NEURAL NETWORK RESIDUAL STOCHASTIC COSIMULATION FOR ENVIRONMENTAL DATA ANALYSIS

V. Demyanov, M. Kanevsky, E. Savelieva, V. Timonin, S. Chernov, V.Polishchuk

Nuclear Safety Institute (IBRAE), Russia, 113191, Moscow, B. Tulskaya 52,

e-mail: vasia@ibrae.ac.ru, http://www.ibrae.ac.ru/~mkanev/

M. Maignan

University of Lausanne, Institute of Mineralogy and Petrology

Abstract

The present work introduces a new original method of stochastic simulation of environmental data – Neural Network Residual Sequential Gaussian Cosimulation (NNRSGCS). It is based on Artificial Neural Networks (ANN) and stochastic cosimulation of spatially distributed correlated variables. Multivariate ANN model – a multi-layer perceptron – is applied for modelling large-scale non-linear trend. Stochastic cosimulation is performed on the remaining residuals, which feature small-scale spatial correlation structure. The method is applied to the real data on radioactive soil contamination from the Chernobyl fallout.

Introduction

The problem of analysing environmental data is being solved nowadays with the help of different methods. Most of them fall into the following major approaches: deterministic, statistical and artificial intelligence. Statistical approach is based on considering the samples as realisations of a random process. This allows to apply a wide selection of linear and non-linear regression models (kriging family) and a variety of stochastic simulation algorithms (Gaussian, annealing). Geostatistics provides efficient tool for exploration of spatial correlation structure – variography. Possibility of joint consideration of correlated variables improves quality of the estimates/simulations in comparison with single variable estimation/simulation methods. Important advantage of geostatistical predictors/simulators is that they deliver exact estimation/simulation values at the sampling points.

Artificial neural networks (ANN) have been actively used for a wide set of problems nowadays. ANN are model-free estimators, which depend only on the data used for training and ANN architecture. Being applied for spatial estimates ANN produce fairly good results.

Combined model ANN+geostatistics is attractive, because of their high capacity in spatial modelling in comparison to of ANN and geostatistics applied separately. ANN fairy well model long-range correlation structures (trends) including non-linear ones. Geostatistics applied to the residuals produce prediction/simulation, honouring the sampling data, accompanied by estimation errors (for kriging) and possibility of development risk maps to support decision making process.

The application of NNRSGCS model is illustrated in the present work in a radioactive soil contamination case study with the data from the Chernobyl fallout. Accident at the Chernobyl Nuclear Power Plant occurred in 1986 and caused large-scale environmental consequences for vast regions in Europe. The contamination pattern originated after the release of the radioactive materials is highly variable and spotty due to the variety of influencing factors, like weather conditions, precipitation, surface characteristics (land use, orography, soil type, source term, etc.). Thus, complicated estimation/simulation methods have to be used to model such spatial pattern. Strontium 90 (^{90}SR) and caesium 137 (^{137}CS) from the fallout are

the most important radionuclides influencing the humans. They are significantly correlated in the considered area – western part of Briansk region, the most contaminated region in Russia. This fact allows to use more numerous and relatively cheap ^{137}CS samples in co-estimation/cosimulation ^{90}SR, which has fewer samples.

Neural Network Residual Stochastic Cosimulation

Neural Network Residual Sequential Gaussian Cosimulation (NNRSGCS) is a two step algorithm. At the first step ANN is used to model large-scale spatial structures (trend). The second step includes sequential Gaussian cosimulation performed on ANN residuals, obtained from the ANN estimates at the first step. Stochastic realisations produced by cosimulation describe spatial variability and uncertainty on the pattern. Lets look more closely into the original NNRSGCS algorithm.

1. The data is prepared for the analysis: split into training and validation set, checked for outliers, analysed with variography tools. The data is scaled on the interval [0.1, 0.9] for further use in ANN to facilitate the training procedure.

2. Feedforward multi-layer ANN (perceptron) is trained using back propagation algorithm. The ANN used for spatial prediction consists of 2 input neurones representing co-ordinates (x and y), hidden layer and 2 output neurones carrying correlated variable (^{90}SR and ^{137}CS) values. Such architecture allows to use additional information on the secondary correlated variable (^{137}CS) in order to improve estimates of the primary variable (^{90}SR). Using secondary variable sampling information as one more input neurone is restricted by the absence of the measurements at the estimation points (regular grid). Usually, one hidden layer is enough to model non-linear effects of the spatial pattern structure. The number of neurones in the hidden layer can vary. The optimal number of hidden neurones is chosen according to the best results obtained at the training points (accuracy test). It must be noted that over training is as dangerous as not sufficient training. Often, when there are two many neurones in the hidden layer, overtraining leads to modelling minor peculiarities, which represent not spatial correlation but noise. Accuracy test is ANN estimation at training points. It shows how well the ANN has been trained. Validation procedure – when ANN estimates values at validation points, which have not been used for training, – is a test of overall ANN performance, its ability to generalise and is especially used to avoid overtraining.

3. Accuracy test provides ANN residuals (*estimated - measured*) which are the base of the further analysis. Two cases are possible:

- residuals are not correlated with the measurements, which means, that ANN has modelled all spatial structures represented in the raw data;

- residuals show some correlation with the samples, than further analysis must be performed on the residuals to model this correlation.

The remaining spatial correlation represents short-range correlation structures. Long-range correlation (trend) in the whole area beyond the hot spots is very well modelled by ANN.

4. ANN residuals are explored using variography tools. First, normal score transformation is performed to prepare data for further Gaussian simulations. Data transformation is followed by bi-normality test, which allows to accept multi-normality hypothesis in a certain extent [8]. Structural analysis of normalised residuals showed presence of anisotropic structures.

5. Sequential Gaussian cosimulation algorithm requires generation of a random path through all nodes of the grid. The following procedure is performed at every node of the path [5]:

   - Simple cokriging estimate is calculated as follows:

$$Y_{SCK}^{(1)*}(x) - m_1 = \sum_{\alpha_1=1}^{n_1(x)} \lambda_{\alpha_1}^{SCK}(x)\left(Y_1(x_{\alpha_1}) - m_1\right) + \sum_{\alpha_2=1}^{n_2(x)} \lambda_{\alpha_2}^{SCK}(x)\left(Y_2(x_{\alpha_2}) - m_2\right)$$

where the weights λ are obtained from the following system of equations:

$$\begin{cases} \sum_{\beta_1=1}^{n_1(x)} \lambda_{\beta_1}^{SCK}(x) \cdot C_{11}(x_{\alpha_1} - x_{\beta_1}) + \sum_{\beta_2=1}^{n_2(x)} \lambda_{\beta_2}^{SCK}(x) \cdot C_{12}(x_{\alpha_1} - x_{\beta_2}) = \\ = C_{11}(x_{\alpha_1} - x), \alpha_1 = 1, \ldots n_1(x) \\ \sum_{\beta_1=1}^{n_1(x)} \lambda_{\beta_1}^{SCK}(x) \cdot C_{21}(x_{\alpha_2} - x_{\beta_1}) + \sum_{\beta_2=1}^{n_2(x)} \lambda_{\beta_2}^{SCK}(x) \cdot C_{22}(x_{\alpha_2} - x_{\beta_2}) = \\ = C_{21}(x_{\alpha_{21}} - x), \alpha_2 = 1, \ldots n_2(x) \end{cases}$$

where C_{ij} are auto- and cross-covariances of primary ($Y^{(1)}$) and secondary ($Y^{(2)}$) variables.

Simple cokriging uses the data from the neighbourhood – both samples and previously simulated values. Simple cokriging is an exact predictor – it honours the data at the sampling points.

Estimation variance is given by:

$$\sigma_{SCK}^2(x) = C_{11}(0) - \sum_{\alpha_1=1}^{n_1(x)} \lambda_{\alpha_1}^{SCK}(x) \cdot C_{11}(x_{\alpha_1} - x) - \sum_{\alpha_2=1}^{n_2(x)} \lambda_{\alpha_2}^{SCK}(x) \cdot C_{21}(x_{\alpha_2} - x)$$

- Simple cokriging estimate and estimation variance are used to build a Gaussian distribution with the mean $Y^{(1)*}_{SCK}$ and standard deviation σ_{SCK}^2.

- The cosimulated value is dragged from the normal distribution built:

$$Y^{(l)}(x) = Y^*_{SCK}(x) + E(x)$$

where error component $E(x)$ is independent from $Y^*_{SCK}(x)$, $E\{E(x)\}=0$ and $Var\{E(x)\}= \sigma_{SCK}^2$.

- Final cosimulation value appears after back normal transformation.

6. The cosimulated value is added to the data and then used for cosimulation in next nodes of the pass.

7. Final NNRSGCS simulated value is a sum of ANN estimate and sequential Gaussian cosimulation value of the residual.

Case Study: Radioactive Soil Contamination by ^{90}Sr and ^{137}Cs

Radioactive soil contamination caused by the Chernobyl fallout feature anisotropic highly variable and spotty spatial pattern. The multi-scale character of the pattern is due to numerous influencing factors. Structural analysis of sample data discovers limitation for use of stationary estimation/simulation models, like kriging or stochastic simulation. Idea of applying ANN for modelling large-scale spatial tends before using geostatistical models, was first proposed in [1] and then was successfully developed in [3, 4]. Joint consideration of correlated variables allows to improve the quality of the primary variable estimation. Sequential Gaussian Cosimulation model applied to the correlated variables showed significant advantages over the results provided by single variable simulations [6]. Multivariate ANN+geostatistica model Neural Network Residual Cokriging (NNRCK) proposed in [2] showed excellent performance on ^{90}SR and ^{137}CS data from the Chernobyl fallout. ^{90}SR and ^{137}CS feature significant correlation (about 70-80%) due to the source term. Thus, application of NNRSGCS model is expected to provide good results as well. Variogram analysis of normal score data discovered long-range structures (80 km) and local correlation (10-15 km) (see Fig. 1). This conclusion leads to ANN use for trend modelling. Another problem with not de-trended data is that normal score variogram does not reach the *sill=1*, which is required for normally distributed variable.

ANN with 5 neurones in 1 hidden layer is chosen for final estimation, based on the accuracy test performance. ANN estimates captures major hot spot and fairly well represent long range structure of the pattern (see Fig. 1). The residuals appear correlation with the samples and the variograms of their normalised values represent local correlation structure at 15-20 km (see Fig. 3). The normal score residuals variograms for ^{90}SR and ^{137}CS are very easy to model and satisfy the condition for normal score variogram to possess level of at *sill=1*. The elaborated variogram models, which include linear model of co-regionalisation, are used in sequential Gaussian cosimulation procedure.

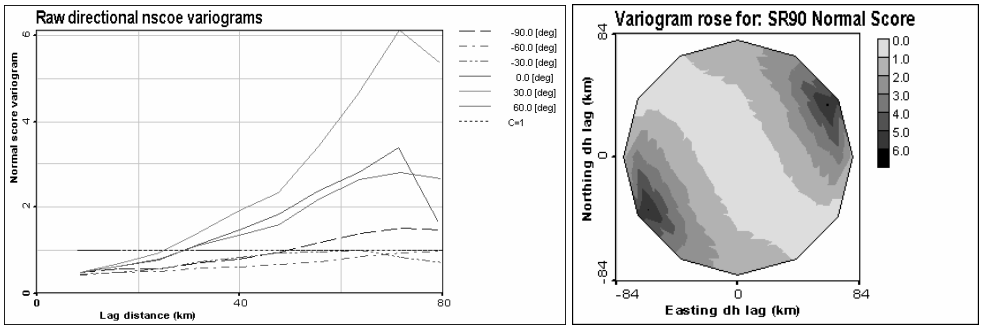

Figure 1. Raw directional (left) and raw variogram rose (right) variograms for normal score[90] SR samples.

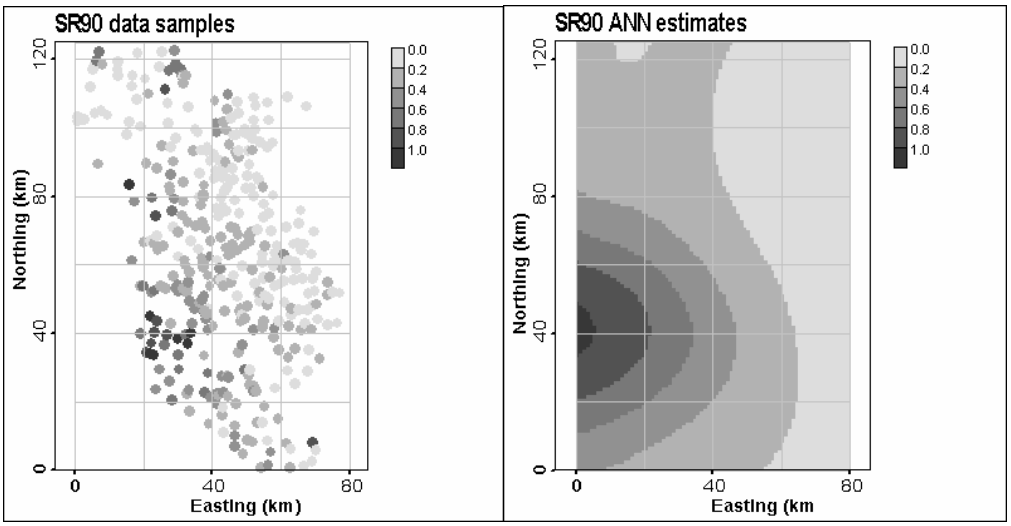

Figure 2. [90]SR: samples values (left) and ANN [2-5-0-2] estimates (right).

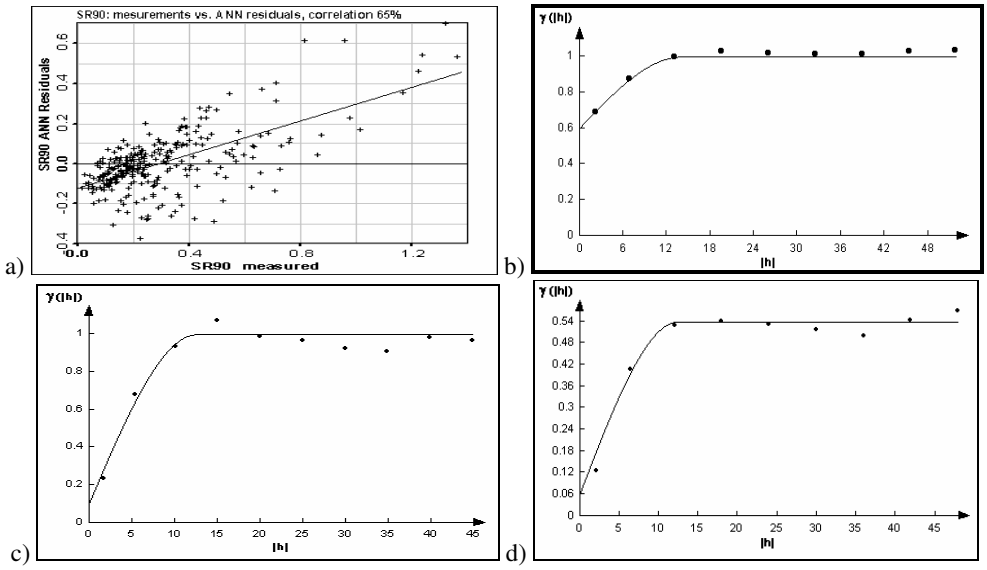

Figure 3. Correlation between [90]SR measurements and ANN residuals (a); variograms for ANN residuals - [90]SR (b),

^{137}CS (c), cross-variogram (d).

Final NNRSG cosimulation results are presented as equeprobable realisations (see Fig. 4) The realisation show preservation of the histogram and spatial correlation structure represented by the variograms of simulations vs. raw variogram (see Fig. 5).

Average E-type estimate is obtained from 40 realisations (see Fig. 6). It can be compared with other estimates (cokriging, NNRCK). Further processing of the NNRSGCS results provided risk maps – estimates which can be exceeded with a certain probability. Maps of such estimates are presented in Fig. 7 for 50% and 90% probability of exceeding.

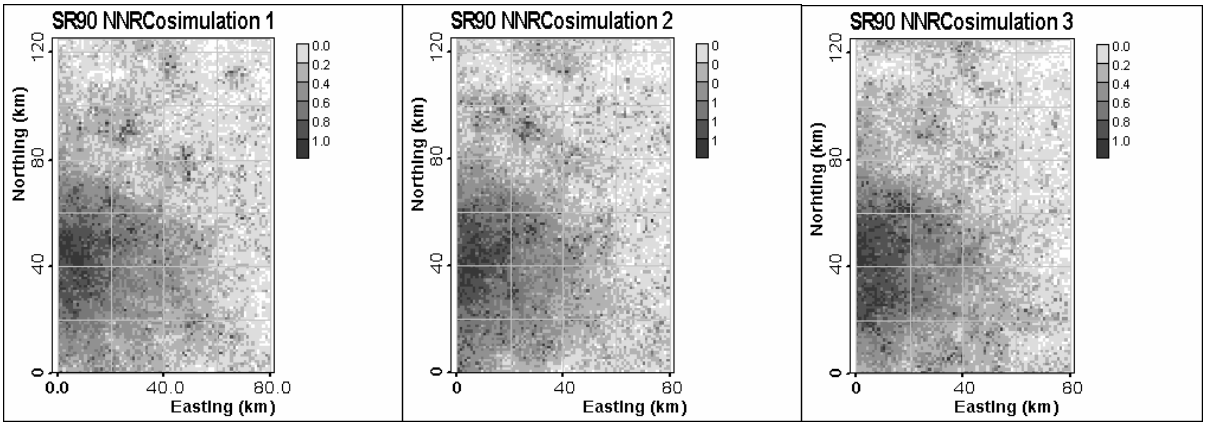

Figure 4. ^{90}SR Neural Network Residual Sequential Gaussian Cosimulation realisations.

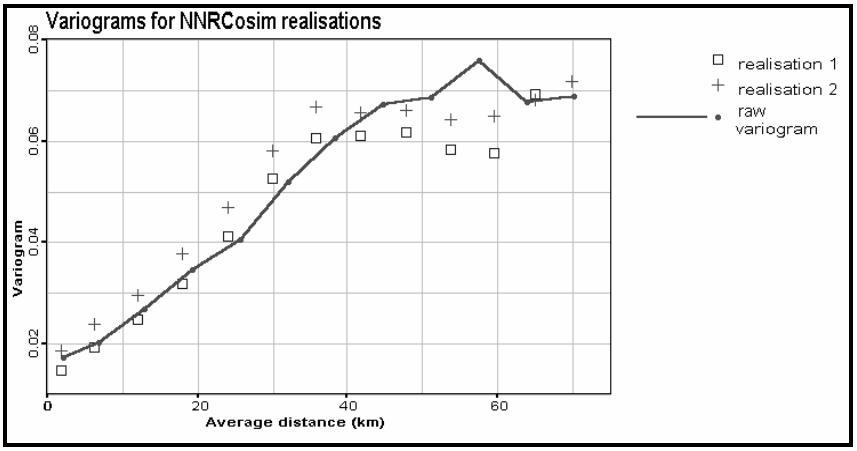

Figure 5. Variograms of simulated realisations vs. raw data variogram.

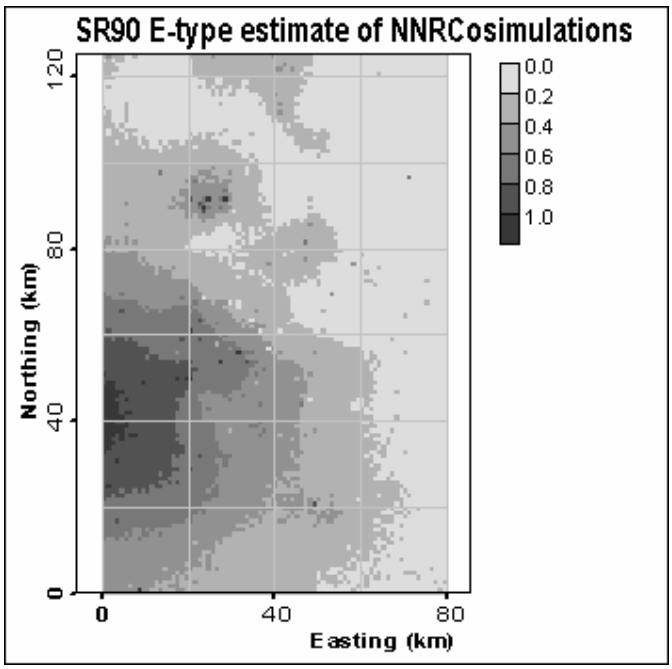

Figure 6. ^{90}SR Neural Network Residual Sequential Gaussian Cosimulation E-type estimates.

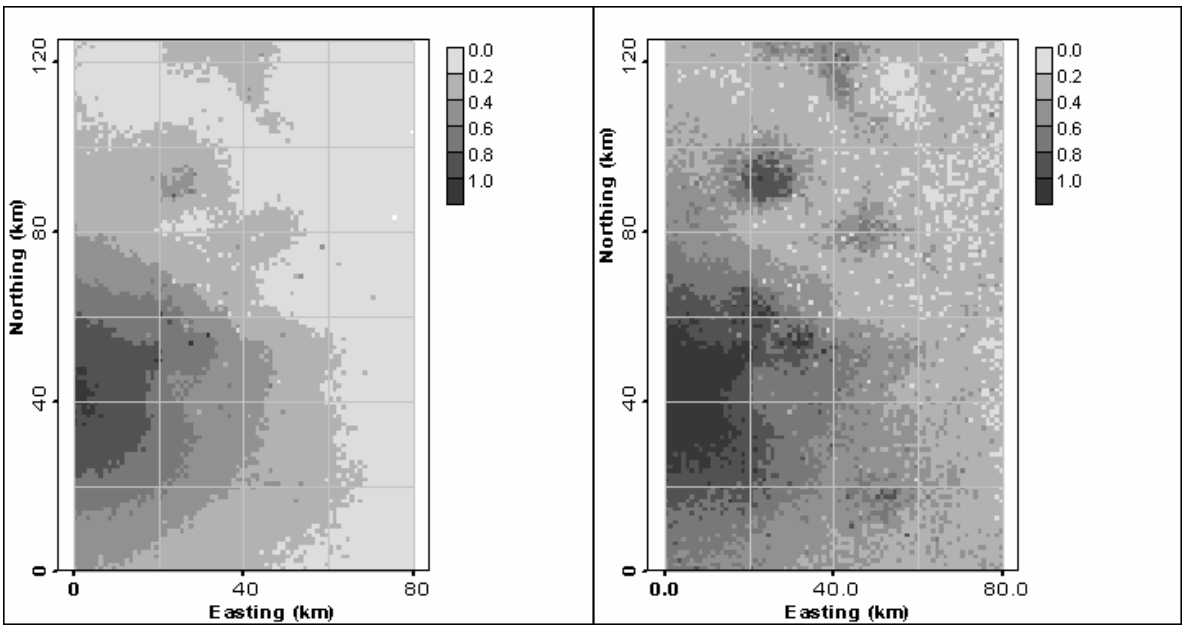

Figure 7. ^{90}SR Neural Network Residual Sequential Gaussian Cosimulation estimates with a probability of being

exceeded: 0.5 (left) and 0.1 (right).

Conclusions

A new stochastic simulation method based on ANN is proposed in the work. NNRSGCS includes ANN model for large-scale trend and sequential stochastic cosimulation of residuals, which model variability and uncertainty of the pattern, preserving its univariate distribution function and spatial correlation. Using additional samples of the secondary variable (^{137}CS) allows to improve quality of the simulated pattern and to discover additional hot spots and extend monitoring network of the primary variable ^{90}SR. NNRSGCS provided fairly well realisation of ^{90}SR spatial pattern, reproducing its spatial variability and uncertainty. Risk maps of estimates that can be exceeded with a certain probability are obtained from NNRSG cosimulation results to support decision-making process.

Acknowledgements

The work was supported in part by INTAS Grants 96 1957 and 97 31726.

The results analysis and presentation as well as ANN modelling was performed with the help of GEOSTAT OFFICE software (http://www.ibrae.ac.ru/~mkanev) [7].

References

[1] Kanevsky, R. Arutyunyan, L. Bolshov, V. Demyanov, M. Maignan. Artificial neural networks and spatial estimations of Chernobyl fallout. Geoinformatics, 1996, vol. 7, nos. 1-2., p. 5-11.

[2] Kanevski M., Demyanov V. and Maignan M. Mapping of Soil Contamination by Using Artificial Neural Networks and Multivariate Geostatistics. In: Artificial Neural Networks ICANN '97. 7th International Conference, Switzerland. Proceedings W. Gerstner, A. Germond, M. Hasler, J.-D. Nicould (eds.). Lecture Notes in Computer Science, Springer,1997, p. 1125.

[3] Kanevski, V. Demyanov and M. Maignan. Spatial estimations and simulations of environmental data by using geostatistics and artificial neural networks. In: IAMG'97 Proceedings of The Third Annual Conference of the International Association for Mathematical Geology. Ed. V. Pawlowsky Glan. Barcelona, Spain, CIMNE, 1997, ISBN 84-87867-97-9, vol. 2, p.527.

[4] Kanevsky, R. Arutyunyan, L. Bolshov, S. Chernov, V. Demyanov, I. Linge, N. Koptelova, E. Savelieva, T. Haas, M. Maignan. Chernobyl Fallouts: Review of Advanced Spatial Data Analysis. In: geoENV I — Geostatistics for Environmental Applications, ed. A. Soares, J. Gomez-Hernandes, R. Froidvaux, Kluwer Academic Publishers, 1997, pp. 389-400.

[5] Goovaerts P. Geostatistics for Natural Resources Evaluation. Oxford University Press, 1997.

[6] M. Kanevski, V. Demyanov, E. Savelieva, S. Chernov, M Maignan. *Conditional Stochastic Cosimulations Of The Chernobyl Fallout.* geoENV II - Geostatistics for Environmental Applications, ed. J. Gomez-Hernandez, A. Soares, R. Froidevaux. Kluwer Academic Publishers, 1998, pp. 453-464.

[7] M. Kanevski, V. Demyanov, S. Chernov, E. Savelieva, A. Serov, V. Timonin. *Geostat Office for Environmental and Pollution Spatial Data Analysis*. Mathematische Geologie, band 3, April, 1999, pp. 73-83, CPress Publishing House, 1999.

[8] Deutsch C., Journel A.G. GSLIB. Geostatistical Software Library and User's Guide. Oxford University Press. 1992, 1996, 340p.