данных на обучающее и тестовое подмножества.

Основываясь на собственном опыте [1] и результатах других исследователей [7-11], в качестве опорной модели выбрали тип ИНС многослойный перцептрон (МЛП) прямого распространения с алгоритмом обучения Левенберга-Марквардта. Это эффективный метод, широко применяемый в исследованиях, касающихся распределения химических элементов в почве, в частности тяжелых металлов [12, 13].

Целью работы является оптимизация метода однократного разбиения пространственно размещенной выборочной совокупности на обучающее и тестовое подмножества для задачи интерполяции методом многослойного персептрона.

Материалы и методы

Область исследования. Данные для исследования были получены по результатам скрининга поверхностного слоя почвы двух субарктических городов – Ноябрьска (N63,2°, E75,5°) и Тарко-Сале (N64,9°, E77,8°) – Ямало-Ненецкого автономного округа, Россия (рис. 1). Этот регион является субарктическим климатическим районом (тип Dfc по классификации климата Кёппена).

Почвы обследуемых районов преимущественно состоят из песка (размер фракции менее 1 мм) и относятся к торфяно-подзолистому типу [14].

Всего было отобрано 338 проб поверхностного слоя почвы в жилых зонах двух городов: 101 проба – в Тарко-Сале и 237 проб – в Ноябрьске.

Отбор проб почвы в черте городов выполнен по точкам, расположенным в узлах квадратной сетки с шагом 250 м. Фактическое их расположение определялось при проведении опробования непосредственно на местности, исходя из необходимости отбора проб почвы на естественных участках исследуемой территории. Географическая привязка осуществлялась с помощью GPS-приемника. Поверхность места предполагаемого отбора пробы почвы размечалась в виде квадрата со стороной 1 метр. В вершинах, серединах сторон и в центре размеченного квадрата пробоотборником из нержавеющей стали диаметром 0,05 м отбирались от пяти до девяти кернов почвы на глубину 0,05 м. Отобранные керны объединялись в одну пробу и запаковывались в двойные полиэтиленовые пакеты для пищевых продуктов. На внутреннем пакете маркером наносился идентификатор пробы.

Анализ образцов. Содержание химических элементов в пробах получено методом химического анализа. Подготовка образцов почвы и химический анализ проводились в соответствии с действующими нормативными требованиями. Химическая лаборатория, занимающаяся подготовкой и анализом проб почвы, сертифицирована Федеральной системой сертификации России. Лаборатория отвечает общим требованиям к компетентности испытательных и калибровочных лабораторий ISO/IEC 17025:2005. В пробах почвы определены следующие химические элементы: Al, Cr, Mn, Fe, Co, Ni, Zn, Pb. Для модели был выбран Cr. Содержание Cr в г. Тарко-Сале образовывало области с аномально высоким значением. В г. Ноябрьске аномалий в значениях содержания Cr не выявлено.

Алгоритм разбиения. В работе использовано три типа разбиения.

1. Полностью случайный метод, при котором весь набор данных случайным образом делится на обучающее и тестовое подмножество в отношении 70% на 30% соответственно.
2. Пространственное квотирование исходных данных, которое состоит из трех шагов. На первом шаге пространственно-размещенная выборочная совокупность оконтуривается выпуклым многоугольником так, чтобы геодезическая, проведенная между любыми двумя точками (реальные точки пробоотбора),

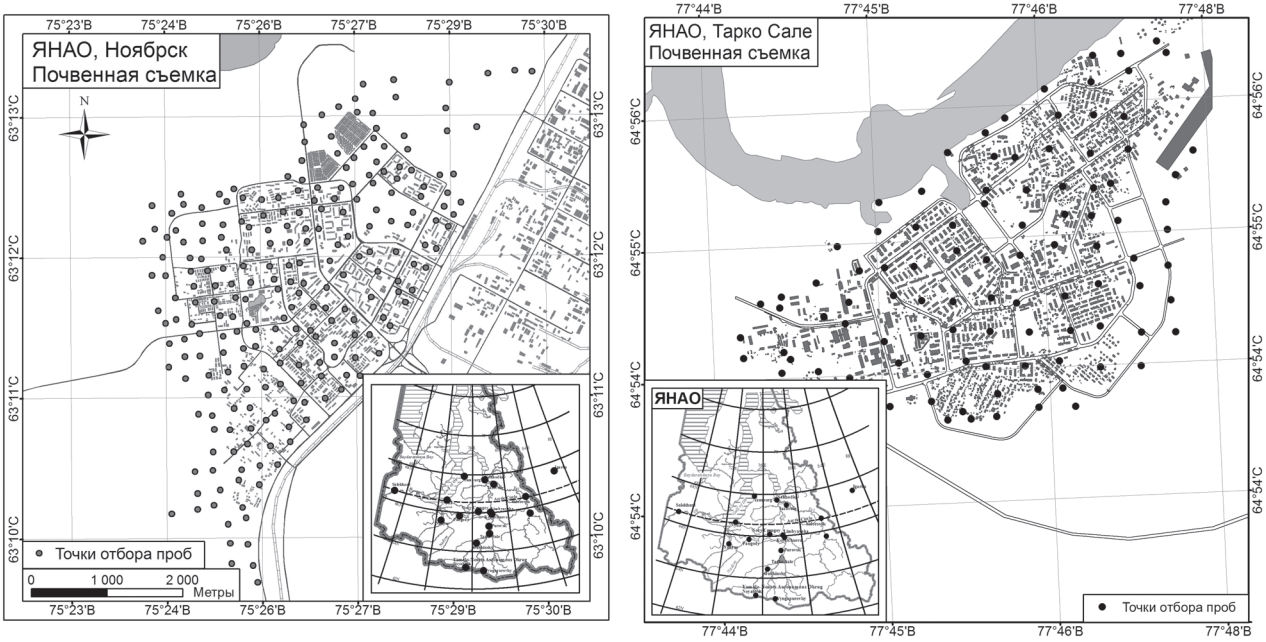


Рис. 1. Карта-схема территорий отбора проб

Fig. 1. The sampling areas

оказалась внутри этого многоугольника. Оконтуривание происходит путем соединения граничных точек. Такая процедура удовлетворяет условию интерполяции, поскольку любое восстановленное значение оказывается внутри этого многоугольника. На втором шаге многоугольник делится на области (пространственные квоты) включающие одинаковое количество наблюдений (реальные точки отбора). Граничные точки многоугольника включаются в обучающее подмножество. Также в обучающее подмножество обязательно попадают максимальное и минимальное значения из каждой области (пространственной квоты). На третьем шаге из каждой пространственной квоты случайным образом происходит отбор значений для обучающего подмножества так, чтобы его доля составляла 70%. Таким образом, итоговое обучающее подмножество состоит из граничных точек, максимальных и минимальных значений из каждой пространственной квоты и случайного добора из каждой пространственной квоты до 70%. Оставшиеся 30% значений попадают в тестовую выборку.

3. Пространственное квотирование исходных данных, учитывающее разброс значений моделируемой переменной. После пространственного квотирования исходных данных оставшиеся значения моделируемой переменной из каждой пространственной квоты (граничные точки и максимальное и минимальное значения попадают в обучающее подмножество) разбиваются на квартили. Значения, попавшие в каждый квартиль также случайным образом, разбиваются на обучающее и тестовое подмножества так, чтобы итоговое соотношение получилось 30% на 70% соответственно. Возможны варианты, при которых граничные точки и максимальное и минимальное значения будут совпадать. В таком случае в обучающее подмножество попадет больше «случайных» значений. Итоговые обучающее и тестовое подмножества есть сумма соответствующих малых подвыборок.

Построение МЛП. Структура МЛП была выбрана методом компьютерного симулирования на основе минимизации среднеквадратической ошибки (*RMSE*). МЛП имел один скрытый слой, количество нейронов варьировалось от 2 до 20. Каждая сеть обучалась 500 раз, и выбиралась лучшая.

Сравнены результаты моделей на основе МЛП, при построении которых применялись разные методы разделения исходных данных на обучающее и тестовое подмножество: случайное разделение, пространственное квотирование исходных данных и пространственное квотирование исходных данных, учитывающее разброс значений моделируемой переменной.

Оценка точности моделей. Для каждого из подходов вычислялись средняя абсолютная ошибка (*MAE*), среднеквадратическая ошибка (*RMSE*) и среднеквадратическая относительная ошибка (*RMSRE*):

$$MAE = \frac{\sum_{i=1}^{n} \left| p(x_i) - o(x_i) \right|}{n}, \quad (1)$$

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

$$RMSRE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{p(x_i) - o(x_i)}{o(x_i)} \right)^2}, \quad (3)$$

где $p(x_i)$ – прогнозируемая концентрация, $o(x_i)$ – наблюдаемая концентрация, а n – количество точек.

Результаты и обсуждение

Представлены описательные статистики концентрации Cr в почве для каждого из городов (табл. 1). Распределения содержания элемента имеют асимметрию. По сравнению с фоновым содержанием в Уральском регионе (Ural Clarke) и в мировых почвах (World Clarke) общее содержание Cr на обследуемых территориях близко к контрольным значениям, в то время как содержание Cr на участках аномалии было в несколько раз выше [15, 16]. Распределение содержания Cr в почве г. Ноябрьска близко к нормальному, в г. Тарко-Сале распределение имеет бимодальный характер (рис. 2a).

Была выбрана структура МЛП, которая включала один скрытый слой с 8 нейронами (рис. 2b). Были вычислены показатели точности моделей на основе МЛП (табл. 2). Показатели с наименьшей ошибкой выделены жирным шрифтом.

Для города Тарко-Сале территория обследования была разделена на 4 области, включающие 25 точек пробоотбора каждая. Для Ноябрьска каждая из 4 областей включала 59 (60) точек (рис. 3).

Построены частотные (вероятностные) распределения ошибок моделей: *MAE*, *RMSE*, *RMSRE* – для разных методов разбиения исходных данных (рис. 4). Модели с контролируемым разбиением более точны, кроме того их распределение «уже», чем модели на основе случайного разбиения. Модель МЛП с пространственно-квартильным разбиением

Таблица 1

Описательные статистики моделируемого элемента (Cr)

Descriptive statistics of the modelled element (Cr)

Участок	Минимальное значение, мг/кг	Максимальное значение, мг/кг	Среднее значение, мг/кг	Медиана, мг/кг	Стандартное отклонение, мг/кг	Коэффициент вариации, %	Коэффициент асимметрии	Коэффициент эксцесса
Тарко-Сале	35	1424	259	87	337	130	1,6	1,2
Ноябрьск	17	140	62	59	24	39	0,8	0,4

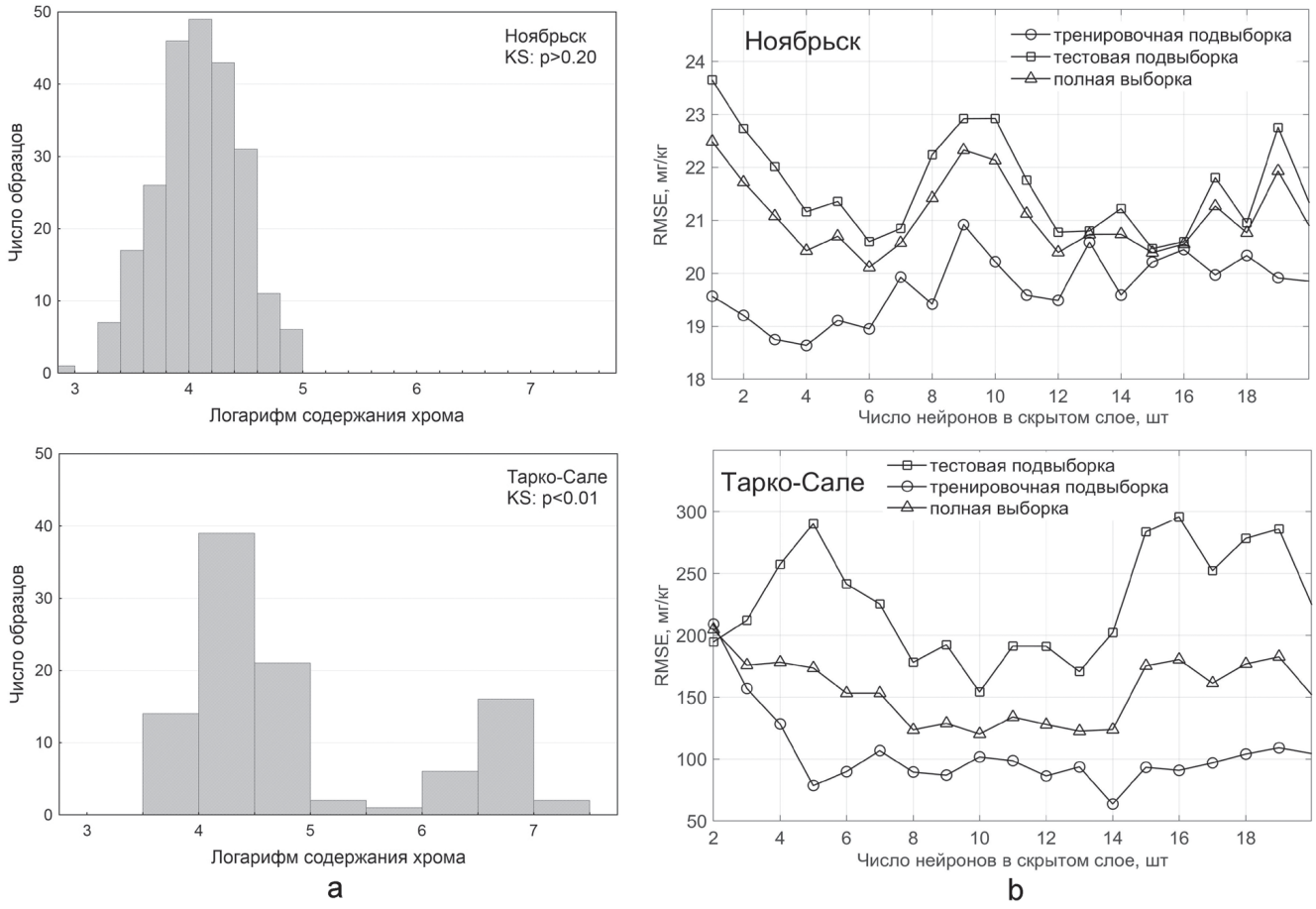


Рис. 2. Гистограмма логарифма содержания Cr (результат теста Колмогорова-Смирнова) (a); выбор структуры МЛП на основе минимизации *RMSE*: тестовые, тренировочные и общие данных для различного числа нейронов в скрытом слое (b)

Fig. 2. Histogram of the Cr content logarithm (the result of the Kolmogorov-Smirnov test) (a); Selection MLP structure based on the minimization of RMSE: test, training and general data for different numbers of neurons in the hidden layer (b)

Таблица 2

Оценка точности показателей моделей МЛП для содержания Cr

Accuracy assessment of MLP model indices of the Cr content

Город	Показатель	Статистический параметр	Тип разбиения на обучающее и тестовое подмножества		
			Случайное разбиение	Пространственное квотирование	Пространственно-квартильное квотирование
Тарко-Сале	*MAE*	Mean, мг/кг	101,4	91,8	**90**
		SD, мг/кг	21	**17,4**	18,5
		Median, мг/кг	102	90,7	**89,6**
Тарко-Сале	*RMSE*	Mean, мг/кг	152,3	138,2	**133,4**
		SD, мг/кг	30,9	**22,6**	25,2
		Median, мг/кг	151,4	135,4	**134,8**
Тарко-Сале	*RMSRE*	Mean	1,3	1,2	**1,1**
		SD	0,4	0,4	**0,4**
		Median	1,3	1,1	**1**
Ноябрьск	*MAE*	Mean, мг/кг	16,1	15,1	**14,9**
		SD, мг/кг	1,5	1,1	**0,9**
		Median, мг/кг	16,2	14,9	**14,9**
Ноябрьск	*RMSE*	Mean, мг/кг	20,6	18,9	**18,8**
		SD, мг/кг	1,9	1,3	**1,1**
		Median, мг/кг	20,6	18,9	**18,9**
Ноябрьск	*RMSRE*	Mean	0,41	0,35	**0,34**
		SD	0,06	0,03	**0,03**
		Median	0,39	0,35	**0,34**

Примечание:
Mean – среднее значение,
SD – стандартное отклонение,
Median – медиана.

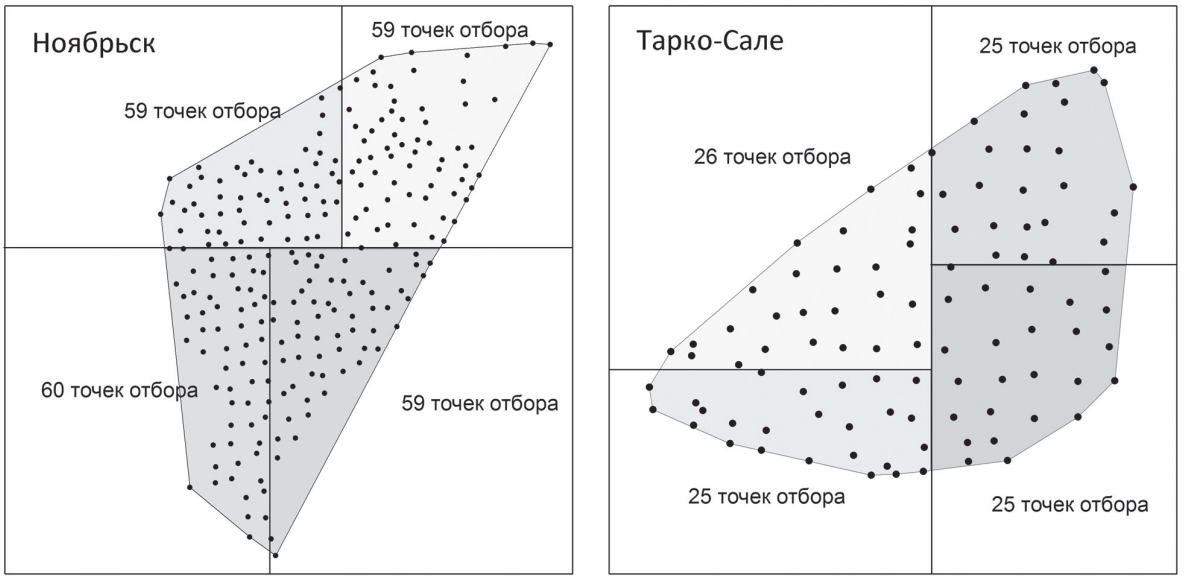


Рис. 3. Места отбора, оконтуренные выпуклым многоугольником

Fig. 3. The survey areas contoured by a convex polygon

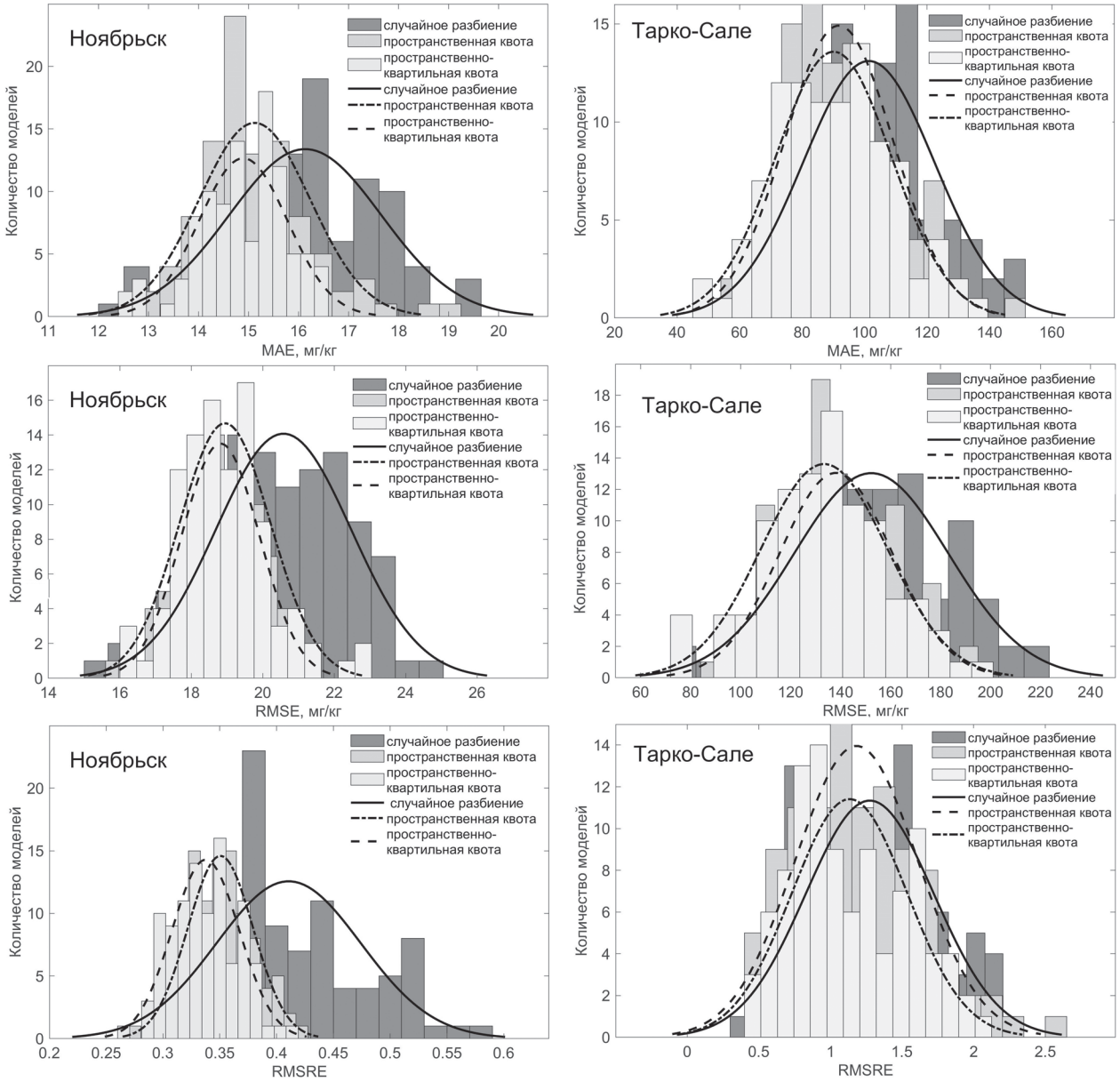


Рис. 4. Частотные (вероятностные) распределения ошибок модели

Fig. 4. Frequency distribution of the model errors

оказалась точнее, чем модель только с пространственным разбиением (разница составила около 5%).

Применяемый подход позволил улучшить точность модели на основе МЛП. Улучшение было особенно заметно для стандартного отклонения (SD) по всем показателям – оно составило около 50% (табл. 2).

Для условно нормально распределенных данных в Ноябрьске (рис. 2а) способ пространственного квотирования исходных данных, учитывающий разброс значений моделируемой переменной, продемонстрировал лучшие результаты по всем показателям. Для бимодально распределенных данных в Тарко-Сале способ только пространственного квотирования улучшил показатель SD для индексов *MAE* и *RMSE*. Способ пространственного квотирования с последующим учетом разброса значений моделируемой переменной позволил улучшить точность для остальных показателей (Mean и Median для *MAE* и *RMSE*; Mean, SD, Median для *RMSRE*). Однако разница между способами пространственного квотирования исходных данных: учитывающим и не учитывающим разброс значений моделируемой переменной, была незначительной (около 5%). Такую незначительную разницу между двумя контролируемыми типами разбиения можно объяснить относительно небольшим количеством точек отбора проб.

Для относительно малых выборок (менее 80 значений) процедура деления на обучающее и тестовое подмножества может быть практически полностью детерминирована. Это значительно упрощает расчет модели, при этом точность модели будет выше – по сравнению со случайным разбиением исходных данных.

Заключение

Оптимизация разбиения исходных данных на обучающее и тестовое подмножества улучшила точность моделей на основе МЛП. Алгоритм был опробован для моделирования пространственного распределения содержания химического элемента Cr в поверхностном слое почвы урбанизированных территорий. Для всех индикаторов и на всех территориях модели МЛП, использующие метод контролируемого разбиения, оказались более точными в сравнении со случайным разбиением выборки. Наиболее точные результаты были у модели, использующей разбиение, которое учитывало пространственную квоту и разброс величин значений содержания Cr в почве. Описанный метод относительно прост в реализации в современных вычислительных пакетах и может использоваться для моделирования пространственных переменных.

Ключевые слова: искусственные нейронные сети, многослойный перцептрон, моделирование, разбиение, выборка.

ЛИТЕРАТУРА

1. 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. – P. 425-435.
2. Guo G.H., Wu F., Xie F., Zhang R. Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China // Journal of Environmental Sciences. – 2012. – Vol. 24, Issue 3. – P. 410-418.
3. Liu F., He X., Zhou L. Application of generalized regression neural network residual kriging for terrain surface interpolation // Proc. SPIE 7492, International Symposium on Spatial Analysis, Spatial-Temporal Data Modeling, and Data Mining. – 2009. – 74925F.
4. Land cover and landscape as predictors of groundwater contamination: a neural-network modelling approach applied to Dobrogea, Romania / R.R. Shaker [et al.] // Journal of Environmental Protection and Ecology. – 2010. – Vol. 11, Issue 1. – P. 337-348.
5. 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, Issue 4. – P. 1-20.
6. Use of trans-Gaussian kriging for national soil geochemical mapping in Ireland / C. Zhang [et al.] // Geochemistry: Exploration Environment Analysis. – 2008. – Vol. 8, Issue 3-4. – P. 255-265.
7. Koike K., Matsuda S., Suzuki T., Ohmi M. Neural Network-Based Estimation of Principal Metal Contents in the Hokuroku District, Northern Japan, for Exploring Kuroko-Type Deposits // Natural Resources Research. – 2002. – Vol. 11, Issue 2. – P. 135-156.
8. Samanta B., Ganguli R., Bandopadhyay S. Comparing the Predictive Performance of Neural Networks with Ordinary Kriging in a Bauxite Deposit // Transactions of Institute of Mining and Metallurgy, Section A, Mining Technology. – 2005. – Vol. 114, Issue 3. – P. 129-139.
9. Spatial prediction of soil organic matter content integrating artificialneural network and ordinary kriging in Tibetan Plateau / F. Dai [et al.] // Ecological Indicators. – 2014. – Vol. 45. – P. 184-194.
10. Falamaki A. Artificial neural network application for predicting soil distribution coefficient of nickel // Journal of Environmental Radioactivity. – 2014. – Vol. 115. – P. 6-12.
11. Li Y., Li C., Tao J.-J., Wang L.-D. Study on Spatial Distribution of Soil Heavy Metals in Huizhou City Based on BP-ANN Modeling and GIS // Procedia Environmental Sciences. – 2011. – Vol. 10. – P. 1953-1960.
12. Qualitative and quantitative investigation of chromium-polluted soils by laser-induced breakdown spectroscopy combined with neural networks analysis / J.-B. Sirven [et al.] // Anal Bioanal Chem. – 2006. – Vol. 385, Issue 2. – P. 256-262.
13. Anagu I., Ingwersen J., Utermann J., Streck T. Estimation of heavy metal sorption in German soils using artificial neural networks // Geoderma. – 2009. – Vol. 152. – P. 104-112.
14. Добровольский Г.В., Урусевская И.С. География почв. – М. : Изд-во МГУ ; Изд-во «КолосС», 2004. – 460 с.
15. Геохимия окружающей среды / Ю.Е. Сает [и др.]. – М. : Недра, 1990. – 335 с.
16. Справочник по геохимии / Г.В. Войткевич [и др.]. – М. : Недра, 1990. – 479 с.

REFERENCES

1. 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. P. 425-435.
2. Guo G.H., Wu F., Xie F., Zhang R. Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China // Journal of Environmental Sciences. 2012. Vol. 24, Issue 3. P. 410-418.
3. Liu F., He X., Zhou L. Application of generalized regression neural network residual kriging for terrain

surface interpolation // Proc. SPIE 7492, International Symposium on Spatial Analysis, Spatial-Temporal Data Modeling, and Data Mining. 2009. 74925F.
4. Land cover and landscape as predictors of groundwater contamination: a neural-network modelling approach applied to Dobrogea, Romania / R.R. Shaker [et al.] // Journal of Environmental Protection and Ecology. 2010. Vol. 11, Issue 1. P. 337-348.
5. 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, Issue 4. P. 1-20.
6. Use of trans-Gaussian kriging for national soil geochemical mapping in Ireland / C. Zhang [et al.] // Geochemistry: Exploration Environment Analysis. 2008. Vol. 8, Issue 3-4. P. 255-265.
7. Koike K., Matsuda S., Suzuki T., Ohmi M. Neural Network-Based Estimation of Principal Metal Contents in the Hokuroku District, Northern Japan, for Exploring Kuroko-Type Deposits // Natural Resources Research. 2002. Vol. 11, Issue 2. P. 135-156.
8. Samanta B., Ganguli R., Bandopadhyay S. Comparing the Predictive Performance of Neural Networks with Ordinary Kriging in a Bauxite Deposit // Transactions of Institute of Mining and Metallurgy, Section A, Mining Technology. 2005. Vol. 114, Issue 3. P. 129-139.
9. Spatial prediction of soil organic matter content integrating artificialneural network and ordinary kriging in Tibetan Plateau / F. Dai [et al.] // Ecological Indicators. 2014. Vol. 45. P. 184-194.
10. Falamaki A. Artificial neural network application for predicting soil distribution coefficient of nickel // Journal of Environmental Radioactivity. 2014. Vol. 115. P. 6-12.
11. Li Y., Li C., Tao J.-J., Wang L.-D. Study on Spatial Distribution of Soil Heavy Metals in Huizhou City Based on BP-ANN Modeling and GIS // Procedia Environmental Sciences. 2011. Vol. 10. P. 1953-1960.
12. Qualitative and quantitative investigation of chromium-polluted soils by laser-induced breakdown spectroscopy combined with neural networks analysis / J.-B. Sirven [et al.] // Anal Bioanal Chem. 2006. Vol. 385, Issue 2. P. 256-262.
13. Anagu I., Ingwersen J., Utermann J., Streck T. Estimation of heavy metal sorption in German soils using artificial neural networks // Geoderma. 2009. Vol. 152. P. 104-112.
14. Dobrovolsky G.V., Urusevskaya I.S. Soils geography. Moscow : MSU Publishing House ; Kolos-s Publishing House, 2004. 460 p.
15. Environment Geochemistry / Yu.E. Saet [et al.]. Moscow : Nedra, 1990. 335 p.
16. Geochemistry handbook / G.V. Voitkevich [et al.]. Moscow : Nedra, 1990. 479 p.