RESEARCH PAPER

INDIVIDUAL AND PAIRWISE REPRESENTATIVENESS OF SAMPLING POINTS IN INTERPOLATION TASKS OF HEAVY METALS DISTRIBUTION IN THE TOPSOIL

Elena M. Baglaeva*, Aleksandr P. Sergeev, Andrey V. Shichkin, Alexander G. Buevich
Institute of Industrial Ecology UB RAS, S. Kovalevskaya str., 20, Ekaterinburg, 620990, Russia
*Corresponding author: elenbaglaeva@gmail.com

Received: : February 8th 2024 / Accepted: November 27th 2024 / Published: March 31st 2025
https://doi.org/10.24057/2071-9388-2025-3240

ABSTRACT. The optimization of environmental soil monitoring based on representative selection of a training subset for an artificial neural network is an unresolved problem in the tasks of interpolation of the distribution of metals in the topsoil. The soil survey data, often used as input for artificial neural network modeling, are datasets at irregular points. Usually, the division of the input data into training and test subsets is carried out randomly in a ratio of 70% to 30% points, respectively. The question of the individual and collective representativeness of local sampling points on the element content in the soil in a given area for a training subset remains beyond the scope of interpolation problems. In this work, the representativeness of the sampling points plays a crucial role in reducing the ANN error and enhancing the correlation between the results of model calculations on the test subset and natural measurements when the points are part of the training subset. When evaluating the pairwise representativeness, we found two types of effects: synergy and anti-synergy. The synergy was achieved with an increase in model accuracy when the pair entered the training subset. The anti-synergy manifested in a decrease informativeness of the point pair for modeling. The various sampling locations have different information and unequal meaning for feature interpolation. The scale-free network structures were found to have pairwise representativeness by *RMSE*.

KEYWORDS: representativeness, sampling, heavy metals, artificial neural networks, training subset

CITATION: Baglaeva E. M., Sergeev A. P., Shichkin A. V., Buevich A. G. (2025). Individual And Pairwise Representativeness Of Sampling Points In Interpolation Tasks Of Heavy Metals Distribution In The Topsoil. Geography, Environment, Sustainability, 6-13
https://doi.org/10.24057/2071-9388-2025-3240

Conflict of interests: The authors reported no potential conflict of interest.

INTRODUCTION

The environmental soil monitoring methods often require preliminary data to be sufficient to represent soil–environment relationships throughout the study area (Zhu 2015). A limited quantity of soil sample data to represent the study area is still an issue to predict soil properties and estimate prediction uncertainty. A large number of publications are devoted to the issues of representative sampling of the components of the environment (Malof 2018; Liu 2022). The task of assessing representativeness and constructing a representative set arises when organizing sampling to assess the quality of environmental components, when statistically processing environmental monitoring data, and when choosing a training subset for artificial neural networks (ANNs) that model the spatial distribution of a feature (Nath 2018; Demyanov 2020; Mello 2022). The existing rules for choosing a training subset do not reflect the picture of pollution (Baglaeva 2020; Malof 2018). The formal structure of the training subset must be determined by the rules governing the origin and maintenance of ecological topologies in order to correctly interpret ecological patterns (Prager 2009).

Insufficient attention is paid to the interpretation of the results of a modeling. Often behind the scenes is the connection between the features of environmental data and landscapes (Boussange 2022). The key challenge is understanding how the connectivity and heterogeneity of the model results relate to environmental characteristics. To characterize environmental connectivity, these tasks are proposed to be solved by spatial graph theory methods (O'Brien 2006, Urban 2009). We are interested in how graph topology is combined with the spatial distribution of element contents in topsoil. The graph topology properties demonstrate landscape complexity and allow us to determine a finite size of local basic landscape diversity. Using graph topology, we evaluate a representativeness training subset to build the element content distribution in the topsoil. Our study suggests a formalization of assessment of individual and collective representativeness for sampling points to explain the connected landscape pattern. If there are individual sampling points that are important at some scale but not at another one, then there can be doublets, triplets, or *n*-lets of the sample that are important for modeling at some scale but not at another one (Dale 2010; Shu 2015).

Monitoring of environmental parameters in the conditions of urban development is not able to provide complete spatial and temporal characteristics of pollution (O'Hare 2020; Pesch 2008; Zhong 2021). For a comprehensive assessment of the levels of environmental pollution in cities

(Wang 2020; Xu 2023), monitoring is often combined with other methods of obtaining data, including models based on ANN. In the scientific literature, works have been published when the selection of points in the training subset occurs using information about the distribution of the feature under study, but the gain in the accuracy of the model turns out to be small (Kramm 2020; Fernandez Jaramillo 2018; Gutierrez-Velez 2020). The prediction accuracy of ANN models is greatly influenced by the choice of points used to train the ANN. Random sampling points should only be used on a homogeneous experimental site (Legendre 2004; Prager 2009). As demonstrated by (Wang 2021; Ziggah 2019; Baglaeva 2021), various sampling points make contribute differently to the ANN forecast error, i.e., have different - representativeness for the purposes of the forecast.

Previously, some authors presented a definition of representativeness. Zhu (2018) uses the representativeness of a single sampling point and a sampling point set to other points as the similarity of these points to the sampling point set. The representativeness is a similarity in geographic configuration between sample point *k* and prediction point *i*, which is then used as the weight in the prediction of the value of the target variable at prediction point *i*, together with the other involved sample points whose weights are determined similarly. And this similarity is also used to measure the uncertainty associated with the prediction (Levin 2002; Zhu 2018). By representativeness, we understand the characteristics of points of the studied statistical population to adequately reflect the characteristics of the trait under study. Representative sampling or a representative selection of points in the training subset provides, within a given accuracy, reliable data on the content of a pollutant in an environmental component (air, water, soil etc.) in a selected area at a given point in time.

We assume that not only points distinguish in different representativeness for the evaluation of a feature, but also sets of points (doublet, triplets, ..., *n*-lets) have different representativeness. In the present work, it is proposed to consider the comparison of individual and collective representativeness when points are included in the training subset. Under the individual representativeness of the sampling point, we mean the frequency of its hits in the training subset, training on which provides the smallest model error. Collective representativeness is the frequency of hits of a collection of points (pairs, triples, quadruples, etc.) in the training subset, training on which provides the smallest model error. To build a representative training subset, it is necessary to 1) determine which *n*-lets are the most representative (*n*-lets size and representativeness level); 2) reveal the relationship between individual and collective representativeness. Determination of the volume of all representative *n*-lets requires large computing power of the computer, so the collective representativeness in this work was evaluated by pair.

MATERIALS AND METHODS

Sampling location

200 soil samples were collected in the residential part of Noyabrsk city (N 63.2°, E 75.5°), Russia. The industry of the city is hydrocarbon energy. The climatic zone is subarctic, or Dfc, by Köppen climate classification. The predominant soil type is gley taiga (Gd 23-1ab) on the FAO-UNESCO soil map[1].

The sampling point map was designed so that a given number of samples (200 in the residential area of Noyabrsk) on average evenly filled the study area. The average distance between sampling points was about 300 m. This distance varied depending on the density of buildings, the location of roads, etc. in order to ensure sampling in undisturbed areas of the open soil surface. The sampling depth was 0.05 m, since we were interested in the top layer of soil. The sampling was carried out with a cylindrical sampler with a diameter of 0.05 m. The soil sampling procedure schema was shown in Fig. 1.

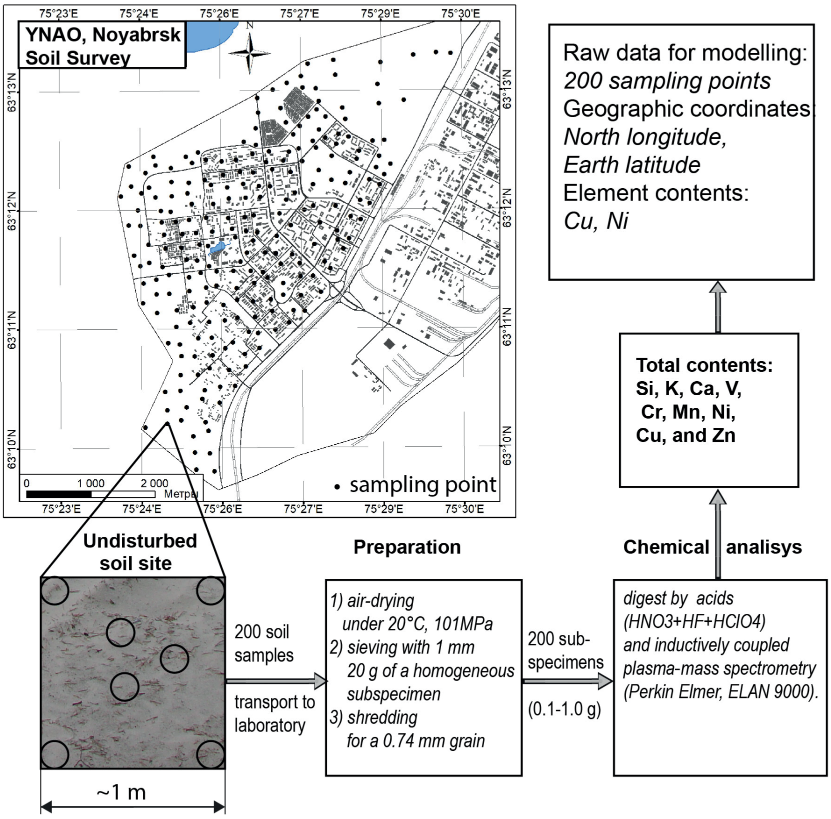

Fig. 1. The soil sampling procedure schema

[1] HWSD (Harmonized World Soil Database), 2009. Soil Units in the Revised Legend of the Soil Map of the World. https://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/ (accessed 15 May 2023).

The raw data preparation

The soil sampling procedure was previously described in detail for Noyabrsk (Baglaeva 2020). Fig. 1 shows the schema of this procedure. The raw data preparation consisted of sampling 200 specimens from undisturbed soil sites, sample preparation, and chemical analysis using inductively coupled plasma-mass spectrometry.

Preparation of soil samples and chemical analysis were conducted in compliance with actual standard requirements. For quality control, standard reference samples were used, certified for the content of determined elements, similar in composition to the samples under study.

The same total element contents were determined in the topsoil. Further modeling involved the total content of *Cuprum* and *Niccolum* in topsoil.

Multilayer perceptron

The input data are the geographic coordinates for the simulation. The output data are the element's contents.

The multilayer perceptron (MLP) with Levenberg-Marquardt learning algorithm was used to demonstrate the possibilities of the method as the easy-to-understand model ANN for modeling the spatial distribution of the element contents in the topsoil. The construction of the MLP model based on the number of neurons inside the hidden layer was chosen after several training cycles and error estimation for the test subset. We used the tangential activation function, which is best suited for predicting the features of the spatial distribution of element content in the topsoil (Baglaeva 2021). The MLP structure had one input layer consisting of two neurons (spatial coordinates x and y), one hidden layer with 9 neurons, and one output layer with one neuron (element content).

Representativeness assessment

Let the representativeness of the sampling point be an ability of this point to provide: 1) a small root-mean-square error *RMSE* (Eq. 1) for estimating accuracy; 2) a high correlation coefficient *Corr* (Eq. 2) to check the synchronism of changes between predicted and observed values with the participation of this point in training.

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

where *p(i)* is predicted data; *o(i)* is observed data; n is the number of subset points. *RMSE* (1) tests the accuracy between predicted and observed data.

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

where $\bar{p}$ is predicted average; $\bar{o}$ is observed average. The correlation coefficient *Corr* shows the linear statistical relationship between the predicted values and the observed ones, how much the changes in the predicted values repeat the systematic changes in the observed ones.

The individual representativeness for each point considers the set of the best (small *RMSE* (Eq. 1) and high *Corr* (Eq. 2)) networks in which the point participated in training. Collective representativeness is the representativeness of the sampling points, which support connections with the neighbors, which makes it possible to provide a small *RMSE* and a high *Corr* when participating in the training of a group of sampling points. Collective representativeness considers the various combinations of training sampling points, such as doublets, triplets, and so on.

Individual representativeness

A four-step (4-step) algorithm was used for individual representativeness assessment of the sampling points involved in the training subset (Fig. 2). The raw data were repeatedly divided randomly into training and test subsets in the ratio of 75%/25%, respectively. 200 points were randomly divided into 150 training and 50 test points. The number of divisions was 100,000.

1. The total raw data set was randomly divided 100,000 times into two non-overlapping sets, training and test subsets, in the ratio of 75%/25%, respectively. Thus, we got 100,000 training and 100,000 test subsets.
2. For each random division, 5 MLP networks were built additionally (500,000 MLP networks in total). For each trained network, the root-mean-square errors (*RMSE*) of the forecast of the training and test subsets were determined. The network with the minimum *RMSE* was chosen.
3. *RMSE* and *Corr* for the training, test, and general subsets were calculated for 100,000 better networks.
4. Each sample point was assigned a set of the best networks in which it participated in training. For each sample point, we calculated the basic statistics of *RMSE* and *Corr* for the training, test, and general subsets for the networks in which the point participated in training.

Individual representativeness was assessed by comparing mean *RMSE* and *Corr* values. The best representative point is the one whose inclusion in the training subset provides a lower mean *RMSE* and a higher mean correlation coefficient with the observed values.

Collective representativeness

To assess the collective representativeness of the sampling points of the training subset, a four-step (4-step) algorithm was also used (Fig. 2). We divided the training and test subsets into 75% and 25%, respectively, used the training results as a set of the best networks for each point, and calculated the corresponding distributions of *RMSE* and correlation coefficients for the training, test, and general subsets.

Collective (paired) representativeness was assessed using samples of two points out of two hundred. For each pair of sampling points, the basic statistics of *RMSE* for the training, test, and total subsets were calculated for networks in which both points participated in training. 1000 pairs from these pairs were selected with the lowest mean *RMSE* for the pair that fell into the training set, which corresponded to the 0.051 quantile. The collective representativeness of the sampling points was assessed by the number of its connections with other points within the 0.051 quantile (1000 point pairs) according to the average *RMSE* or within the 1 - 0.051 = 0.949 quantile (1000 point pairs) according to the average correlation coefficient. We built graphs for a cutoff threshold of 10 connections with other sampling points.

In this work, due to computational difficulties, we limited ourselves to pairs of sampling points. The hypothesis that the synergy effect exists was tested by comparing individual and paired representativeness to predict element content. For verification, we used the conditional distribution of the correlation coefficients and *RMSE* means (provided that a pair of sampling points fell into the best training subset).

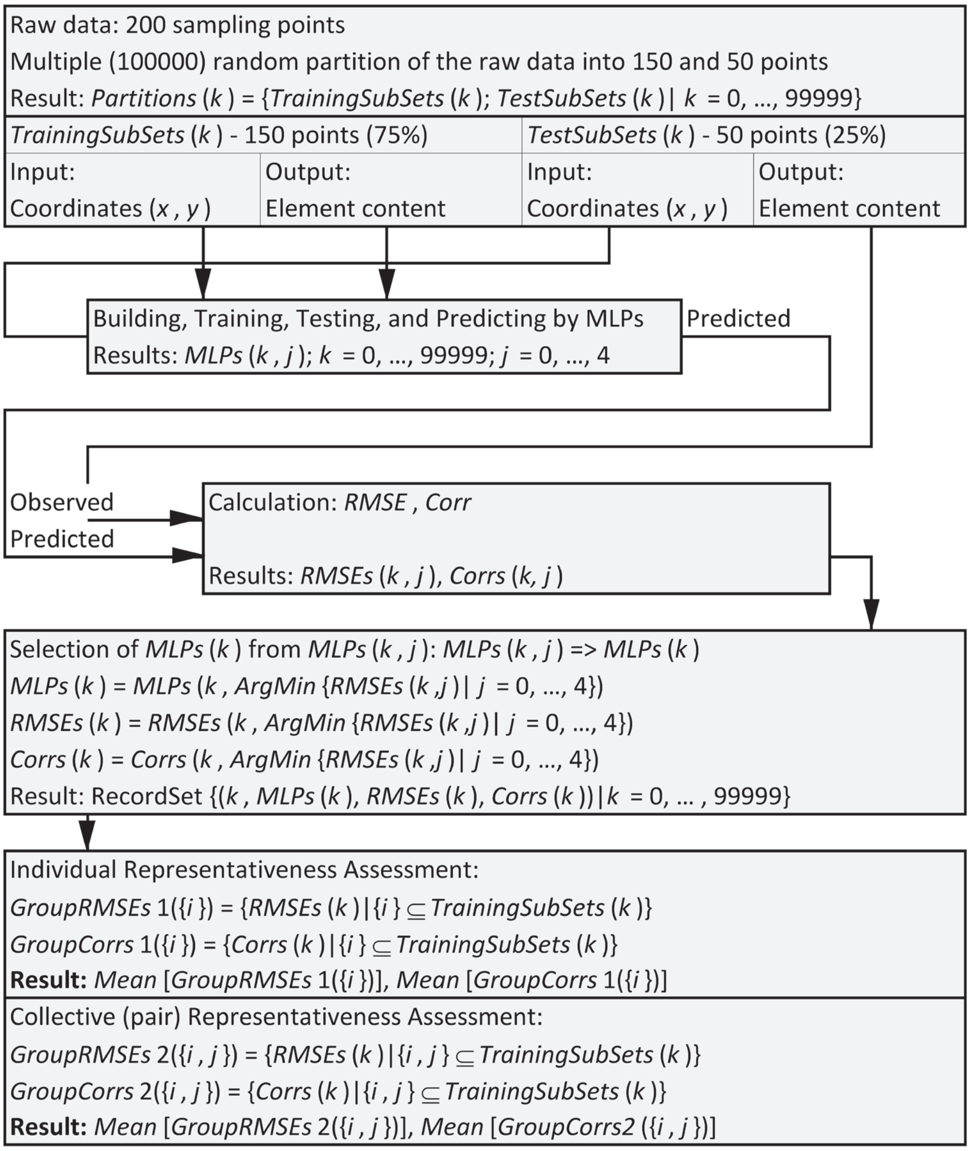
Raw data: 200 sampling points
Multiple (100000) random partition of the raw data into 150 and 50 points
Result: Partitions (k) = {TrainingSubSets (k); TestSubSets (k)| k = 0, ..., 99999}
TrainingSubSets (k) - 150 points (75%)
TestSubSets (k) - 50 points (25%)
Input: Coordinates (x , y)
Output: Element content
Input: Coordinates (x , y)
Output: Element content
Building, Training, Testing, and Predicting by MLPs
Results: MLPs (k , j); k = 0, ..., 99999; j = 0, ..., 4
Predicted
Observed
Predicted
Calculation: RMSE , Corr
Results: RMSEs (k , j), Corrs (k, j)
Selection of MLPs (k) from MLPs (k , j): MLPs (k , j) => MLPs (k)
MLPs (k) = MLPs (k , ArgMin {RMSEs (k ,j)| j = 0, ..., 4})
RMSEs (k) = RMSEs (k , ArgMin {RMSEs (k ,j)| j = 0, ..., 4})
Corrs (k) = Corrs (k , ArgMin {RMSEs (k ,j)| j = 0, ..., 4})
Result: RecordSet {(k , MLPs (k), RMSEs (k), Corrs (k))|k = 0, ... , 99999}
Individual Representativeness Assessment:
GroupRMSEs 1({i }) = {RMSEs (k)|{i } ⊆ TrainingSubSets (k)}
GroupCorrs 1({i }) = {Corrs (k)|{i } ⊆ TrainingSubSets (k)}
Result: Mean [GroupRMSEs 1({i })], Mean [GroupCorrs 1({i })]
Collective (pair) Representativeness Assessment:
GroupRMSEs 2({i , j }) = {RMSEs (k)|{i , j } ⊆ TrainingSubSets (k)}
GroupCorrs 2({i , j }) = {Corrs (k)|{i , j } ⊆ TrainingSubSets (k)}
Result: Mean [GroupRMSEs 2({i , j })], Mean [GroupCorrs2 ({i , j })]

Fig. 2. The representativeness assessment algorithm

RESULTS

Table 1 presents the characteristics of *Cuprum* and *Niccolum* distributions in the study area. The total Cuprum content is in the range from 5.89 to 69.59 mg/kg, *Niccolum* from 3.58 to 41.94 mg/kg, which does not exceed Clarke in the urban soil (Shichkin 2018).

For each data split, MLP models were built, and representativeness characteristics were calculated: *RMSE* and correlation coefficients. Table 2 presents the statistical characteristics of the representativeness assessment. Fig. 3 shows the *RMSE* means and correlation coefficients obtained for 19,900 models.

Table 1. Element statistics in Noyabrsk topsoil

Element	Element Content, mg/kg					*CV*, %	Skewness	Excess Kurtosis
	Minimum	Maximum	Mean	*SD*)*	Median			
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

*) *SD* – standard deviation; **) *CV* – coefficient of variation.

Table 2. Representativeness characteristic statistics

Element	Characteristic	Mean	Median	Minimum	Maximum	*SD*	*CV*, %	Skewness	Excess Kurtosis
Cu	*RMSE*, mg/kg	3.325	3.336	2.771	3.395	0.059	2	-5.5	39
	Corr	0.256	0.256	0.236	0.271	0.003	1	-0.3	2
Ni	*RMSE*, mg/kg	4.312	4.322	3.720	4.340	0.049	1	-7.2	58
	Corr	0.285	0.285	0.271	0.302	0.004	1	0.2	1

As can be seen from Fig. 3, the *RMSE* distribution is split into two clusters: *Cuprum* and *Niccolum*. The lower *RMSE* cluster is associated with the inclusion in the training subset of a single point 129 for *Niccolum* (the point with the highest *Niccolum* content) and 134 for *Cuprum* (the point with the highest *Cuprum* content).

Pair representativeness was visualized as *RMSE* and correlation coefficient graphs for *Niccolum* and *Cuprum* contents (Fig. 4). The sampling points are graph vertices. The graph edges were the best links between pairs of sampling the points named by doublet.

Pair representativeness graphs were built by the least *RMSE* doublets and the largest *Corr* doublets for Ni and Cu (Fig. 4). When constructing the graphs, we were limited to about 20 doublets. That is, the correlation graphs included the doublets with correlation coefficients greater than 0.2667 for Cu and 0.2983 for Ni. The *RMSE* graphs consisted of the doublets with *RMSEs* less than 2.83 for Cu and 3.868 for Ni. Both *Niccolum* and *Cuprum* *RMSE* graphs seem to have a scale-free network structure. The edges of the correlation coefficient graphs are stitching through the area for each element. *RMSE* as the representativeness indicator define the most "polluted" points, and the correlation coefficient adds "important" points for description of the element distributions in the topsoil. Individual and pairwise representativeness comparisons were shown for *Niccolum* and *Cuprum* contents in Table 3.

DISCUSSION

Individual representativeness is not enough to determine the best training subset. A point with low individual representativeness in pairs may be "good" for learning. By analogy with individual characteristics, there can be "good" pairs for training (these are all 199 pairs with 129 points for *Niccolum*, and 134 points for *Cuprum*), there can be "bad" ones, which increase the model error when included in pairs in the training subset. The results are shown in Table 3. 9 points were selected: 8, 12, 14, 63, 67, 116, 165, 168, 199 for *Cuprum* and 49, 52, 84, 102, 103, 104,

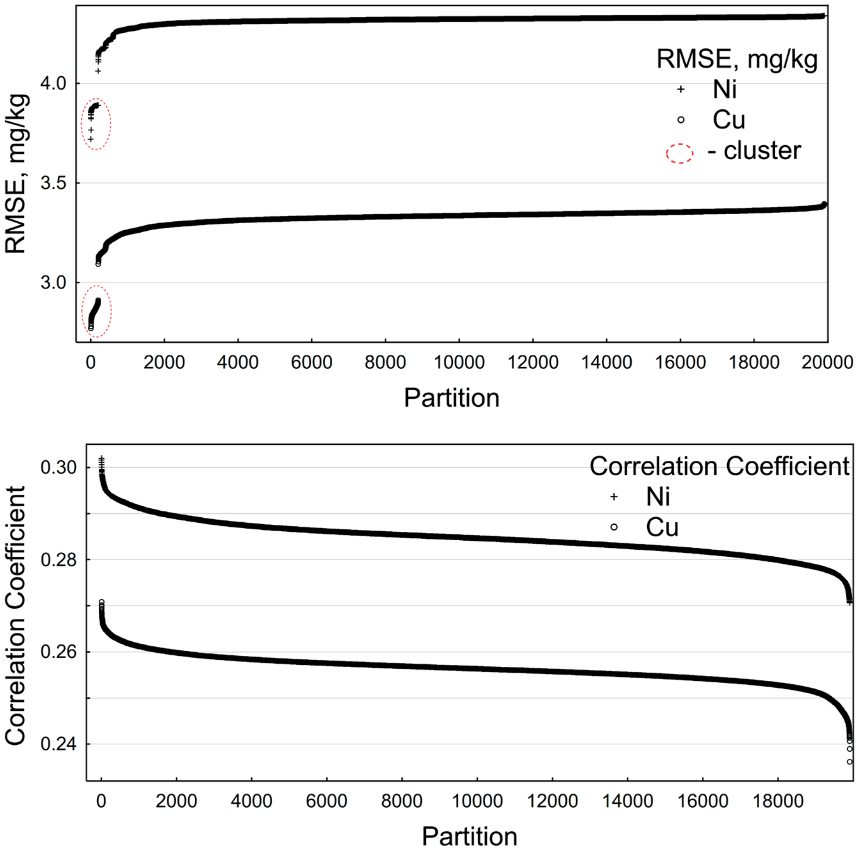

Fig. 3. *RMSE* mean and correlation coefficients mean

Table 3. Individual and pairwise representativeness comparison

Characteristic	«Best» *RMSE* doublet			«Bad» *RMSE* doublet			«Best» *Corr* doublet			«Bad» *Corr* doublet		
Element	Ni (*Niccolum*)											
n	{94;129}	{94}	{129}	{19;56}	{19}	{56}	{13;25}	{13}	{25}	{88;145}	{88}	{145}
RMSE, mg/kg	**3.720**	**4.167**	**3.880**	4.340	4.325	4.326	4.234	4.266	4.281	4.155	4.166	4.304
Corr	0.281	0.278	0.287	0.284	0.284	0.284	**0.302**	**0.294**	**0.293**	0.271	0.278	0.278
Element	Cu (*Cuprum*)											
n	**{134;135}**	**{134}**	**{135}**	{139;165}	{139}	{165}	{22;135}	{22}	{135}	{126;127}	{126}	{127}
RMSE, mg/kg	2.771	2.859	3.319	3.395	3.354	3.358	3.226	3.253	3.319	3.152	3.284	3.148
Corr	0.27	0.259	0.264	0.254	0.258	0.253	**0.271**	**0.263**	**0.264**	0.236	0.248	0.246

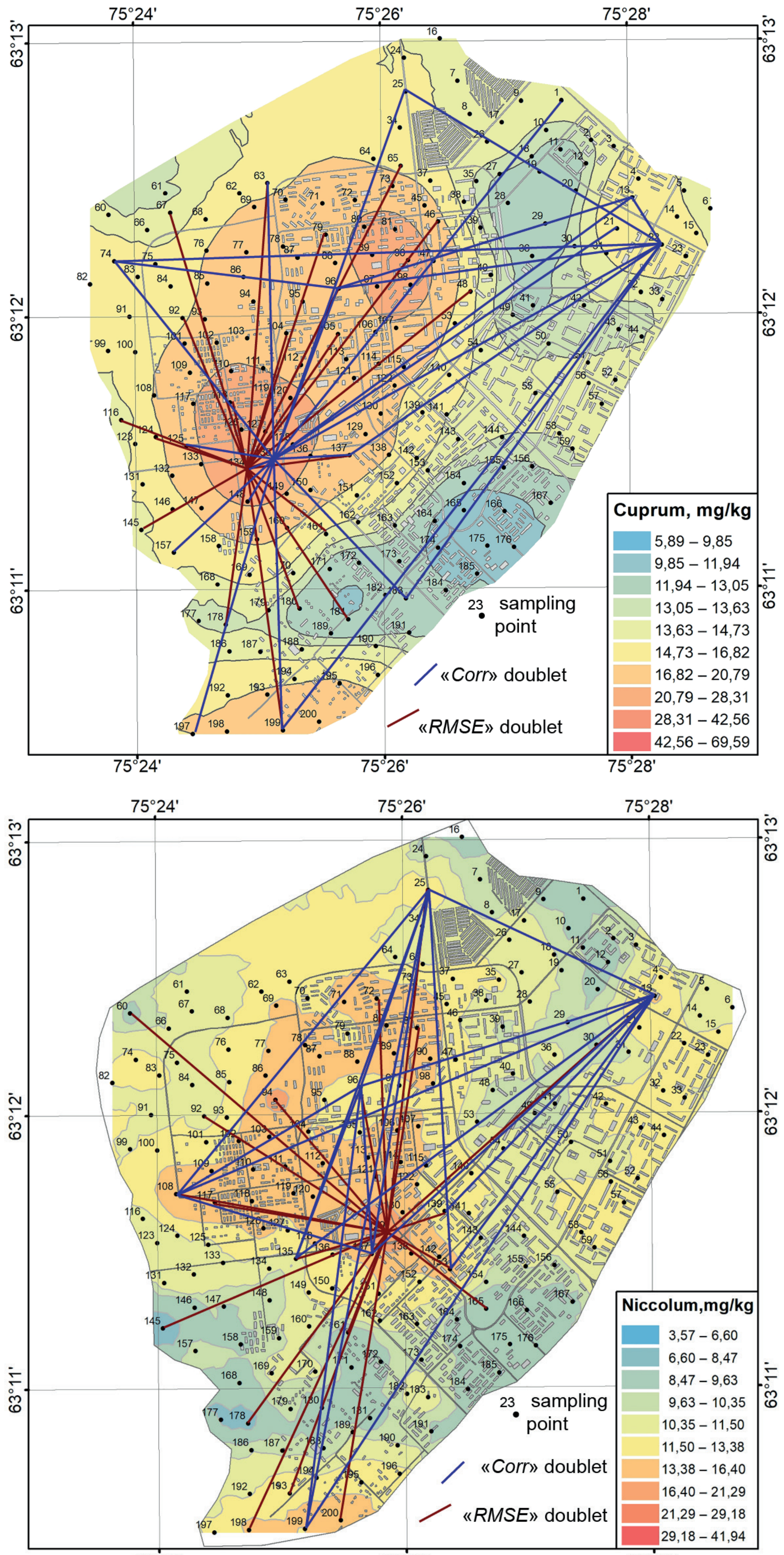

Fig. 4. ***RMSE*** **and correlation coefficient graphs of the pair representativeness**

118, 146, 192 for *Niccolum*. The elemental content at these points is below or close to the average.

Pair representativeness is not always limited to individual. There is a synergy effect, i.e., taking into account the collective (*n*-let) representativeness makes it possible to reduce the model error. Paired (collective) representativeness characterizes the interaction of pairs of points, i.e., the ability of pairs of points, when included in the training subset, to provide *RMSE* and correlation coefficient characteristics that exceed the best individual characteristics. The value added by the pair {94; 129} for *Niccolum* and a pair of {134; 135} for *Cuprum* is higher than the value contributed by individual points {94} and {129} for *Niccolum* and {134} and {135} for *Cuprum*. This synergy effect is created through the mutual influence between the points. Table 3 shows the best pairs with the lowest *RMSE* and the highest correlation coefficient and the worst pairs with the highest *RMSE* and the lowest correlation coefficient for Cu and Ni. Along with the synergy effect, there are relationships that reduce the value of the model. In this case, pairs of points provide less information to describe the element content distribution than individual points included in a pair as the effect of antisynergy. This may be due to the redundant use of points to describe the distribution of the feature.

As can be seen from Table 3, for example, this is {19; 56} for Ni and {139; 165} for Cu. *RMSE* pair {19; 56} for *Niccolum* and {139; 165} for *Cuprum* is greater than the *RMSE* of individual points {19} and {56} for *Niccolum* and {139} and {165} for *Cuprum*. Conversely, the correlation coefficients of a pair {88; 145} for Ni and {126; 127} for Cu are smaller than the correlation coefficients of individual points {88} and {145} for Ni and {126} and {127} for Cu.

The effects found here (synergy and anti-synergy) seem to be useful for predicting spatial variability and predicting the content of elements in the topsoil in areas with complex geographical conditions. This benefit may be expressed as a reduction in the uncertainty of the results of future field studies when they are planned.

The scale-free network structures *RMSE* graphs of the pair's representativeness are the same for both *Niccolum* and *Cuprum* (Fig. 4), and the central points of these graphs are territorial characteristics. For each pair of the points, the best graph topology characteristic of the territory is identified. This topology can be explained by man-made activity.

The obtained results do not contradict the hypothesis that different locations (geolocations) carry different information and an unequal value for the interpolation of the feature distribution. Evaluation of the representativeness of the points will allow you to choose the most representative points for the areas.

CONCLUSIONS

Comparison of individual and pair (collective) representativeness when points were included in the training subset showed their unequal value for interpolating the distribution of heavy metals in the topsoil. The most representative in terms of individual representativeness were the points with the maximum element content in the selected area. Including these points in the ANN training subset reduces the error and increases the correlation between the results of model calculations and field measurements on the test subset. The graph topology of the best collective representativeness (it looks like a constellation) can be used as territory characteristics associated with man-made activity. The volume of the *n*-let can be analogous to the dimension of the phase space. Although it is impossible to predict every detail of the evolution of such a system, it is possible to develop statistical mechanics with heterogeneous ensembles of interacting agents (Levin 2002), similar to the description of statistical ensembles in gas dynamics.

In this work, we have limited ourselves to pair representativeness; determining the volume of all representative *n*-s requires huge computational costs and remains a task for future research. Complex adaptive systems are limited in their predictability because multiscale interactions and evolutionary processes are linked through non-linear interactions. ■

REFERENCES

Baglaeva E.M., Sergeev A.P., Shichkin A.V., Buevich A. G. (2020). The Effect of Splitting of Raw Data into Training and Test Subsets on the Accuracy of Predicting Spatial Distribution by a Multilayer Perceptron. Math. Geosci., 52, 111–121.

Baglaeva E.M., Sergeev A.P., Shichkin A.V., Buevich A.G. (2021). The Extraction of the training subset for the spatial distribution modelling of the heavy metal in topsoil. Catena, 207, 105699. https://doi.org/10.1016/j.catena.2021.105699

Boussange V., Pellissier L. (2022). Eco-evolutionary model on spatial graphs reveals how habitat structure affects phenotypic differentiation. Communications Biology, 5, 668. https://doi.org/10.1038/s42003-022-03595-3

Dale M.R. and Fortin M. (2010). From graphs to spatial graphs. Annu. Rev. Ecol. Evol. Syst., 41, 21–38.

Demyanov V., Gloaguen E., Kanevski M. (2020). A special issue on data science for geosciences. Math. Geosci., 52, 1–3.

Fernandez Jaramillo J.M. and Mayerle R. (2018). Sample selection via angular distance in the space of the arguments of an artificial neural network. Computers and Geosciences, 114, 98–106.

Gutierrez-Velez V.H. and Wiese D. (2020). Sampling bias mitigation for species occurrence modeling using machine learning methods. Ecological Informatics, 58, 101091. https://doi.org/10.1016/j.ecoinf.2020.101091.

Kramm T. and Hoffmeister D. (2020). 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, 22, e00316. https://doi.org/10.1016/j.geodrs.2020.e00316.

Legendre P., Dale M.R.T., Fortin M.J., et. al. (2004). Effects of spatial structures on the results of field experiments. Ecology, 85(12), 3202–3214.

Levin S.A. (2002). Complex adaptive systems: Exploring the known, the unknown and the unknowable. Bull. Am. Math. Soc., 40, 3–20.

Liu Q., Li H., Guo L., et.al. (2022). Digital mapping of soil organic carbon density using newly developed bare soil spectral indices and deep neural network. Catena, 219, 106603. https://doi.org/10.1016/j.catena.2022.106603.

Malof J.M., Reichman D., Collins L.M. (2018). 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. 10628, Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIII, 106280C. https://doi.org/10.1117/12.2305793

de Mello D.C., Ferreira T.O., Veloso G.V., et. al. (2022). Pedogenetic processes operating at different intensities inferred by geophysical sensors and machine learning algorithms. Catena, 216, Part A, 106370, ISSN 0341-8162. https://doi.org/10.1016/j.catena.2022.106370.

Nath A. and Subbiah K. (2018). 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, 272, 294–305.

O'Brien D., Manseau M., Fall A., Fortin M-J. (2006). Testing the importance of spatial configuration of winter habitat for woodland caribou: An application of graph theory. Biological Conservation, 130, 70–83.

O'Hare M.T., Gunn I.D.M., Critchlow-Watton N., et. al. (2020). Fewer sites but better data? Optimising the representativeness and statistical power of a national monitoring network. Ecological Indicators, 114, 106321. https://doi.org/10.1016/j.ecolind.2020.106321.

Pesch R., Schröder W., Dieffenbach-Fries H., et. al. (2008). Improving the design of environmental monitoring networks. Case study on the heavy metals in mosses survey in Germany. Ecological Informatics, 3(1), 111 121. https://doi.org/10.1016/j.ecoinf.2007.11.001.

Prager S.D. and Reiners W.A. (2009). Historical and emerging practices in ecological topology. Ecological complexity, 6, 160–171. doi:10.1016/j.ecocom.2008.11.001

Shichkin A.V., Buevich A.G., Sergeev A.P., et. al. (2018). Prediction of the content of anomalously distributed chromium in the soil by hybrid models based on artificial neural networks. Geoecology. Engineering geology. Hydrogeology. Geocryology, 3, 86 96. [in Russian]

Urban D.L., Minor E.S., Treml E.A., Schick R.S. (2009). Graph models of habitat mosaics. Ecology Letters, 12, 260–73.

Wang I.J. and Bradburd G.S. (2014). Isolation by environment. Molecular Ecology, https://doi.org/10.1111/mec.12938

Wang X., An Sh., Xu Y., et. al. (2020). 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, 10(51), 1 17. https://doi:10.3390/app10010051

Wang Y., Ma H., Wang J., et. al. (2021). Hyperspectral monitor of soil chromium contaminant based on deep learning network model in the Eastern Junggar coalfield. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 257, 119739. https://doi.org/10.1016/j.saa.2021.119739.

Xu Q., Zhu A-X., Liu J. (2023). Land-use change modeling with cellular automata using land natural evolution unit. Catena, 224, 106998. https://doi.org/10.1016/j.catena.2023.106998.

Zhu A.X., Liu J., Du F., et.al. (2015). Predictive soil mapping with limited sample data. European Journal of Soil Science. 66, 535–547. doi: 10.1111/ejss.12244

Zhu A.X., Lu G., Liu J., et.al. (2018). Spatial prediction based on Third Law of Geography. Annals of GIS, 24 (4), 225–240. https://doi.org/10.1080/19475683.2018.1534890

Zhu A.X., Lv G.N., Zhou C.H., et al. (2020). Geographic similarity: Third Law of Geography? Journal of Geoinformation Science, 22(4), 673–679. https://doi.org/10.12082/dqxxkx.2020.200069.

Zhong L., Guo X., Xu Zh., Ding M. (2021). Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. Geoderma, 402, 115366.

Ziggah Y.Y., Youjian H., Tierra A.R., Laari P.B. (2019). Coordinate Transformation between Global and Local Data Based on Artificial Neural Network with K-Fold Cross-Validation in Ghana. Earth Sciences Research Journal, 23(1), 67 77. https://doi.org/10.15446/esrj.v23n1.63860.