Chromium Distribution Forecasting in Subarctic Noyabrsk Using Cokriging, Generalized Regression Neural Network, Multilayer Perceptron, and Hybrid Technique

Alexander G. Buevich[1,2] (ORCID: 0000-0003-4964-2787),
Alexander P. Sergeev[1,2] (ORCID: 0000-0001-7883-6017),
Andrey V. Shichkin[1,2] (ORCID: 0000-0002-0081-1853),
Alexandra I. Kosachenko[1] (ORCID: 0000-0001-8896-3837),
and Anastasia S. Moskaleva[1] (ORCID: 0000-0002-1570-6642)

[1] Ural Federal University, Mira str., 19, Ekaterinburg, RUSSIA 620002
corresponding author: bagalex3@gmail.com
[2] Institute of Industrial Ecology UB RAS, S. Kovalevskoy str., 20, Ekaterinburg, RUSSIA 620990

Abstract. Combination of geostatistical interpolation techniques (*e.g.* kriging) and machine learning (*e.g.* neural networks) leads to better prediction accuracy and productivity. Application of an artificial neural network residual kriging (ANNRK) for spatial prediction of soil contamination with Chromium (Cr) is considered in the paper. We examined and compared two neural networks: Generalized Regression Neural Network (GRNN) and Multilayer Perceptron (MLP). We consider them as classes of neural networks widely used for the continuous function mapping, as well as a combined technique Multilayer Perceptron Residual Kriging (MLPRK). The case study is based on the survey on surface contamination by Cr at the subarctic city Noyabrsk, Russia. Structures of used models have been developed using a computer simulation based on a minimization of the RMSE. Each technique has its own benefits and drawbacks; however both demonstrated fast training and good prediction possibilities. The MLPRK showed the best predictive accuracy.

Keywords: Artificial Neural Networks, Chromium, Residual kriging, Cokriging, GRNNRK, MLPRK

1 Introduction

Methods for predicting the spatial distribution of impurities on the basis of sampling (monitoring, screening) are important part of assessing the state of the environment. In the case when the medium is highly heterogeneous, such methods have advantages over deterministic ones that require a large number of input variables.

Kriging methods have been frequently used for soil and sediments properties prediction [2], [5]. Cokriging is a multivariate kind of the ordinary kriging

(OK), which calculates estimates for a poorly sampled element with help of well-sampled highly correlated elements (co-elements) [11].

Accuracy of kriging depends on the density and size of the sample grid, since the method is based on interpolation. However, it is not always possible to collect the required number of samples due to time-related or resource-related constraints. To overcome these shortcomings and improve the accuracy, a more effective method is required.

Currently, one of the applicable methods of prediction is the machine learning and, in particular, the artificial neural networks (ANN), which provide many powerful techniques for predicting, pattern recognition, data analysis, and many other operations.

Overviews [1], [6] showed the high versatility of the ANNs. Recently, this method is widely used in handling environmental issues [8]. The ANN models successfully predict the pollutants content at unmonitored locations [14].

The most frequently used ANN in environmental studies is the multi-layer perceptron (MLP) and generalized regression neural networks (GRNN). Perceptrons are widely used in studies on soil chemical elements content assessment [9], [4], [2]. Many researchers have explored perceptrons for resource estimation [12], [13] and most of them proved the superiority of the MLP over the geostatistical and deterministic methods.

The GRNNs are variations of the radial basis functions (RBF) neural networks, which are based on the kernel regression networks. The GRNNs are used as interpolators and are known as universal function approximators, which can approximate any continuous nonlinear function. The key difference between GRNNs and MLPs is that the GRNNs do not require the learning process using long-term iterative procedures as back propagation networks (MLPs *etc.*).

To neutralize weaknesses and to multiply dignities of the mentioned models, it has been offered to combine different techniques. Researchers successfully exploit the hybridization of geostatistical and neural approaches, which lead to better predictions and lower errors [3], [7], [15].

In this work, we examine two solo ANN-based prediction models (MLP, GRNN), as well as a hybrid model combining the ANN based forecasting and cokriging (MLPRK) for prediction a soil pollutant Cr content at a particular location of the Subarctic Noyabrsk, Russia. We examine the results obtained by applying the models and compare the models output.

2 Materials and Methods

2.1 Study Area

Data for the study were obtained from the results of the soil survey in Noyabrsk (N63.1926°, E75.5066°) and Yamalo-Nenets Autonomous Okrug (YNAO), Russia (see Fig. 1(a)). The area of sampling was approximately 16.5 km^2. The detailed spatial location of sampling points is shown in Fig. 1(b). The terrain was flat and covered with sandy soil (Cryosols soil type). Totally, 237 topsoil samples

at a depth of 0.05 m were collected. Concentration indicators for the elements were obtained by chemical analysis.

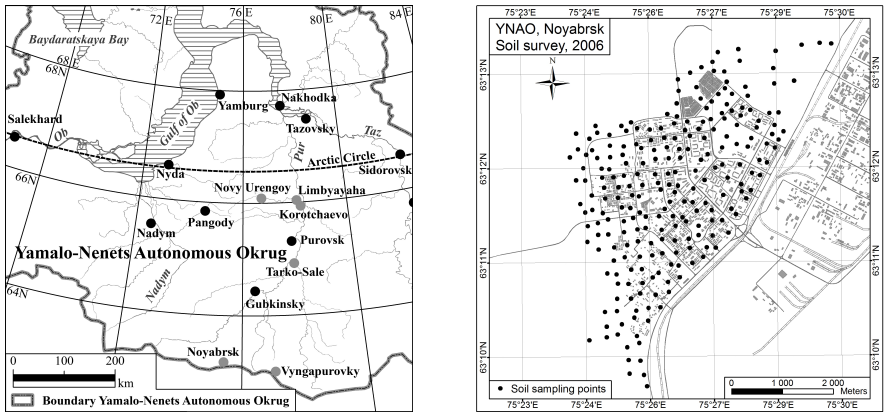


Fig. 1. The sampling area: (a) Yamalo-Nenets Autonomous Okrug, Russia; (b) Noyabrsk city (dots are sampling points)

2.2 Chemical Analysis

Preparation of the soil specimens and chemical analysis were conducted in compliance with actual standard requirements. The chemical laboratory involved with the soil sample preparation and analysis passed through the Russian Federal Certification System.

2.3 Spatial Prediction of Cr Content by ANN

Study Algorithm. To estimate the pollutant content and to predict its distribution at the unknown locations, three competing techniques was applied: two ANN methods (MLP and GRNN), as well as a hybrid ANN-geostatistical model Multi-layer Perceptron Residual Kriging (MLPRK).

The MLPRK is a three-step algorithm combining two different interpolation techniques in one ensemble. The first step implies estimating large-scale nonlinear trends using neural networks (MLP). The second step is analysis of the stationary residuals by ordinary kriging (exponential model), which is able to provide local estimates. The final step is estimation produced as a sum of the ANN predictions and ordinary kriging estimates of the residuals. In the work, the ANN predictions were carried out in MATLAB; the ArcGIS application was performed to predict the values by kriging.

Since further analysis implies the use of the ANN, a method of selecting input variables (IVS) based on the estimation of partial mutual information (PMI) was used [10].

All the samples were randomly split into independent training and test data sets. The training data set (165 samples) was used for building cokriging, training the networks, and for interpolating the surface pollutant distribution. The test data set (72 samples) was used for testing the models only.

Data Isotropy. Values of the experimental semivariogram were calculated depending on the distance between the points in the pair within the lag h. The lag was selected according to the size of the spatial correlation. With this approach, the semivariogram value did not depend on the orientation of the pair in space, which means isotropy of the structure. The variogram constructed in all directions (omnidirectional) depends on all pairs of points in the domain. To identify differences in spatial structure depending on the direction the experimental variograms in various directions were applied.

Building ANNs. As first ANN type, a feed-forward multi-layer perceptron (MLP) with the Levenberg -Marquardt training method was used. The network structure was determined during computer simulation. The input layer of MLP was compiled with sampling points; the hidden layer consists of a several neurons, and the output layer represents the element concentration in the relevant sample. Selection of the number of neurons in the hidden layer was carried out by the lower total root mean squared error (RMSE) (5) of prediction of the element (Cr) concentration for the training (165 samples), test (72 samples), and a complete set of data (237 samples). The number of neurons was varied from 2 to 20. Each network was trained 500 times and the best of them was selected. The network education quality was checked by the Spearmans correlation coefficient, mean absolute error (MAE) (4), and RMSE between the results of the network predictions and the training data set.

As the second ANN, the GRNN was chosen. The first layer in the GRNN resembles the RBF with the amount of neurons that is equivalent to the quantity of input vectors. Choice of the SPREAD parameter of the RBF, that is known as a smoothing parameter, determines the width of the input area, to which each basis function responds. It is the distance from the center of a Gaussian where the value is one-half of the peak value. The GRNN network had 165 input neurons according to 165 sampling points formed the training data set. During the simulation, the SPREAD parameter varies from 0.01 to 0.30 with step 0.01; totally, 300 simulations were done.

2.4 Residuals Estimation by the Ordinary Kriging

The starting procedure for the residual kriging is the prediction of residual values by the neural network in the test points. Residuals in the neural network can be defined as follows:

$$r(x_i) = Z(x_i) - Z_{\mathrm{ANN}}(x_i), \tag{1}$$

where $r(x_i)$ are the residuals of data set (x_i), $Z(x_i)$ are the measured values, $Z_{\mathrm{ANN}}(x_i)$ are the values estimated by the neural network. The resulting residuals

were estimated using kriging. Evaluation in ordinary kriging (OK) is constructed as a linear combination of input data

$$r_{\mathrm{OK}}(x) = \sum \lambda_i r(x_i), \tag{2}$$

where r_{OK} is the estimated value at the point x using OK, $\lambda_i(x)$ are the optimal weights with the condition $\sum \lambda_i = 1$, and $r(x_i)$ is the residual of the neural network for the point (x_i). The final evaluation of the pollutant content $Y(x_i)$ was obtained as the sum of the neural network evaluation and residuals evaluation by kriging

$$Y(x_i) = Z_{\mathrm{ANN}}(x_i) + r_{\mathrm{OK}}(x_i). \tag{3}$$

2.5 Evaluation of Interpolation Accuracy

The performance of prediction models was based on the model error statistics. The predictive accuracy of each selected approach was verified by the Spearmans rank correlation coefficient *r, MAE* (4) and *RMSE* (5) between the prediction and raw data from the training data set.

$$MAE = \frac{\sum_{i=1}^{n} |z_{\mathrm{mod}}(x_i) - z(x_i)|}{n}, \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (z_{\mathrm{mod}}(x_i) - z(x_i))^2}{n}}, \tag{5}$$

where $z_{\mathrm{mod}}(x_i)$ is a predicted concentration (ANNs, cokriging), $z(x_i)$ is a measured concentration, n is a number of points.

3 Results

3.1 Descriptive Statistics of the Content

During the analysis, the contents of eight elements were obtained (Al, Cr, Mn, Fe, Co, Ni, Zn, Pb). Correlation analysis (Table 1) revealed possible co-elements for Cr (in bold). Three of them (Fe, Co, Ni) we used for the cokriging.

Table 1. Correlation matrix of element contents

	Al	**Cr**	**Mn**	**Fe**	**Co**	**Ni**	**Zn**	**Pb**
Al	1							
Cr	-0.01	1						
Mn	0.39	0.54	1					
Fe	0.16	**0.83**	0.69	1				
Co	0.04	**0.65**	0.59	0.64	1			
Ni	-0.01	**0.72**	0.60	0.75	0.70	1		
Zn	0.34	0.10	0.35	0.18	0.23	0.36	1	
Pb	0.34	0.07	0.14	0.07	-0.03	0.17	0.50	1

The descriptive statistics of Cr and its co-elements concentrations are shown in the Table 2.

Table 2. Descriptive statistics of the modeled element (Cr) and co-elements (Fe, Co, Ni), mg/kg

Element	**Min**	**Max**	**Mean**	**SD**	**CV**	**SK**	**RKu**	**MED**
Cr	16.6	140	62.4	24.2	0.388	0.805	0.438	58.8
Fe	4751	28270	13268	4397	0.331	0.786	0.748	12673
Co	2.00	11.4	4.50	1.49	0.332	0.989	2.13	4.42
Ni	3.58	41.9	11.3	4.52	0.398	2.23	12.0	11.0

SD is a standard deviation; CV is a coefficient of variation; SK is a Skewness; RKu is a Kurtosis; MED is a Median.

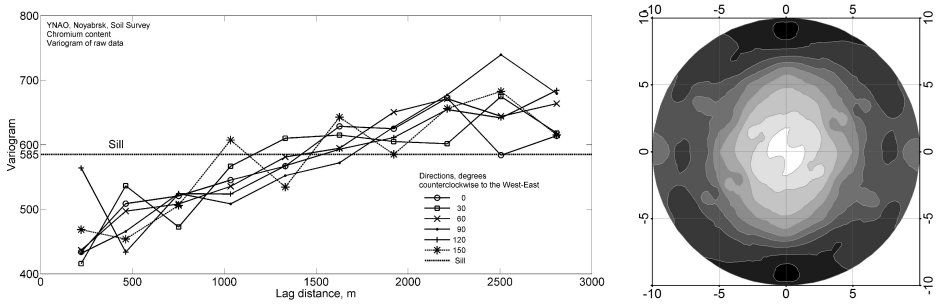
From the basic statistics table, it is observed that the element attributes are erratic and positively skewed. The Cr concentrations in all sampling points were from 16.6 to 140 mg/kg, with an average value of 62.4 mg/kg and a standard deviation of 24.2 mg/kg. Due to the skewness of the distribution, the median value (58.8 mg/kg) is more representative of the average Cr content in the study area than the arithmetic mean. Co-elements demonstrate similar characteristics when the medians (12673 mg/kg for Fe, 4.4 mg/kg for Co, 11.0 mg/kg for Ni) are more representative than mean values.

3.2 Spatial Prediction of Cr Concentration

The probability distribution of the Cr concentration for the training sites is positively skewed and leptokurtic (Table 1). The result of the Chi-Square test shows that this variable is close to normal distribution (p=0.18).

To demonstrate differences in the spatial structure depending on the direction, variograms are constructed in six directions (0°, 30°, 60°, 90°, 120° and

150°) (Fig. 2(a)). The anisotropy of the raw data in all these directions is invisible on the direction variograms (Fig. 2(a)) and variogram surface (Fig. 2(b)).

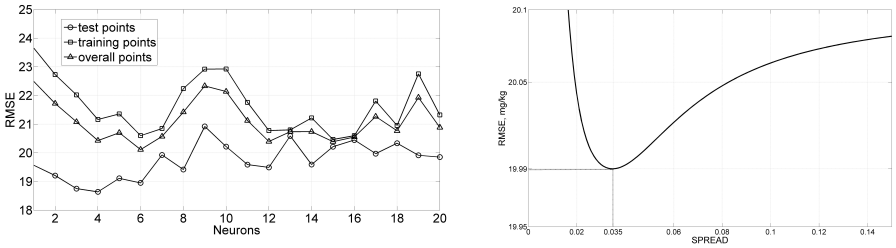


(a) (b)

Fig. 2. Variograms in six directions (a); variogram surface for Cr concentration (b)

The final configuration of the MLP network selected was 2-6-1, *i.e.* the hidden layer contains 6 neurons (see Fig. 3(a)). In our case, 165 sampling points formed the training data set that was applied for networks training.

During the simulation for GRNN building, the SPREAD parameter varies from 0.01 to 0.30 with step 0.01; totally, 300 simulations were done. The minimal RMSE was achieved with the SPREAD parameter of 0.035 (see Fig. 3(b)).



(a) (b)

Fig. 3. MLP (a) and GRNN (b) frameworks selection based on RMSE minimization: root mean square error (RMSE) of the neural network for test, training, and overall data under different neuron number in the hidden layer for Cr

3.3 Assessment of Accuracy of the Interpolation Methods

Table 3. Accuracy assessment indices of the (Cr) concentration

Method	SRCC	RMSE, mg/kg	MAE, mg/kg
Cokriging	0.15	20.7	15.3
MLP	0.41	19.0	**14.7**
GRNN	0.09	20.0	15.0
MLPRK	0.42	**18.8**	14.8

Here, the SRCC is the Spearman's rank correlation coefficient.

MLP and MLPRK have shown significant increase in modeling accuracy comparing to a geostatistical method (cokriging) and even to GRNN. As Table 3 reveals, the MLP-based models had smaller RMSE than cokriging (9.5% improvement). MAE index of the MLP-based models are about 4% better than cokriging ones. The basic GRNN model demonstrated an unexpectedly low correlation coefficient. This means that the method cannot be applied to modeling in our case.

It is found that application of the hybrid approach (MLPRK) gives an increase in the accuracy of prediction, which corresponds to the previous suggestion [2].

4 Discussion

We compared approaches to modeling the spatial distribution of the chemical element concentrations in the surface layer of soil (the geostatistical technique, ANNs, and hybrid model of MLPRK). A quality of models prediction could be analyzed with the help of the test data set, which is not applied to training networks or kriging estimates. Directional variograms of raw data and a variogram surface of training data are shown in Fig. 4. They confirm absence of any anisotropy of data. Table 3 shows the statistical parameters used to assess the performance of the different methods (the best values demonstrated by MLP-based models are written in bold).

The MLPRK model reproduces the spatial structure of the Cr distribution quite well. This model has extracted structured information leaving out unexplained noise and local variability. This is shown on directional residuals variograms (Fig.4). The study of the residuals confirms importance of the variography for analysis and modeling of spatial data with using the neural network algorithms.

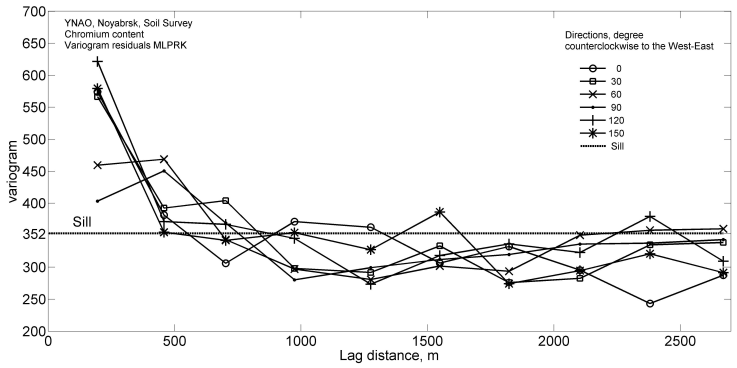


Fig. 4. Variograms for residuals in directions for MLPRK

5 Conclusion

Comparison of different approaches to prediction of the chemical elements distribution in the surface layer of soil is carried out. Estimation of the ANN with prediction of residuals by ordinary kriging reduces the ANN prediction errors, and increases accuracy of the models. The results show that the MLP-based models usually are more accurate than the kriging-based ones. In comparison with cokriging, the most significant improvement of RMSE (9.5%) is observed in the MLPRK model.

The results confirm possibility of the hybrid ANN-Kriging methods that can be used to improve the accuracy of modeling the spatial distribution of concentrations of chemical elements in the upper layer of soils in urban areas characterising by high heterogeneity. We assume that using (as input) not only spatial coordinates, but also additional variables will improve the predictive ability of ANN based models. This is since the variables have a significant correlation with the predicted variable, for example, the concentration of joint elements, geographic data, etc.

Acknowledgment. The reported study was funded by RFBR according to the research project N°18-55-18002.

References

1. Bishop, C.: Neural networks for pattern recognition. Clarendon, Oxford. 504p. (1995)
2. Dai, F., Zhoua, O., Lva, Z., Wang, X., Liu, G.: Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. Ecological Indicators. 45, 184–194 (2014)
3. Demyanov, V., Kanevsky, M., Chernov, S., Savelieva, E., Timonin, V.: Neural Network Residual Kriging Application for Climatic Data. Journal of Geographic Information and Decision Analysis. 2, 215–232 (1998)

4. Falamaki, A.: Artificial neural network application for predicting soil distribution coefficient of nickel. Journal of Environmental Radioactivity. 115, 6–12 (2013)
5. Forsythe, K. W., Marvin, C. H., Valancius, C. J., Watt, J. P., Aversa, J. M., Swales, S. J., Jakubek, D. J., Shaker, R. R.: Geovisualization of Mercury Contamination in Lake St. Clair Sediments. Journal of Marine Science and Engineering. 4(1), 19 (2016)
6. Graupe, D.: Principles of artificial neural networks. 2nd Ed. Advanced series of circuits and systems. World Scientific Publishing Co: Singapore (2007)
7. Lakes, T., Mller, D., Krger, C.: Cropland change in southern Romania: A comparison of logistic regressions and artificial neural networks. Landscape Ecology. 24(9), 1195–1206 (2009)
8. Leuenberger, M., Kanevski, M.: Extreme Learning Machines for spatial environmental data. Computers & Geosciences. Vol. 85, Part B, 64–73 (2015)
9. 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. 10, 1953–1960 (2011)
10. May, R. J., Maier, H. R., Dandy, G. D., Fernando, T. M.: Nonlinear variable selection for artificial neural networks using particle mutual information. Environmental Modelling & Software. 23(10–11), 1312–1326 (2008)
11. Meer, F. D.: Introduction to geostatistics. ITC Lecture Notes. 72p. (1993)
12. Samanta, B., Ganguli, R., Bandopadhyay, S.: Transactions of the Institution of Mining and Metallurgy. 114, 129–139 (2005)
13. Sergeev, A. P., Buevich, A. G., Medvedev, A. N., Subbotina, I. E., Sergeeva, M. V.: Artificial neural network and kriging interpolation for the chemical elements contents in the surface layer of soil on a background area. 15th International Multidisciplinary Scientific GeoConference SGEM 2015, Conference Proceedings (2015)
14. 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. 5(4), 1–20 (2014)
15. Tarasov, D. A., Buevich, A. G., Sergeev, A. P., Shichkin, A. V.: High Variation Topsoil Pollution Forecasting in the Russian Subarctic: Using Artificial Neural Networks Combined with Residual Kriging. Applied Geochemistry. Vol. 88, Part B, 188–197 (2017) https://doi.org/10.1016/j.apgeochem.2017.07.007