Available online at www.sciencedirect.com
SCIENCE DIRECT®
Environmental Modelling & Software 19 (2004) 845–855
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Environmental data mining and modeling based on machine learning algorithms and geostatistics

M. Kanevski [a,b], R. Parkin [c,*], A. Pozdnukhov [a,c,d], V. Timonin [c], M. Maignan [b], V. Demyanov [c], S. Canu [e]

[a] *IDIAP Dalle Molle Institute for Perceptual Artificial Intelligence, Simplon 4, 1920 Martigny, Switzerland*
[b] *Lausanne University, Lausanne, Switzerland*
[c] *IBRAE, Nuclear Safety Institute, Russian Academy of Sciences, Environmental Modelling and System Analysis Laboratory, 52 B, Tulskaya, Moscow 113191, Russia*
[d] *Physics Department, Moscow State University, Math. Division, Moscow, Russia*
[e] *INSA, Rouen, France*

Received 30 September 2002; received in revised form 13 January 2003; accepted 5 March 2003

Abstract

The paper presents some contemporary approaches to spatial environmental data analysis. The main topics are concentrated on the decision-oriented problems of environmental spatial data mining and modeling: valorization and representativity of data with the help of exploratory data analysis, spatial predictions, probabilistic and risk mapping, development and application of conditional stochastic simulation models. The innovative part of the paper presents integrated/hybrid model—machine learning (ML) residuals sequential simulations—MLRSS. The models are based on multilayer perceptron and support vector regression ML algorithms used for modeling long-range spatial trends and sequential simulations of the residuals. ML algorithms deliver non-linear solution for the spatial non-stationary problems, which are difficult for geostatistical approach. Geostatistical tools (variography) are used to characterize performance of ML algorithms, by analyzing quality and quantity of the spatially structured information extracted from data with ML algorithms. Sequential simulations provide efficient assessment of uncertainty and spatial variability. Case study from the Chernobyl fallouts illustrates the performance of the proposed model. It is shown that probability mapping, provided by the combination of ML data driven and geostatistical model based approaches, can be efficiently used in decision-making process.

Keywords: Environmental data mining and assimilation; Geostatistics; Machine learning; Stochastic simulation; Radioactive pollution

1. Introduction

Environmental data feature complex spatial pattern at different scales due to combination of several spatial phenomena or various influencing factors of different origins. In some cases, the original observations are taken with significant measurement errors and may contain significant uncertainty as well as a number of outliers. Non-linear spatial trends corresponding to large-scale processes complicate geostatistical modeling as long as stationary models (e.g. ordinary kriging) are concerned. Trend removal is also necessary for comprehensive spatial correlation analysis and modeling (variography). Variogram modeling for such data and using common geostatistical approaches will result in incorrect results. In the presence of trends, the data can be decomposed into two parts:

$$Z(x) = M(x) + e(x) \quad (1)$$

where $M(x)$ represents large-scale deterministic spatial variations (trends), and $e(x)$ represents small-scale spatial stochastic variations. Contemporary geostatistics offers several possible approaches to handle spatial trends (spatial non-stationarity): universal kriging (implying a polynomial trend model), residual kriging, moving window regression residual kriging (see Cressie, 1991; Deutsch and Journel, 1998; Dowd, 1994; Neuman

* Corresponding author. Tel.: +7-095-955-2231; fax: +7-095-958-1151.
E-mail addresses: kanevski@idiap.ch (M. Kanevski); park@ibrae.ac.ru (R. Parkin); pozd@idiap.ch (A. Pozdnukhov).

1364-8152/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2003.03.004

and Jacobson, 1984; Gambolati and Galeati, 1987; Hass, 1996). All these approaches imply a certain formula based trend model, which is not necessarily in a good agreement with the data. An alternative way for trend modeling is to use a data-driven approach, which relies only on data. One of such approaches was developed by partitioning heterogeneous study area into some smaller homogeneous subareas and analyzing the spatial structure within them separately (Pélissier and Goreaud, 2001).

In the present paper, we propose a newly developed model—machine learning residuals sequential Gaussian simulations (MLRSGS) as an extension of the ideas presented by Kanevsky et al. (1996b) and Demyanov et al. (2000). In these papers, a hybrid model—neural network residuals kriging (NNRK)—was first introduced and then extended for use in a combination with different geostatistical models. The basic idea is to use feedforward neural network (FFNN), which is a well-known global universal approximator to model large-scale non-linear trends, and then to apply geostatistical estimators/simulators for the residuals. Machine learning algorithms unite a wide family of data-driven models. Here, we will focus on two of them: multilayer perceptron (MLP) and support vector regression (SVR). Another type of hybrid models (expert systems) was developed by using geographical information systems and modeling integrated into a decision support system for environmental and technological risk assessment and management (see Fedra and Winkelbauer, 1999).

One of the principal advantages of machine learning algorithms is their ability to discover patterns in data, which exhibit significant unpredictable non-linearity. Being a data-driven approach ("black-box" models), ML depends only on the quality of data and the architecture of model, particularly for MLP—number of hidden neurons, activation functions types of connections. ML can capture spatial peculiarities of the pattern at different scales describing both linear and non-linear effects. Performance of MLA is based on solid theoretical foundations, which were considered by Bishop (1995), Haykin (1999) and Vapnik (1998).

Stochastic simulation is an intensively developed and used approach to provide uncertainty and risk assessment for spatial and spatio-temporal problems. Stochastic simulation models are preferable over estimators as they are able to provide a joint probabilistic distributions rather than a single value estimates. Sequential Gaussian simulation (SGS) is one of the widely used methods that is able to handle highly variable data but still is sensitive to trend, thus formally requires to some extent spatial stationarity assumed. SGSs are based on *modeling* of spatial correlation structures—variography.

A mixture of ML *data driven* and geostatistical *model based* approaches is also attractive for decision-making process because of their interpretability.

The real case study on soil pollution from the Chernobyl fallout illustrates the application of the proposed model. The accident at Chernobyl nuclear power plant caused large-scale contamination of environment by radiologically important radionuclides. Large-scale consequences of the Chernobyl fallout were considered in the past decade and one of the comprehensive mapping work was presented in De Cort and Tsaturov (1996). Geostatistical analysis and prediction modeling of radioactive soil contamination data was presented in Kanevsky et al. (1996a).

2. Machine learning residual Gaussian simulations

2.1. Methodology of ML residual Gaussian simulations

The basic idea is to use ML to develop a non-parametric, robust model to extract large-scale non-linear structures from data (detrending) and then to use geostatistical models to simulate the residuals at local scales. In brief, the MLRSGS algorithm follows the steps given below (extended after Kanevsky et al., 1996b):

1. Data preprocessing and exploratory analysis: in general, split data into training, testing and validation sets, checking for outliers, exploratory data analysis, estimations and modeling of spatial correlation—experimental and theoretical variography. Training set is used for the ML algorithms training, validation set is used to tune hyperparameters (e.g. number of hidden neurons) while testing set is applied to assess MLA generalization ability.
2. Training and testing of ML algorithm. In the present paper, MLP and SVR are used. They are well-known function approximators and are described briefly below.
3. Accuracy test—comprehensive analysis of the residuals provides the ML residuals at the training points (*measured–estimated*) which are the base for further analysis. Two further cases are possible:

- the residuals are not correlated with the measurements correlated (both 1D and 2D), which means that MLA has modeled all spatial structures presented in the raw data;
- the residuals show some correlation with the samples, then further analysis should be performed on the residuals to model this correlation.

4. ML residuals are explored using variography. The remaining spatial correlation presents short-range correlation structures, once long-range correlation (trend) in the whole area was modeled by MLA.
5. Normal score transformation (non-linear transformation from raw data to Nscore values, distributed

$N(0,1)$ is performed to prepare data for further Gaussian simulations. Nscore variogram model describing spatial correlations of Nscore values is built. SGS is then applied to the MLA residuals and stochastic realizations are generated using the training dataset.

The idea of stochastic simulation is to develop a spatial Monte Carlo model that will be able to generate many, in some sense equally probable, realizations of a random function (in general, described by a joint probability density function). Any realization of the random function is called an unconditional simulation. Realizations that honor the data are called conditional simulations. Basically, the simulations try to reproduce the first (univariate global distribution) and the second (variogram) moments. The similarities and dissimilarities between the realizations describe spatial variability and uncertainty. The simulations bring valuable information on the decision-oriented mapping of pollution. Postprocessing of the simulated realizations provides probabilistic maps: maps of probabilities of the function value to be above/below some predefined decision levels. Gaussian random function models are widely used in statistics and simulations due to their analytical simplicity, they are well understood and are limit distributions of many theoretical results. SGS algorithm used in this work was described in detail in Deutsch and Journel (1998).

6. Simulated values of the residuals appear after back normal score transformation. Final ML residual simulations value is a sum of ML estimate and SGS realization.

2.2. Description of multilayer perceptron model

MLP is a type of artificial neural network with a specific structure and training procedure described in Bishop (1995) and Haykin (1999).

The key component of MLP is the *formal neuron*, which sums the inputs, and performs a non-linear transform via the activation function f (Fig. 1). The activation function (or non-linear transformer) can be any continuous, bounded and non-decreasing function. Exponential sigmoid or hypertangent are commonly used in practice. The weights $W(w_0, \ldots, w_n)$ are adaptive parameters which are optimized by minimizing the following quadratic cost function:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(t_i - o_i)^2 \tag{2}$$

where MSE is the mean square error, N is the number of samples, o_i is the net output (prediction) and t_i is the real function desired value. Backpropagation error algorithm is applied to calculate gradient of MSE on adaptive weight, $\partial E/\partial W$. Various optimization algorithms, which employ backpropagation, can be used, such as the conjugate gradient descend method, second-order pseudo-Newton Levenberg-Marquardt method, or the resilient propagation method.

In a standard MLP, the neurons are arranged in input, hidden and output layers. The values of the exploratory variables (X and Y co-ordinates) are exposed to the input layer, the output layer produces and compares the target estimate of the function value (^{137}Cs concentration), hidden layers (one or two) allow(s) to handle non-linearity (Fig. 2). The number of neurons in the hidden layers can vary and is the subject to the optimum configuration. As long as the aim of MLP in the present work is to extract a large-scale trend, as few hidden neurons as possible were chosen to extract non-linear trends. Further increase of the number of hidden neurons leads to extracting more detailed local peculiarities and even noise from the pattern: choosing too many hidden neurons will lead to over-fitting (or over-learning) when MLP loses its ability to generalize the information from the samples. On the other hand, using too few hidden neurons does not provide explicit extraction of the trend; hence some large-scale correlations will remain in the residuals restricting further procedure. Thus, geostatistical variogram analysis becomes the key tool to control the MLP performance for trend extraction.

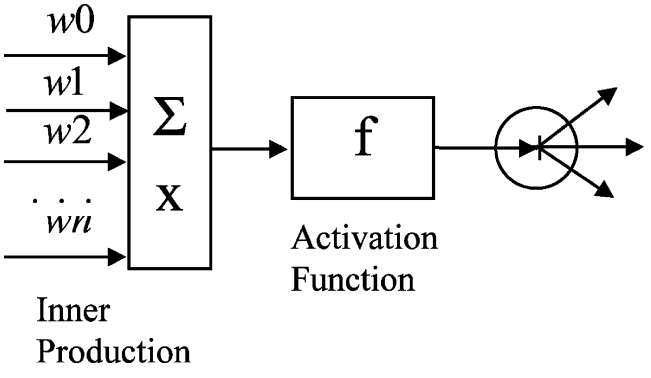

Fig. 1. Formal neuron.

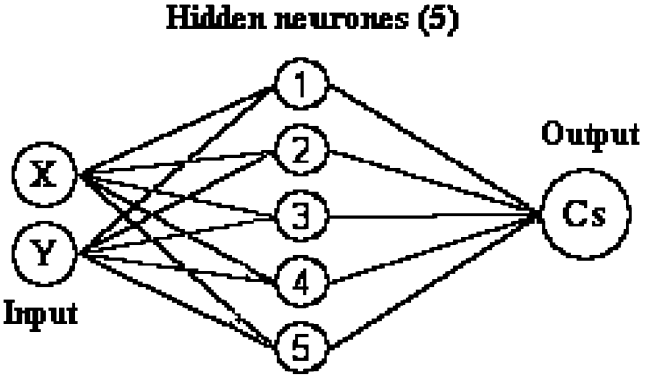

Fig. 2. Multilayer perceptron.

2.3. Description of support vector regression model

SVR is a recent development of the statistical learning theory (SLT) (Vapnik, 1998). It is based on structural risk minimization and seems to be promising approach for spatial data analysis and processing (see Scholkopf and Smola, 1998; Gilardi and Bengio, 2000; Kanevski et al., 2001). There are several attractive properties of the SVR: robustness of the solution, sparseness of the regression, automatic control of the solutions complexity, good generalization performance (Vapnik, 1998). In general, by tuning SVR hyper-parameters, it is possible to cover a wide range of spatial regression functions from over-fitting to over-smoothing (Kanevski et al., 2001).

First, we state a general problem of regression estimation as it is presented in the scope of SLT. Let $\{(\boldsymbol{x}_1,y_1),\ldots(\boldsymbol{x}_N,y_N)\}$ be a set of observations generated from an unknown probability distribution $P(x, y)$ with $x_i \in \mathbf{R}^n$, $y_i \in \mathbf{R}$, and $F = \{f|\mathbf{R}^n \rightarrow \mathbf{R}\}$ a class of functions. The task is to find a function f from the given class of functions that minimizes a risk functional:

$$R[f] = \int Q(y-f(x),x)\mathrm{d}P(x,y) \tag{3}$$

where Q is a loss function indicating how the difference between the measurement value and the model's prediction is penalized.

As $P(x, y)$ is unknown, one can compute an empirical risk:

$$R_{\mathrm{emp}} = \frac{1}{N}\sum_{i=1}^{N} Q(y_i-f(x_i),x_i) \tag{4}$$

When it is only known that noise-generating distribution is symmetric, the use of linear loss function is preferable, and results in a model from robust regression family. For simplicity, we also assume loss to be the same for all spatial locations.

The Support vector regression model is based on a new type of loss functions, the so-called ε-insensitive loss functions. Symmetric linear ε-insensitive loss is defined as:

$$Q(y-f(x),x) = \begin{cases} |y-f(x)|-\varepsilon, & \text{if } |y-f(x)| > \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The asymmetrical loss function can be used in applications where underestimations and overestimations are not equivalent.

Let us start from the estimation of regression function in a class of linear functions $F = \{f(x)|f(x) = (w,x) + b\}$. Support vector regression is based on the structural risk minimization principle, which results in penalization of the model complexity simultaneously with keeping small empirical risk (training error). The complexity of linear functions can be controlled by the term $\|w\|^2$, see Eq. (6) (Vapnik, 1998). Also, we have to minimize the empirical risk (training error). With selected symmetrical linear ε-insensitive loss, empirical risk minimization is equivalent to adding the slack variables ξ_i,ξ_i^* into the functional with the linear constraints (7). Introducing the trade-off constant C, we arrive at the following optimization problem:

$$\text{minimize } \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{6}$$

$$\text{subject to } \begin{cases} f(x_i)-y_i-\varepsilon \leq \xi_i \\ -f(x_i) + y_i-\varepsilon \leq \xi_i^* \\ \xi_i,\xi_i^* \geq 0, \qquad \text{for } i = 1,\ldots,N \end{cases} \tag{7}$$

The slack variables ξ_i,ξ_i^* measure the distance between the observation and the ε tube. The distance between the observation and the ε and ξ_i,ξ_i^* is illustrated by the following example: imagine you have a great confidence in your measurement process, but the variance of the measured phenomena is large. In this case, ε has to be chosen a priori very small while the slack variables ξ_i,ξ_i^* are optimized and thus can be large. Remember that inside the ε tube ($[f(x)-\varepsilon,f(x) + \varepsilon]$) loss function is zero.

Note that by introducing the couple (ξ_i,ξ_i^*), the problem now has $2n$ unknown variables. But these variables are linked since one of the two values is a necessary equal to zero. Either the slack is positive ($\xi_i^* = 0$) or negative ($\xi_i = 0$). Thus, $y_i \in [f(x_i)-\varepsilon-\xi_i,f(x_i) + \varepsilon + \xi_i^*]$.

A classical way to reformulate the constraint based minimization problem is to look for the saddle point of Lagrangian L:

$$\begin{aligned} L(w,\xi,\xi^*\alpha) = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}\alpha_i(y_i \\ & -f(x_i) + \varepsilon + \xi_i) - \sum_{i=1}^{N}\alpha_i^*(f(x_i)-y_i + \varepsilon + \xi_i^*) \\ & - \sum_{i=1}^{N}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{8}$$

where $\alpha_i,\alpha_i^*,\eta_i,\eta_i^*$ are Lagrange multipliers associated with the constraints; α_i,α_i^* can be roughly interpreted as a measure of the influence of the constraints on the solution. A solution with $\alpha_i = \alpha_i^* = 0$ can be interpreted as "the corresponding data point has no influence on this solution". Other points with non-zero α_i or α_i^* are the "support vectors (SVs)" of the problem.

The dual formulation of the optimization problem is solved in practice:

$$\text{maximise } -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(x_i\cdot x_j) \quad (9)$$

$$-\varepsilon\sum_{i=1}^{N}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{N}y_i(\alpha_i^*-\alpha_i)$$

$$\text{subject to } \begin{cases}\sum_{i=1}^{N}(\alpha_i^*-\alpha_i)=0\\ 0\le\alpha_i^*,\alpha_i\le C, \text{ for } i=1,\ldots,N\end{cases}$$

This is a quadratic programming (QP) problem, hence has an unique solution. It can be solved numerically by a number of methods. After we get the values α_i and α_i^*, we can compute b from the constraints of the primary problem (7) and make predictions:

$$f(x)=\sum_{i=1}^{N}(\alpha_i^*-\alpha_i)(x_i\cdot x)+b \quad (10)$$

Note that both the solution (10) and the optimization problem (9) are written in the terms of dot products. Hence, we can use a so-called "kernel trick" to achieve non-linear regression model. We substitute the dot products (x_i,x_j) with a suitable function $\{K\in L^2(R^n)\oplus L^2(R^n),K|(R^n\oplus R^n)\rightarrow R\}$. If the kernel function satisfies the Mercer's conditions:

$$\iint K(x',x'')g(x')g(x'')\mathrm{d}x'\mathrm{d}x''>0 \quad (11)$$

for any $g(x)\in L^2(R^n)$, then it can be expanded in a uniformly converging series

$$K(x',x'')=\sum_j\lambda_j\Phi_j(x')\Phi_j(x'') \quad (12)$$

where $\{\lambda_i,\Phi_j(.)\}$ is an eigensystem of K. We may regard $\Phi_j(x)$ as some j-th feature of vector x, then kernel K is a dot product in some feature space. As (11) determines positively defined kernels, the substitution of K instead of dot products in (9) results in a still convex QP problem:

$$\text{maximise } -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j) \quad (13)$$

$$-\varepsilon\sum_{i=1}^{N}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{N}y_i(\alpha_i^*-\alpha_i)$$

$$\text{subject to } \begin{cases}\sum_{i=1}^{N}(\alpha_i^*-\alpha_i)=0\\ 0\le\alpha_i^*,\alpha_i\le C, \text{ for } i=1,\ldots N\end{cases}$$

and the prediction is a non-linear regression function:

$$f(x)=\sum_{i=1}^{N}(\alpha_i^*-\alpha_i)K(x_i,x)+b \quad (14)$$

3. Case study

Radioactive soil contamination caused by the Chernobyl fallout features anisotropic highly variable and spotty spatial pattern. Multiscale character of the pattern is due to numerous influencing factors: the source term, weather conditions (especially rainfall), dry and damp precipitations, surface properties (orography, ground cover, soil type, land use, etc.). The most significant influence on the long-term contamination was provided by the radionuclide cesium ^{137}Cs. The half-life period of this isotope is about 30 years.

The selected region is rectangular covering 7428 km^2 with 845 populated sites. The basic statistical parameters of the data (^{137}Cs concentration at 684 points) presented in Fig. 3 are the following: minimum value 5.9 kBq/m^2,

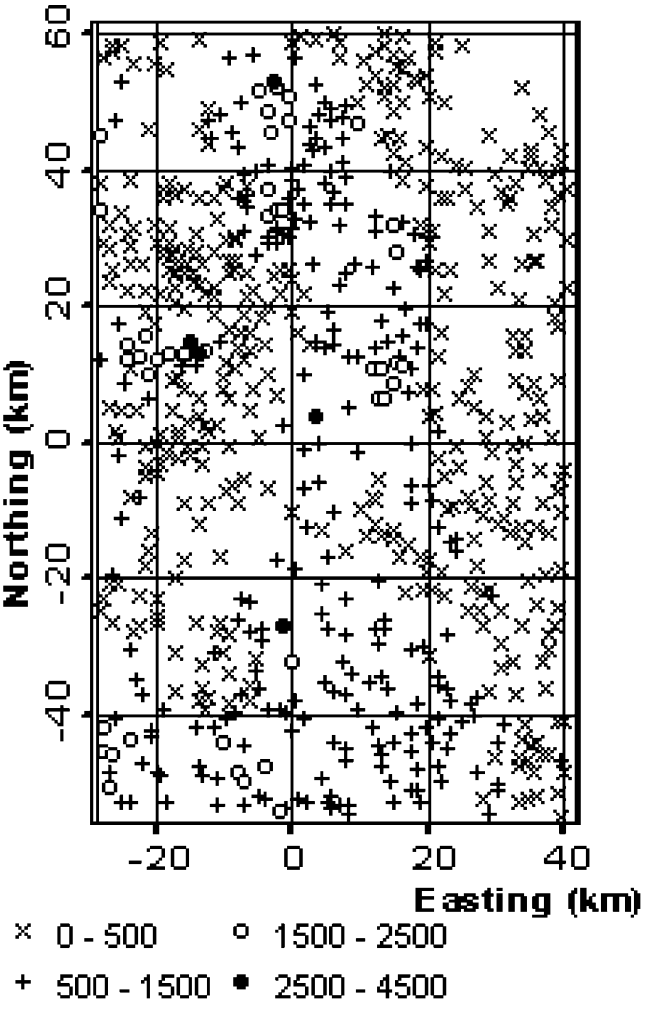

Fig. 3. Raw data on ^{137}Cs concentration in the Bryansk region.

mean value 571.8 kBq/m^2, maximum value 4334 kBq/m^2, variance 315,372 kBq^2/m^4, skewness 2.7 and kurtosis 16.9. As usually, environmental data are positively skewed and their distributions are far from normal.

The samples reflecting spatial contamination pattern are the subject of exploratory spatial data analysis to address spatial continuity. Spatial continuity is a feature of spatial processes, which have some underlying origin in the physics of the process. The presence of spatial continuity means that closer samples are more likely similar than farther ones (Issaks and Shrivastava, 1989).

Because the samples represent only one realization of the spatial process, some kind of stationarity assumption is required to use statistical methods. Strong stationarity means that for any finite number n of sample points x_i ($i = 1,\ldots,n$) and any lag h, the joint finite-dimensional distribution functions of $Z(x_1),Z(x_2),\ldots,Z(x_n)$ are the same as of $Z(x_1 + h),Z(x_2 + h),\ldots,Z(x_n + h)$. In practice, this proposition is very hard to be detected, and as we are usually interested only in two first moments, the second-order stationarity assumption is enough. It is the stationarity only of the two first moments: the mean is constant ($E[Z(\boldsymbol{x})] = m = \text{const}$) and covariance ($Cov(x_1,x_2) = E[Z(x_1),Z(x_2)] - m^2 = C(h)$) exists and does not depend on x, but only on h.

Rather often, the real data do not follow even second-order stationarity model. Intrinsic hypothesis, which is weaker than second-order stationarity, is enough to apply geostatistical tools. The intrinsic hypothesis is a process with second-order stationarity applied for the increments. It means that the mean of increments (also named the drift):

$$D(h) = E[Z(x) - Z(x + h)] \tag{15}$$

is constant and $D(h) = 0$ and does not depend on x and h, and the variance of increments ($2\gamma(h) = \text{var}[Z(x + h) - Z(x)]$) exists and does not depend on x, only on h.

The drift $D(h)$ can be an indicator of data obedience to the intrinsic hypothesis. Such a deduction can be made, for example, when the value of the drift $D(h)$ fluctuates around zero (the drift is supposed to be zero whatsoever the position of h in the domain). If $D(h)$ increases/decreases with the augmentation of the length of the separation vector h (see Fig. 4), then the data do not follow the intrinsic hypotheses. It can mean that the data have systematic trend. In such cases, the variogram modeling and the following common geostatistical prediction will result in misleading results. To handle this problem, the trend must be removed from the data in the first place. Here, the machine learning algorithms have been used to model the trend in the data.

Cell declustering was used for splitting data into training and testing sets to provide efficient ML learning. The region was divided into rectangular cells by a regular grid and one or several points were selected at random from each cell. The testing dataset was obtained in this way, and the rest of the data formed the training set. Thus, testing set represents regional data. Of course, the training set in this case is somewhat clustered that is not so good for the MLA training. However, using the backward selection (i.e. picking out the points for training dataset by the declustering and to consider the rest as testing one) is impossible to obtain a representative testing set. The procedure of selection was carried out several times with different cell sizes and with varying numbers of the selected points from each cell. Since it is difficult to control both testing and training datasets, more attention was paid to the similarity of the training data set to the initial data structures of all data. The similarity was controlled by comparing summary statistics, histograms and spatial correlation structures. Similarity of spatial structures for the obtained datasets to the initial data is even more important than statistical factors. Comparison of the spatial structure was carried out with the help of variogram roses, which show anisotropy. Such comparison provides grounds that split (see Fig. 5) with 484 training and 200 testing points is quite suitable for the following ML modeling and it is the best of all obtained.

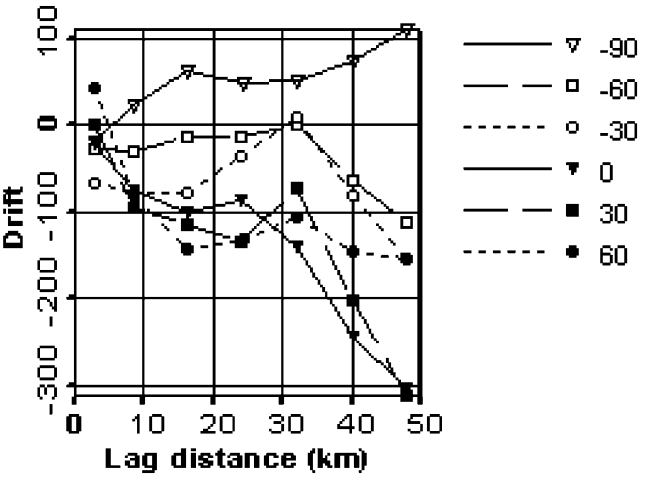

Fig. 4. The drift of ^{137}Cs data.

In the present study, MLP models with the following parameters were used: two input neurons, describing spatial co-ordinates (X, Y), one hidden layer and output neuron describing ^{137}Cs contamination. Backpropagation training with Levenberg-Marquardt followed by conjugate gradient algorithm was used in order to avoid local minima (Masters, 1995).

The variogram analysis of the obtained residuals for the trained neural networks with varying number of neurons in the hidden layer showed that the optimal results (in the sense of modeling non-linear trends) was obtained by using MLP with five neurons in a single hidden layer. Further increase of the number of hidden neurons leads to extracting more detailed local peculiarities of the pattern, reflected by multiple correlation range

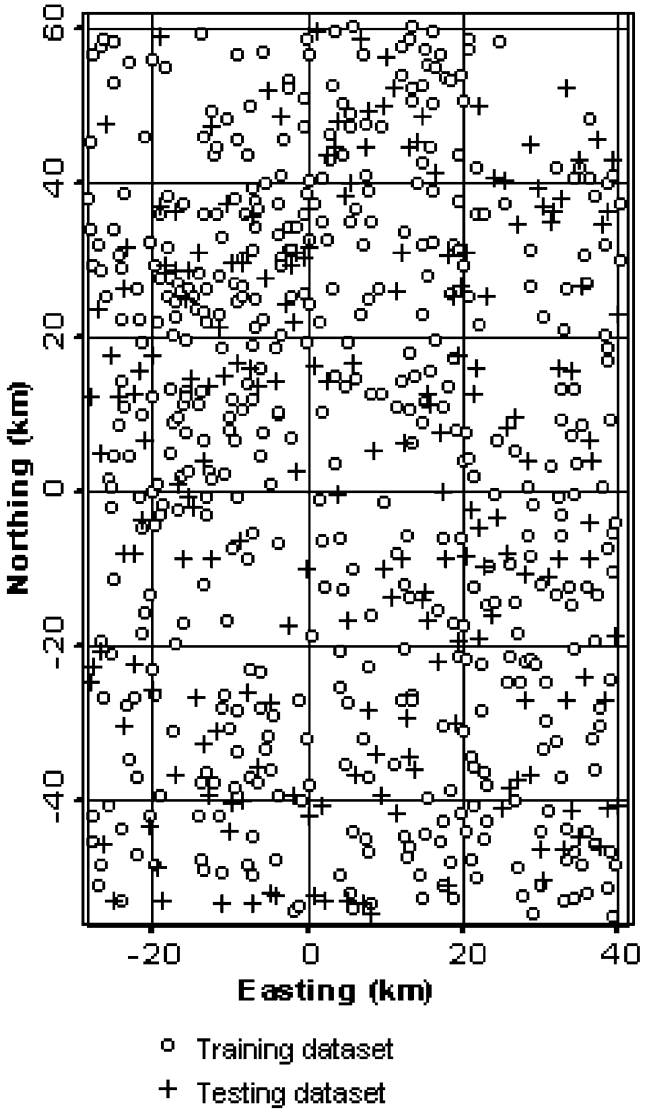

Fig. 5. Location of the training and testing points.

of the variogram of trend estimates. Then, MLP is used for ^{137}Cs spatial prediction mapping. Predictions were performed on a rectangular regular grid with cell size 1×1 km. The result for the ^{137}Cs MLP large-scale mapping is presented in Fig. 6.

Let us present the results of the large-scale modeling using support vector regression approach. Several user-defined (hyper) parameters influence on the SVR model: kernel function, C, ε. Gaussian radial basis functions (RBF) were found to be well suited for spatial environmental modeling:

$$K(x,x') = \exp\left(-\frac{|x-x'|^2}{2\sigma^2}\right) \quad (16)$$

Kernel parameter—bandwidth σ is related to some characteristic correlation scales of trend model. Kernel bandwidth of 20 km is used for the presented model. Other parameters were defined as: $C = 20$, $\varepsilon = 200$. This

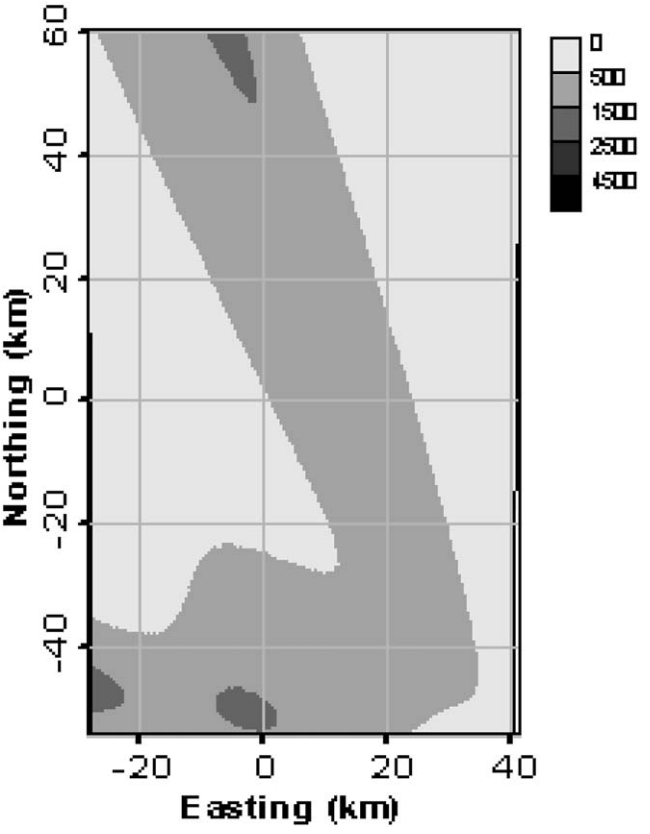

Fig. 6. ^{137}Cs, artificial neural network (one hidden layer with five neurons) spatial predictions.

choice is based both on the analysis of training and testing errors and the analysis of the variogram of the resulting trend model. Detailed description of the influence of the parameters on the solution and tuning procedure can be found in Kanevski et al. (2001). The mapping results (trend model) are presented in Fig. 7.

Trained multilayer perceptron and support vector regression were able to extract some information from data described by large-scale spatial correlations. The rest of the information—small scale spatially structured residuals—was analyzed and modeled using geostatistical conditional stochastic simulations. The obtained residuals are correlated with the original data and are not correlated with the MLA estimates (see Figs. 8 and 9). Correlation coefficients between the residuals and ^{137}Cs sample values are equal to 0.77 (for MLP residuals) and 0.79 (for SVR residuals).

Exploratory variography of spatial correlation structures of the Nscore transformed residuals are presented in Figs. 10 and 11. Variograms of the Nscore transformed residuals can be easily modeled (fitting to theoretical model) and SGSs can be applied (variogram reaches a sill and levels off). Final ML residual sequential Gaussian simulation results are presented as equiprobable realizations in Figs. 12 and 13. They keep the large-scale trend structure (from Figs. 6 and 7) and also feature dis-

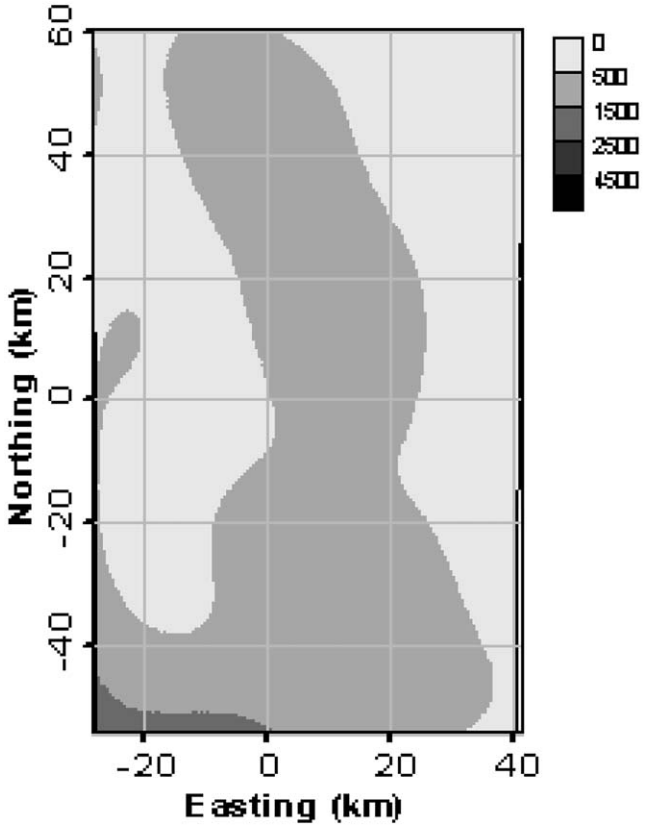

Fig. 7. ^{137}Cs, support vector regression trend modeling.

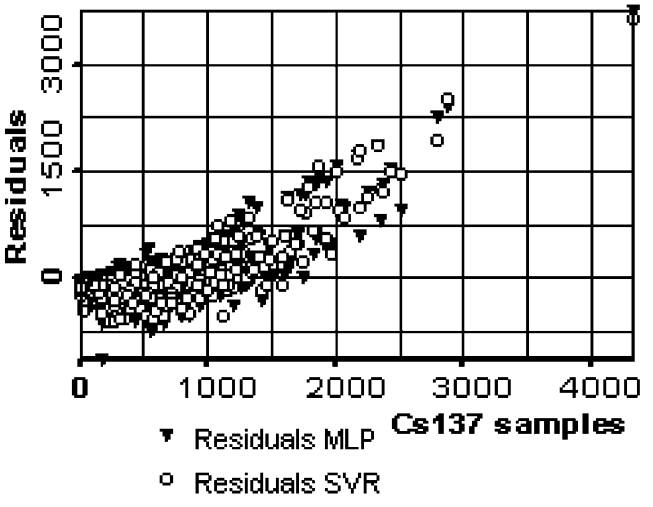

Fig. 8. Scatterplot of the MLP and SVR residuals vs. ^{137}Cs sample values.

tinctive spatial variability and small-scale effects ignored by ML models.

The similarity and dissimilarity between the realizations describe spatial variability and uncertainty. The next step deals with the probabilistic mapping: probability mapping of to be above/below some predefined decision level. This topic relates to decision-oriented mapping of contaminated territories. Usually, hundreds

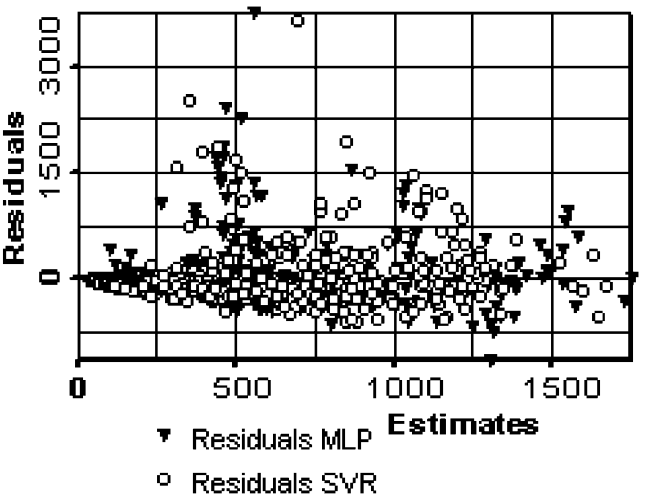

Fig. 9. Scatterplot of the MLP and SVR residuals vs. MLP and SVR estimates, respectively.

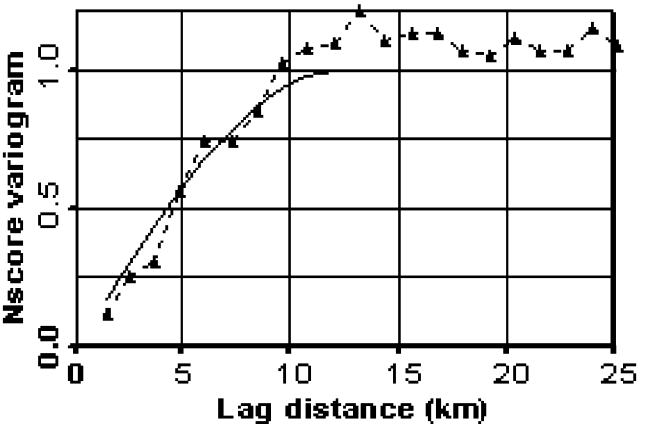

Fig. 10. Nscore omni-directional variogram and the variogram model of the MLP residuals.

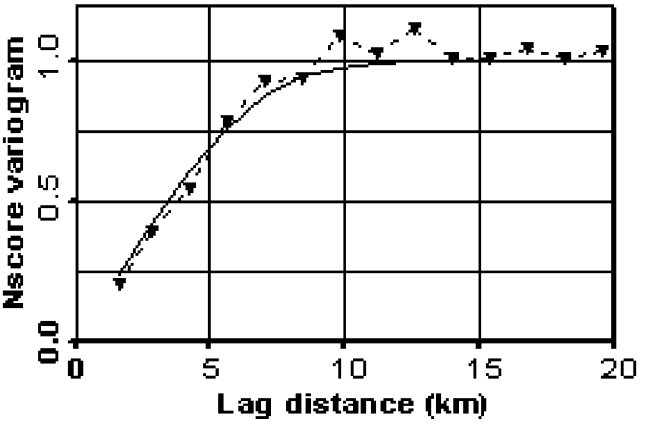

Fig. 11. Nscore omni-directional variogram and the variogram model of the SVR residuals.

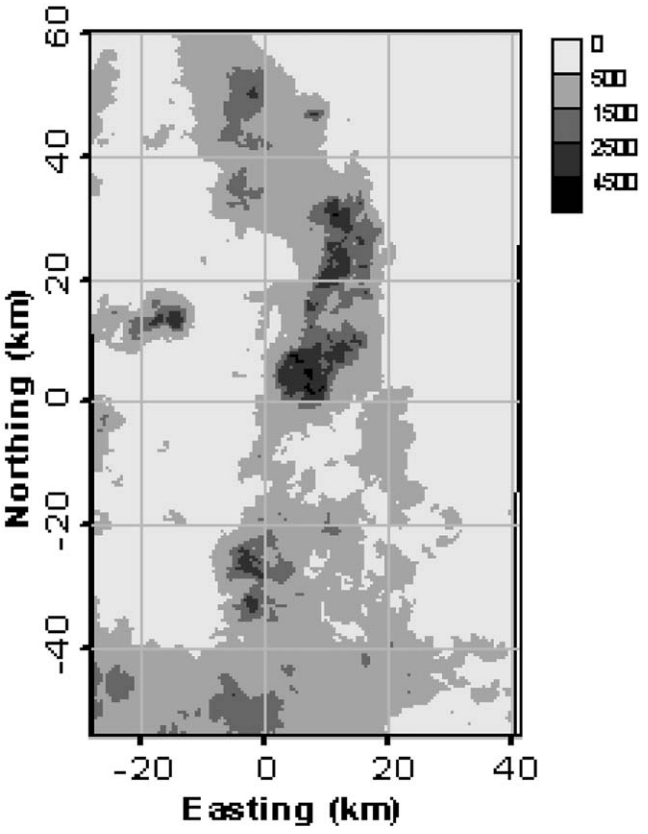

Fig. 12. Mapping of ^{137}Cs with neural network residual sequential Gaussian simulations model (NNRSGS).

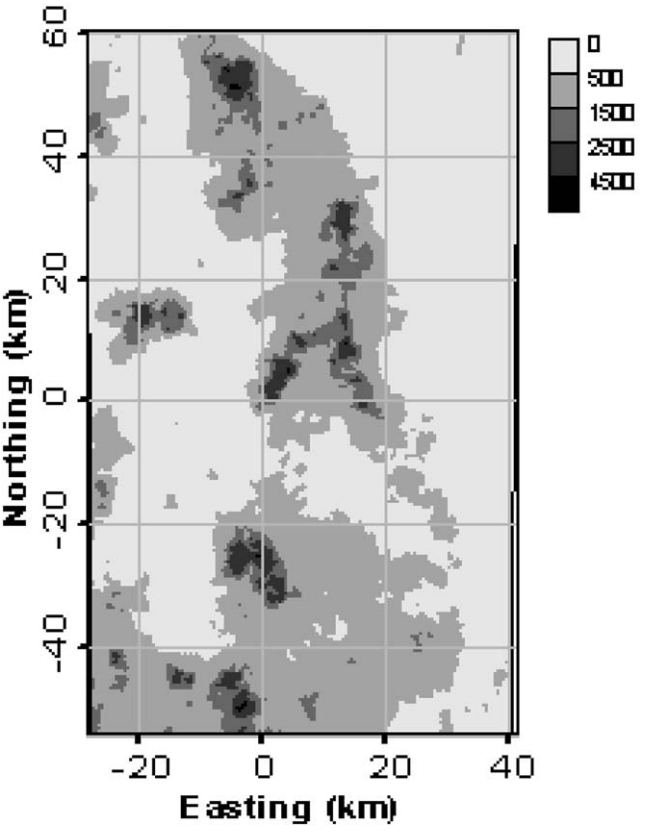

Fig. 13. Mapping of ^{137}Cs with support vector regression residual sequential Gaussian simulations model (SVRRSGS).

of simulated models (realizations) are generated. Post-processing of realizations gives rich variety of outputs, one of them is the probability/risk map. Probability maps of exceeding level 800 kBq/m^2 obtained with neural network/support vector regression residual sequential Gaussian simulation models are presented in Figs. 14 and 15, respectively. This is an important advanced information for the real decision-making process.

4. Discussion

The final stage deals with the validation of the ML residual sequential Gaussian simulation results. comparisons with geostatistical prediction models were carried out. The proposed models give comparable or better results on different data sets. A comparison between proposed models (NNRSGS and SVRRSGS) was also carried out at the testing points. As a result, the NNRSGS model gives better results than the SVRRSGS model in terms of testing error and summary statistics of testing distribution. Comprehensive comparison with other ML methods is a topic of further research.

Several important points should be mentioned:

(1) Analysis of the residuals is important also in case when only ML mapping is applied. This helps to understand the quality of the results. If there is no spatial correlation in the residuals, it means that all spatial information from data have been extracted and ML can be used for prediction mapping as well.

(2) Robustness of the approach: how it is sensitive to the selection of the ML architecture and learning algorithm. Chernov et al. (1999) demonstrated the robustness of MLP with varying number of neurons on validation data. Also, it was shown that MLP is more sensitive towards selection of the training set than towards the number of neurons. The same robust behavior in the case presented in this study has been obtained both for MLP and SVR (varying model parameters). So, we can choose the simplest ML models capable to learn and catch non-linear trends.

Usually, accuracy test (analysis of the residuals) has been used for the analysis and description of what was learned by ML. Accuracy test measures correlation between the training data and the MLA predictions at the same points.

(3) Data clustering is a well-known problem in spatial data analysis (Deutsch and Journel, 1998). This

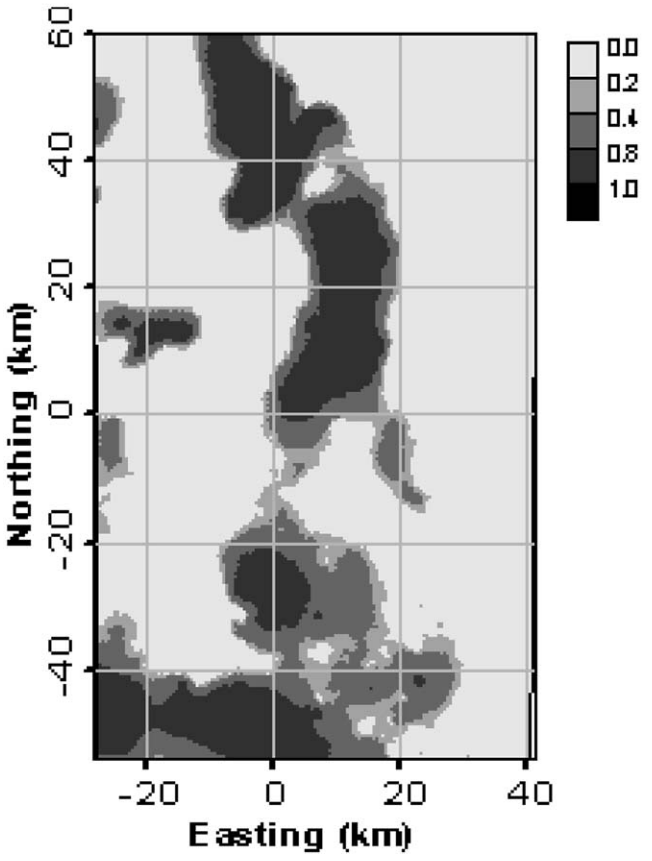

Fig. 14. Probability of exceeding level of 800 kBq/m^2 for NNRSGS model.

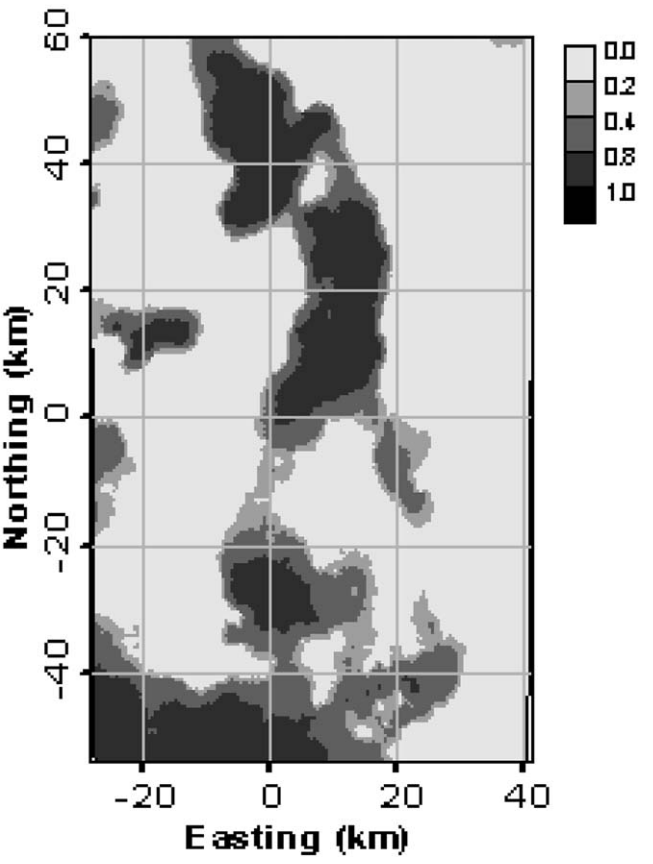

Fig. 15. Probability of exceeding level of 800 kBq/m^2 for SVRRSGS model.

problem is related to the spatial representativity of data. The influence of clustering on the efficiency of ML algorithms should be studied in detail.

5. Conclusions

New non-stationary NNRSGS and SVRRSGS models for the analysis and mapping of spatially distributed data were developed. Non-linear trends in environmental data can be efficiently modeled by a three layer perceptron. Combinations of ML and geostatistical models gave rise to the decision-oriented risk and probabilistic mapping. Promising results presented are based on the unique case study: soil contamination by the most radiologically important Chernobyl radionuclide. Other kinds of ANN models (in particular local approximators) can be used with possible modifications in the proposed framework. ML based models are preferable to pure geostatistical methods because the latter have limitations due to presence of non-linear trends in data, which are difficult to model. Computational costs of the method are rather cheap for a typical geostatistical problem. But the application of the method needs deep expert knowledge in geostatistical modeling. Further, extensions of the approach may deal with multivariate cases as long as ML algorithms are capable of dealing with multivariate information and can integrate different types of data. Extension of the model to image processing requires improving and adaptation of the algorithms, especially from ML side. Recent developments in ML algorithms implementations, see e.g. http://www.torch.ch, are promising from the computational point of view.

The analysis and presentation of the results as well as MLP and Gaussian simulation modeling were performed with the help of GEOSTAT OFFICE software (Kanevski et al., 1999). Support vector regression modeling was carried out with the help of GeoSVM (http://www.ibrae.ac.ru/~mkanev).

Acknowledgements

The work was supported in part by the INTAS grants 99-00099, 97-31726, INTAS Aral Sea project #72, CRDF grant RG2-2236, and Russian Academy of Sciences grant for young scientists research N84, 1999.

References

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Chernov, S., Demyanov, V., Grachev, N., Kanevski, M., Kravetski, A., Savelieva, E., Timonin, V., Maignan, M., 1999. Multiscale Pollution Mapping with Artificial Neural Networks and Geostastistics. Proceedings of the 5th Annual Conference of the International Association for Mathematical Geology (IAMG' 99). Ed. Lippartd, S.J., Nass, A., Sinding-Larsen, R., August 1999, 325-330.

Cressie, N., 1991. Statistics for Spatial Data. John Wiley & Sons, New York.

De Cort, M., Tsaturov, Yu.S., 1996. Atlas on caesium contamination of Europe after the Chernobyl nuclear plant accident. European Commission, Report EUR 16542 EN.

Demyanov, V., Kanevski, M., Savelieva, E., Timonin, V., Chernov, S., Polishuk, V., 2000. Neural Network Residual Stochastic Cosimulation for Environmental Data Analysis. Proceedings of the Second ICSC Symposium on Neural Computation (NC'2000), May 2000, Berlin, Germany, 647-653.

Deutsch, C.V., Journel, A.G., 1998. GSLIB Geostatistical Software Library and User's Guide. Oxford University Press, New York, Oxford.

Dowd, P.A., 1994. In: Dimitrakopoulos, R. (Ed.), The Use of Neural Networks for Spatial Simulation, Geostatistics for the Next Century. Kluwer Academic Publishers, pp. 173–184.

Fedra, K., Winkelbauer, L., 1999. A hybrid expert system, GIS and simulation modeling for environmental and technological risk management. Environmental Decision Support Systems and Artificial Intelligence, Technical Report WS-99-07. AAAI Press, Menlo Park, CA, pp. 1–7.

Gambolati, G., Galeati, G., 1987. Comment on "analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels" by Neuman and Jacobson. Mathematical Geology 19, 249–257.

Gilardi, N., Bengio, S., 2000. Local machine learning models for spatial data analysis. IDIAP-RR 00-34.

Haas, T.C., 1996. Multivariate spatial prediction in the presence of nonlinear trend and covariance nonstationarity. Environmetrics 7.

Haykin, S., 1999. Neural Networks. A Comprehensive Foundation, second ed. Prentice Hall International, Inc.

Isaaks, Ed.H., Shrivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, Oxford.

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

Kanevski, M., Pozdnukhov, A., Canu, S., Maignan, M., Wong, P., Shibli, S., 2001. Support vector machines for classification and mapping of reservoir data. In: Soft Computing for Reservoir Characterization and Modeling. Springer-Verlag, pp. 531–558.

Kanevsky, M., Arutyunyan, R., Bolshov, L., Demyanov, V., Linge, I., Savelieva, E., Shershakov, V., Haas, T., Maignan, M., 1996a. Geostatistical Portrayal of the Chernobyl fallout. In: Baafi, E.Y., Schofield, N.A. (Eds.), Geostatistics '96, Wollongong, vol. 2. Kluwer Academic Publishers, pp. 1043–1054.

Kanevsky, M., Arutyunyan, R., Bolshov, L., Demyanov, V., Maignan, M., 1996b. Artificial neural networks and spatial estimations of Chernobyl fallout. Geoinformatics 7, 5–11.

Masters, Timothy, 1995. Advanced Algorithms for Neural Networks. A C++ Sourcebook. John Wiley & Sons, Inc.

Neuman, S.P., Jacobson, E.A., 1984. Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. Mathematical Geology 16, 499–521.

Pélissier, R., Goreaud, F., 2001. A practical approach to the study of spatial structure in simple cases of heterogeneous vegetation. Journal of Vegetation Science 12, 99–108.

Scholkopf, B., Smola, A., 1998. Learning with Kernels. MIT Press, Cambridge, MA.

Vapnik, V., 1998. Statistical Learning Theory. John Wiley & Sons, New York.