IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

ADVANCED SPATIAL DATA ANALYSIS AND MODELLING WITH SUPPORT VECTOR MACHINES

Mikhail Kanevski
Aleksey Pozdnukhov[1]
Michel Maignan[3]
Stephane Canu[2]

IDIAP-RR-00-31

Submitted to
International Journal of Fuzzy Systems.

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[1] Institute of Nuclear Safety, Russian Academy of Sciences, B. Tulskaya 52, 113191 Moscow
[2]INSA; Place Emile Blondel, 76131 Mont-Saint-Aignan, France
[3] Institute of Mineralogy and Petrology, University of Lausanne, BFSH2, 1015 Lausanne, Switzerland

Advanced Spatial Data Analysis and Modeling with Support Vector Machines

M. Kanevski, A. Pozdnukhov, S. Canu, M. Maignan

***Abstract*-- The present paper deals with novel developments and application of Support Vector Machines (Support Vector Classifier SVC and Support Vector Regression SVR) for the analysis and modeling of spatially distributed environmental and pollution information (categorical and/or continuous data). SVC/SVR models are based on the Statistical Learning Theory or Vapnik-Chervonenkis (VC)-theory. The SVC provide non-linear classification by mapping the input space into high dimensional feature spaces where a special type of hyper-planes with maximal margins (giving rise to good generalizations) are constructed. SVR provide robust non-linear regression of spatially distributed data. Real case studies of the present paper deal with binary classification problem of indicator variables, multi-class classification of soil types, and prediction mapping of radioactively contaminated territories. Geostatistical tools (variography) is used to control the performance of the machines and for better understanding of the results. The SVC/SVR are well adapted to fuzzy environmental and pollution data.**

Index Terms—environmental spatial data classification and mapping, support vector machines, geostatistics

I. INTRODUCTION

Recently the analysis and processing of spatially distributed and time dependent information have become a very important problem due to the comprehensive development of environmental and pollution monitoring networks even leading to data mining problems from one side and much better understanding of data analysis approaches (both model dependent and data driven) from another side.

The present paper deals with novel developments and adaptation of SVC and SVR, the models based on Statistical Learning Theory or Vapnik-Chervonenkis (VC)-theory for the analysis and modeling of spatially distributed environmental and pollution information (categorical and/or continuous data).

Statistical Learning Theory is a general mathematical framework for estimating dependencies from empirical and finite data sets [1]-[4]. The basic idea of SVM is to determine a classifier or regression machine that minimizes Structural Risk consisting of the empirical error and the complexity of the model leading to good generalization error.

The SVM provides non-linear classification (or regression) by mapping the input space into high dimensional feature spaces where a special type of hyper-planes with maximal margins (giving rise to good generalizations – low errors on validation data sets) are constructed.

In case of classification SVM are focusing on the marginal data (support vectors - SV) and not on statistics such as means and variances. Only data points close to the classification decision boundaries are important for the solution of the problem. Essentially the method is non-linear, robust and does not depend on the dimension of input space.

Recently first promising results on application of SVC/SVR for the spatially distributed data were published [5]-[7]. The main attention was paid to binary classification problems and to understanding of SVR application to spatial data and interpretation of SVR hyper-parameters. Results of the SVM classification were compared with indicator kriging [6] as well. It was demonstrated that the use of geostatistical spatial correlation measures like variogram improved both understanding of the machine performance and interpretation of the results.

The main attention in the present paper is paid to: 1) the problem of SVC multi-class classification of environmental data – soil types that is important, e.g. in modeling of radionuclides vertical migration, and 2) to SVR mapping of radioactively contaminated territories by Sr90 Chernobyl radionuclide.

Originally SVC were developed for the binary (2 class) classification. Different generalization schemes of 2-class classification problem to multi-class classification are considered in the present study. The methodology and the results on soil types classification are considered in detail. Radial basis Gaussian functions (both isotropic and anisotropic) were used as the SVC kernels. Error surfaces (training and testing errors and number of support vectors versus regularization parameter and kernel

M. Kanevski is with the IDIAP Dalle Molle Institute of Perceptual Artificial Intelligence, CP 592, 1920 Martigny, Switzerland, kanevski@idiap.ch and with INSA, Rouen, France.
A. Pozdnukhov is with the Moscow State University, Physics Department
S. Canu is with the INSA, Rouen, France

bandwidth) are used in order to tune hyper-parameters. Particularly, kernel bandwidths are tuned using testing data sets and taking into account spatial variability of classes.

Several approaches for the classification of spatially distributed data that were developed within the framework of geostatistics can be found in the review [8].

The last part of the paper presents results on application of Support Vector Regression to the problem of prediction mapping of spatial data. Real case study is based on data on soil contamination by Sr90 Chernobyl radionuclide in the most contaminated Briansk region of Russia. Sr90 has 30 years half-time decay and is radiologically important.

II. INTRODUCTION TO SUPPORT VECTOR MACHINES

The main concepts and principles of SVM are described shortly, starting from lineally separable dichotomies. The presentation of the SVM theory is based on [1]-[4].

A. Principles of SVM

The following problem is considered. A set S of points (X_i) is given in R^2 (we are working in a two dimensional [X_1, X_2] space). Each point X_i belongs to either of two classes and is labeled by $Y_i \in \{-1,+1\}$. The objective is to establish an equation of a hyper-plane that divides S leaving all the points of the same class on the same side while maximizing the minimum distance between either of the two classes and the hyper-plane – maximum margin hyper-plane.

Optimal hyper-plane with the largest margins between classes is a solution of the constrained optimization problems considered below [1]-[4].

B. Linearly separable case

Let us remind that data set S is linearly separable if there exist $W \in R^2, b \in R$, such that

$$Y_i(W^T X_i + b) \geq +1, \quad i = 1,...N \qquad (1)$$

The pair (W,b) defines a hyper-plane of equation

$$(W^T X + b\ ) = 0\ .$$

Linearly separable problem: Given the training sample $\{X_i,Y_i\}$ find the optimum values of the weight vector W and bias b such that they satisfy constraints

$$Y_i(W^T X_i + b) \geq +1, \quad i = 1,...N \qquad (2)$$

And the weight vector W minimizes the cost function (maximization of the margins)

$$F(W) = W^T W / 2 \qquad (3)$$

The cost function is a convex function of W and the constraints are linear in W.

This constrained optimization problem can be solved by using Lagrange multipliers. Lagrange function is defined by

$$L(W,b,\alpha) = W^T X\ / 2 - \sum_{i=1}^{N} \alpha_i \left[Y_i(W^T X_i + b) - 1 \right]$$

where Lagrange multipliers $\alpha_i \geq 0$.

The solution of the constrained optimization problem is determined by the saddle point of the Lagrangian function $L(W,b,\alpha)$ which has to be minimized with respect to W and b and to be maximized with respect to α .

Application of optimality condition to the Lagrangian function yields

$$W = \sum_{i=1}^{N} \alpha_i Y_i X_i \qquad (4)$$

$$\sum_{i=1}^{N} \alpha_i Y_i = 0 \qquad (5)$$

Thus, the solution vector $\boldsymbol{W}$ is defined in terms of an expansion that involves the N training data.

Because of constrained optimization problem deals with a convex cost function, it is possible to construct dual optimization problem. The dual problem has the same optimal value as the primal problem, but with the Lagrange multipliers providing the optimal solution.

The dual problem is formulated as follows:

Maximize the objective function

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - (1/2) \sum_{i=1}^{N} \alpha_i \alpha_j Y_i Y_j X_i^T X_j \qquad (6)$$

Subject to the constraints

$$\sum_{i=1}^{N} \alpha_i Y_i = 0 \qquad (7)$$

$$\alpha_i \geq 0, \mathrm{i} = 1,...\mathrm{N} \qquad (8)$$

Note that the dual problem is presented only in terms of the training data. Moreover, the objective function $Q(\alpha)$ to be maximized depends only on the input patterns in the form of a set of dot products $\{X_i^T X_j\}_{i=1,2,...N}$.

After determining optimal Lagrange multipliers α_{i0}, the optimum weight vector is defined by (4) and bias is calculated as $b = 1 - W^T X_i^S$, for $Y^{(s)} = +1$

Note, that from the Kuhn-Tucker conditions it follows that

$$\alpha_i \left[Y_i (W^T X_i + b) - 1 \right] = 0 \qquad (9)$$

Only α_i that can be nonzero in this equation are those for which constraints are satisfied with the equality sign. The corresponding points X_i, called *Support Vectors*, are the points of the set $\boldsymbol{S}$ closest to the optimal separating hyper-plane. In many applications number of support vectors is much less that original data points.

The problem of classifying a new data point X is simply solved by computing

$$F(X) = sign(W^T X_i + b) \qquad (10)$$

with the optimal weights *W* and bias *b*.

C. SVM classification of non-separable data. Soft margin classifier (allowing for training errors)

In case of linearly non-separable set it is not possible to construct a separating hyper-plane without allowing classification error. The margin of separation between classes is said to be soft if training data points violate the condition of linear separability.

In case of non-separable data the primal optimization problem is changed by using slack variables.

Problem is posed as follows: Given the training sample $\{X_i, Y_i\}$ find the optimum values of the weight vector W and bias b such that they satisfy constraints

$$Y_i (W^T X_i + b) \geq +1 - \xi_i, \ \xi_i \geq 0, \forall i \qquad (11)$$

The weight vector *W* and the slack variable ξ_i minimize the cost function

$$F(W) = W^T W / 2 + C\sum_{i=1}^{N} \xi_i \qquad (12)$$

where C is a user specified parameter (regularization parameter is proportional to 1/C).

The dual optimization problem is the following: Given the training data maximize the objective function (find the Lagrange multipliers)

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - (1/2)\sum_{i=1}^{N} \alpha_i \alpha_j Y_i Y_j X_i^T X_j \qquad (13)$$

Subject to the constraints (7) and

$$0 \le \alpha_i \le C, \mathrm{i} = 1,...\mathrm{N} \qquad (14)$$

Note that neither the slack variables nor their Lagrange multipliers appear in the dual optimization problem.

The parameter C controls the trade-off between complexity of the machine and the number of non-separable points.

The parameter C has to be selected by user. This can be done usually in one of two ways: 1) C is determined experimentally via the standard use of a training and testing data sets, which is a form of re-sampling; 2) It is determined analytically by estimating VC dimension and then by using bounds on the generalization performance of the machine based on a VC dimension [1].

D. SVM non-linear classification

In most practical situations the classification problems are non-linear and the hypothesis of linear separation in the input space are too restrictive.

The basic idea of Support Vector Machines is 1) to map the data into a high dimensional feature space (possibly of infinite dimension) via a non-linear mapping and 2) construction of an optimal hyper-plane (application of the linear algorithms described above) for separating features. The first item is in agreement of Cover's theorem on the separability of patterns which states that input multidimensional space may be transformed into a new feature space where the patterns are linearly separable with high probability, provided: 1) the transformation is non-linear; 2) the dimensionality of the feature space is high enough [1]-[4]. Cover's theorem does not discuss the optimality of the separating hyper-plane. By using Vapnik's optimal separating hyper-plane VC dimension is minimized and generalization is achieved. Let us remind that in the linear case the procedure requires only the evaluation of dot products of data.

Let $\{\varphi_j(x)\}_{j=1,...m}$ denote a set of non-linear transformation from the input space to the feature space; *m* – is a dimension of the feature space. Non-linear transformation is defined a priori.

In the non-linear case the optimization problem in the dual form is following:

Given the training data maximize the objective function (find the Lagrange multipliers)

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - (1/2)\sum_{i=1}^{N} \alpha_i \alpha_j Y_i Y_j K(X_i^T X_j) \qquad (15)$$

Subject to the constraints (7) and (14)

The kernel

$$K(X,Y) = \varphi^T(X)\varphi(Y) = \sum_{j=1}^{m} \varphi_j(X)\varphi_j(Y) \qquad (16)$$

Thus, we may use inner-product kernel $K(X,Y)$ to construct the optimal hyper-plane in the feature space without having to consider the feature space itself in explicit form.

The optimal hyper-plane is now defined as

$$f(X) = \sum_{j=1}^{N} \alpha_j Y_j K(X, X_j) + b \qquad (17)$$

Finally, the non-linear decision function is defined by the following relationship:

$$F(X) = sign\left[W^T K(X, X_j) + b\right] \qquad (18)$$

The requirement on the kernel $K(X, X_j)$ is to satisfy Mercer's conditions [1]. Three common types of Support Vector Machines are widely used:

1. Polynomial kernel

$$K(X, X_j) = (X^T X_i + 1)^p$$

where power p is specified a priori by the user. Mercer's conditions are always satisfied.

2. Radial basis function RBF kernel

$$K(X, X_j) = \exp\left\{-\left\|X - X_j\right\|^2 / 2\sigma^2\right\}$$

Where the kernel bandwidth σ (sigma value) is specified a priori by the user. In general, Mahalanobis distance can be used. Mercer's conditions are always satisfied.

3. Two-layer perceptron

$$K(X, X_j) = \tanh\left\{\beta_0 X^T X_j + \beta_0\right\}$$

Mercer's conditions are satisfied only for some values of β_0 , β_1.

For all three kernels (learning machines), the dimensionality of the feature space is determined by the number of support vectors extracted from the training data by the solution to the constrained optimization problem. In contrast to RBF neural networks, the number of radial basis functions and their centers are determined automatically by the number of Support Vectors and their values. In the present study only the results obtained with the RBF kernel are presented.

III. SPATIAL DATA CLASSIFICATION. CASE STUDIES

Two classification case studies are considered:

- Binary non-linear classification of radioactively contaminated territories (Briansk region, Sr90). This part of the study is of methodological nature and follows the ideas presented in [5]-[6]. Training algorithms are extended with k-fold cross-validation (leave-k-out).
- Multi-class classification of real soil types data in Briansk region, Russia. This case study is important for prediction mapping of radioactively contaminated territories, when taking into account radionuclides vertical migration in soil.

The generic methodology for the analysis, modeling and presentation of spatially distributed data follows the basic ideas presented in [10]. The main phases (steps) of the study are following:

- Visualization of data. Monitoring network analysis and description. Understanding of spatial clustering (results of preferential sampling) and representativity of data.
- Comprehensive exploratory data analysis.
- Comprehensive exploratory structural analysis (variography). Modeling of anisotropic spatial correlation.
- Splitting data into training, testing, and validation subsets. In case of clustered data spatial declustering procedures can be sued.
- Training of SVC/SVR. Selection of the optimal SVC/SVR hyper-parameters.
- Spatial data classification - categorical data mapping.
- Spatial data mapping – spatial regression.
- Comprehensive analysis of the residuals (statistical analysis, correlation, variography)
- Understanding, interpretation, and presentation of the results.

A. Two class classification problem

Let us consider binary classification problem applied to Sr90 indicator transformed variable

I(Sr90=0.3 Ci/km^2.). Indicator transformation means, that I(Sr90=0.3 Ci/km^2.) = I = 1 if Sr90≤0.3 Ci/km^2 (class 1) and I=0 if Sr90>0.3 Ci/km^2 (class 2). Thus, the problem is posed as a binary classification problem after the indicator transformation of Sr90 concentration. Here, the indicator is chosen close to the median of Sr90 concentration.

In non-parametric geostatistics indicator transformation is widely used when modeling local probability density functions: expected value of indicator at unsampled point is an estimation of the probability density function at this point with a given cut [9].

The post plot of indicator values are presented in Figure 1. Variogram rose for the indicator variable is presented in Figure 2. Let us remind, that variogram (semivariogram) is an important measure of spatial continuity describing spatial correlation and widely used in geostatistics [9]-[10]:

$$\gamma(\mathbf{h}) = 0.5\,\mathrm{Var}\{Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})\}$$

where **h** – is a separation vector between points in space. Variogram, estimated using indicator variable for several lag distances and in several directions is presented as a Variogram rose in Figure 2. Geostat Office software [10] was used for computations. Anisotropic structure – different correlations in different directions is evident.

Information on spatial correlation can be used in data pre-processing: one objective can be a transformation of input space (**X**) in order to have more isotropic spatial correlation structures. Also this information can be used to tune anisotropic SVM kernels when Mahalanobis distance is used.

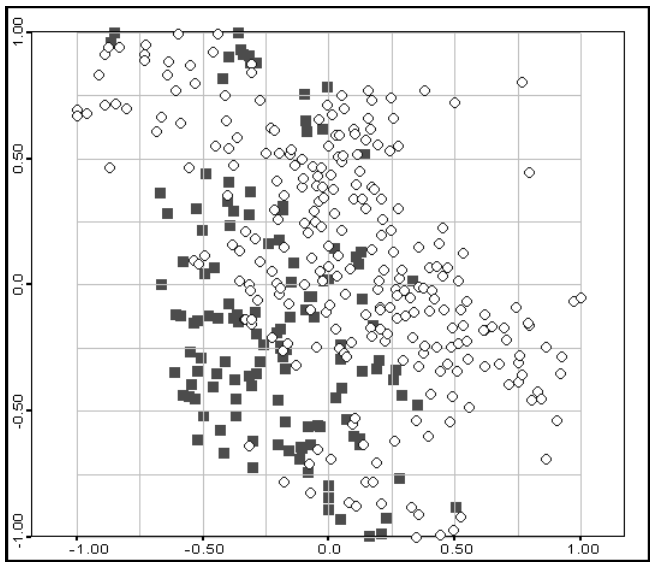

Figure 1. Two class classification problem. Training data set postplot. "O" – class 1, "■" – class 2.

1) SVC training

Two basic strategies were applied for the SVM training: 1) splitting of original data set into training, testing and validation subsets; 2) leave-k-out cross-validation. The first approach is a traditional procedure when training data set is used to develop a model, testing data set is used to tune hyper-parameters of the model, and validation data set is used for the generalization (expected) error estimation. Taking into account spatial clustering – preferential sampling in space, spatial declustering procedures were used to split data in order to have representative data sets. The simplest way to do it is to cover the region under study by a regular grid and to select randomly one data from each grid cell. Random splitting was used as well.

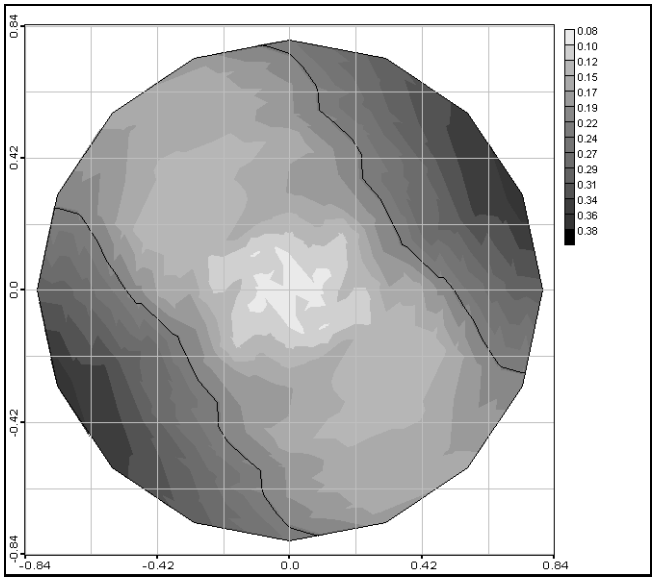

Figure 2. Variogram rose of Sr90 indicator variable.

There are two hyper-parameters in SVM when RBF kernel is fixed: kernel bandwidth (sigma) and regularization parameter C. In general, full covariance matrix (Mahalanobis distance) was used. In the present study the results of the isotropic kernel RE mainly presented.

Basically, there is a general recommendation to put C as a big value when data are not noisy and there is no special need in regularization. In order to find the best (minimizing testing error) C and sigma parameters training and testing error surfaces (training and testing errors versus sigma and C) were estimated. It was found that, after some high C values, when sigma is fixed, training and testing errors do not change. In our case it was about 100 at optimal sigma value. The error curves along with normalized number of Support Vectors (the number of Support Vectors divided by the number of training data) are presented in Figure 3. The minimal testing error was achieved at sigma = 0.1. An important observation, already mentioned in [5] and [6] is that at the optimal point the number of Support Vector has also minimum. This, in general, corresponds to small values of generalization (expected) errors [1].

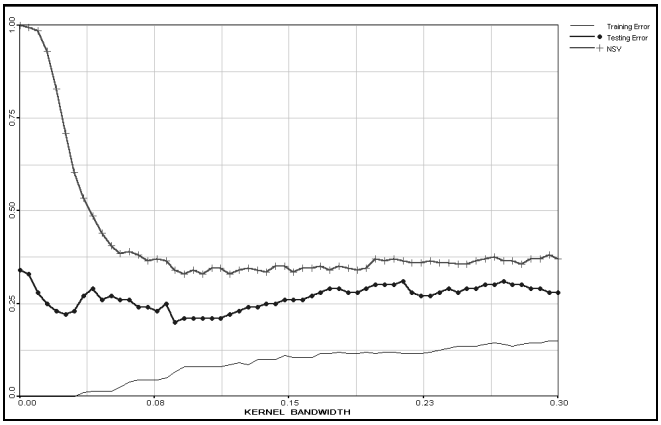

Figure 3. Training and testing error curves and normalized number of Support Vectors. C=100.

2) Binary classification

The optimal SVC hyper-parameters were used for the categorical data mapping (prediction of categorical variable/class at unsampled points). The result is presented in Figure 4. Variogram rose computed using the results of SVC classification is presented in Figure 5. Except with some noise this variogram rose follows the original experimental variogram rose. Thus, classification model correctly reflects basic anisotropic spatial correlations.

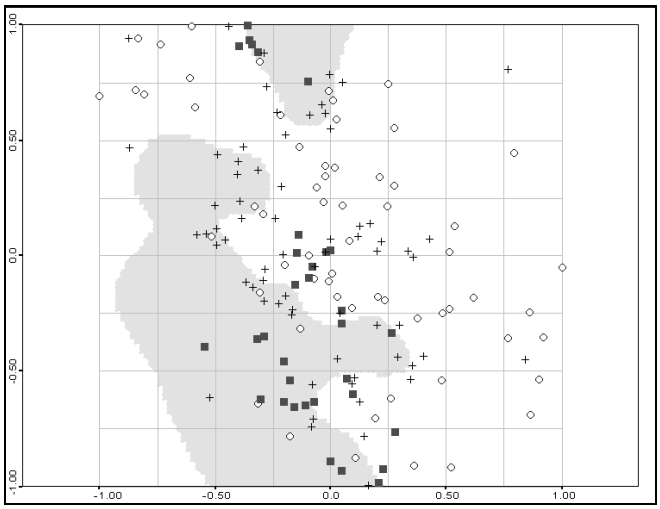

Figure 4. SVM 2 class classification (categorical data mapping). White zone – class 2. Kernel bandwidth = 0.1, C=100. Training error = 0.08; testing error = 0.21; validation error = 0.24. "+" – Support Vectors; "O" – class 2 of validation data; " ■ " – class 1 of validation data.

Basically, by varying kernel bandwidth at some fixed C value, it is possible to cover wide range of model's complexity from overfitting at small sigma values to oversmoothing at high sigma values.

In the following, a real case study on multi class classification using data on soil types in Briansk region, Russia.

IV. SVM MULTI-CLASS CLASSIFICATION

The current section of the work deals with the soil types prediction mapping using Support Vector Machines. The main objective of the study is following: using available categorical data on soil types (measurements on an irregular monitoring networks) develop multi-class classification Support Vector Machine to predict soil types at the unsampled points (spatial prediction of categorical variables). The problem can be considered as a pattern completion task as well.

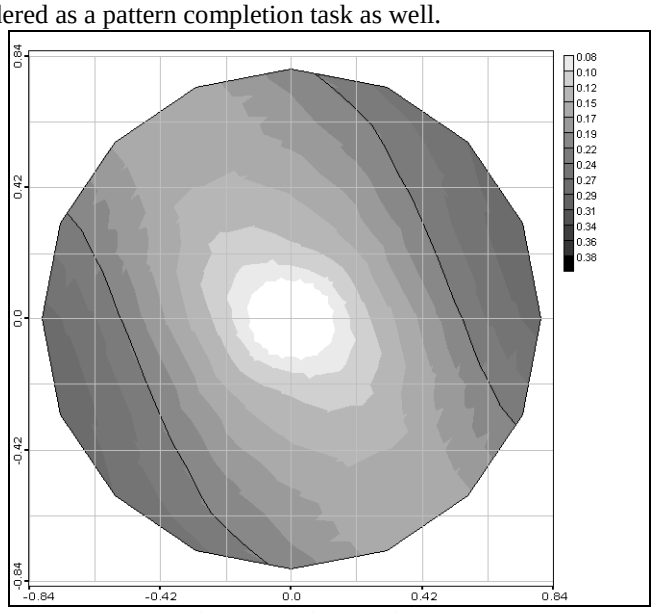

Figure 5. SVM classification. Variogram rose of indicators after classification.

In the present study, SVC are used for environmental spatial data classification. Straightforward generalization of binary SVM classification to multi class classification (m classes) is the following:

$$y_j = \arg \max_m \sum_i \lambda_i^{(m)} y_i K(x, x_i) + b^{(m)} \qquad (19)$$

The real case study deals with the soil types classification in Briansk region. This is the most contaminated part of Russia by Chernobyl radionuclides. Actually, prediction mapping of environment

contamination includes both physico-chemical modeling of radionuclides migration in environment and spatial data analysis and modeling [11]. Migration of radionuclides in soil depends on properties of radionuclides, soil types, precipitation, etc. Variability of environmental parameters and initial fallout at different scales highly complicates the solution of the problem.

The present problem deals with five classes:

5 Classes data	Number of data
Class1	392
Class2	48
Class3	333
Class4	52
Class5	485

The grid for predictions consists of 4321 points (the boundary of the grid follows the boundary of the region).

The influence of soil types on Sr90 vertical migration is presented in Figure 6, where Sr90 profiles after 20 years of fallout are presented.

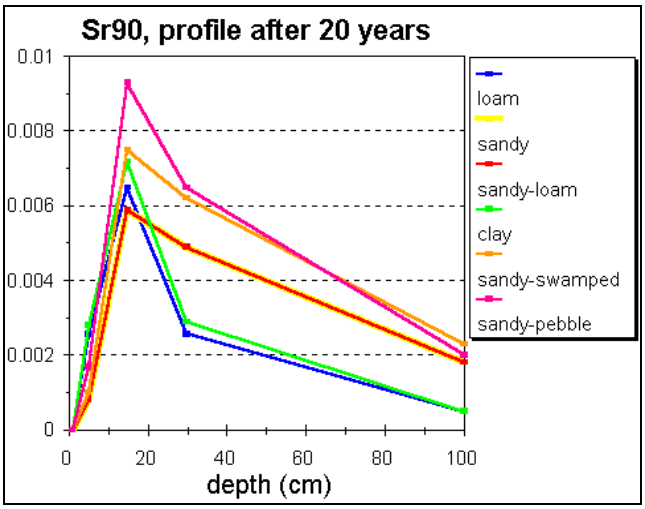

Figure 6. Radionuclide vertical migration in soil. Vertical profile of Sr90 distribution after 20 years of fallout.

The major classes (post plot of training data) are presented in Figure 7.

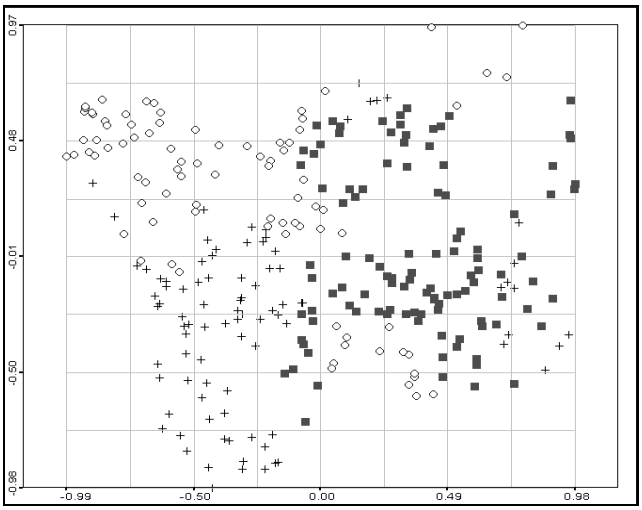

Figure 7. Major classes (soil types data) postplot. “+” – class 1, “O” – class 3, “■” – class 5.

Like in the case of binary classification, original data were split into 3 subsets: training (310), testing (500) and validation (500 data). Data were split several times to understand fluctuations of the results.

Spatial correlation structures for two major classes are presented as Variogram roses in Figures 8 and 9. Classes were coded as indicators with 1 corresponding to class and 0 to all other classes. Different correlation behavior is clearly observed.

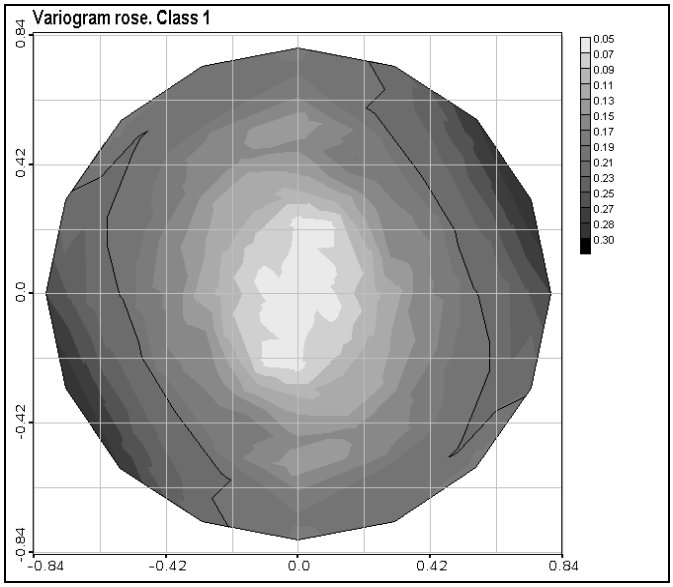

Figure 8. Class 1 variogram rose.

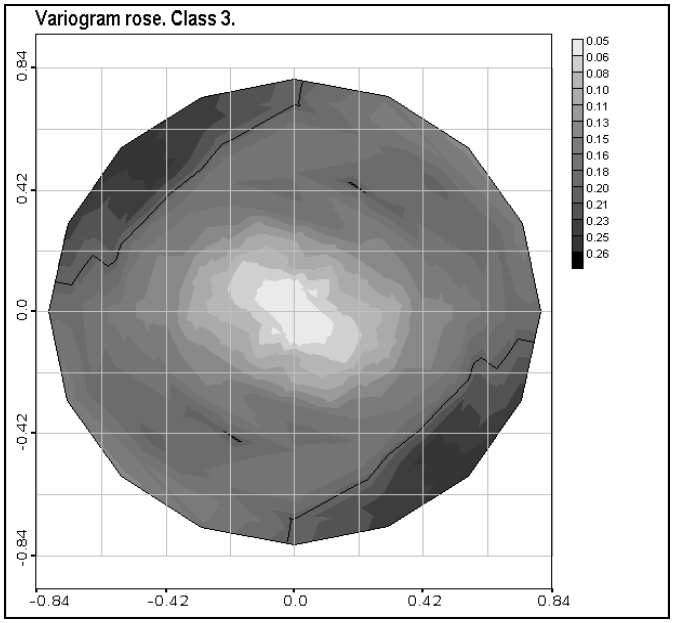

Figure 9. Class 3 variogram rose.

A. SVC Training

There are several possibilities for the multi-class classification with SVM using binary models: one-to-rest classification, pair-wise classification, direct generalization of the SVM to multi-class problems and others [1], [12], [13].

One-to-Rest class-insensitive classification. In this case m- models are developed from binary classification by applying the most simple algorithm. m-classifiers have the same kernel bandwidths.

Error curves give general overview of the problem without taking into account different spatial variability of classes. If classes have different variability at different scales and directions the “optimal kernel bandwidth” characterizes some averaged scale of variability. Of course, what is optimal for one

class, can be over-fitting or over-smoothing for the others. Class insensitive approach is fast and gives general overview of the problem. In some cases it can give satisfactory results. The more interesting approach deals with adaptation of models to spatial variability of classes.

1) Class-Adaptive Approach

In this case for each one-to-rest M models different optimal kernel bandwidths are tuned. Training and testing error curves with class adaptive technique are presented in Figure 10.

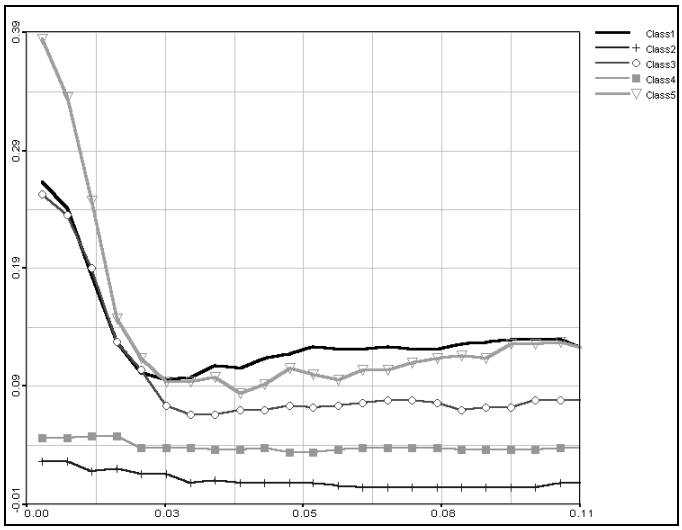

Figure 10. One-to-rest multi-class classification. Testing error curves.

For each one-to-rest model optimal kernel bandwidths minimizing testing errors were selected. Spatial predictions of categorical variable (soil type mapping) with optimal m models are presented in Figure 12.

The same approach was applied with the generalization of binary model using pair-wise classifications, both class insensitive and class adaptive. In this case m(m-1)/2 are developed. Example of training testing and normalized number of Support Vectors curves is presented in Figure 11.

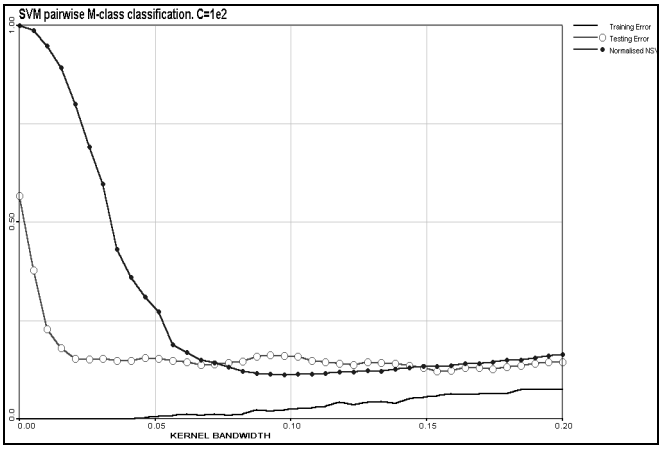

Figure 11. Pair-wise training and testing error curves and normalized number of Support Vectors. C= 100.

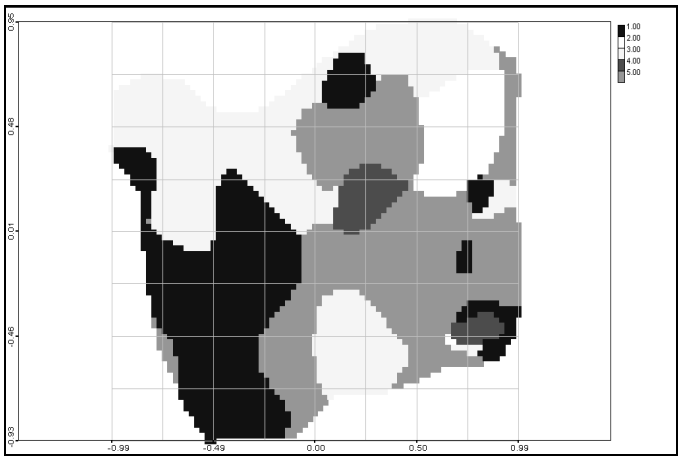

Figure 12. SVM Mapping with class-adaptive bandwidths: Class1 =0.026; Class2 = 0.1; Class3 = 0.14; Class4 = 0.06; Class5 = 0.88.

In the present case pair-wise classification did not improve significantly the results in comparison with simpler one-to-rest adaptive model.

In conclusion, SVC is a promising approach for the classification of spatially distributed environmental and pollution data. The use of simple multi class classification models (generalizations to the binary models) with class adaptive approach efficiently reproduced spatial variability of classes.

V. POLLUTION DATA MAPPING WITH SUPPORT VECTOR REGRESSION

Let us consider application of the Statistical Learning Theory for spatial data mapping of continuous variables using Support Vector Regression model.

Assume $Z \in R$ is a variable to be predicted based on some geographical observations (x,y). Our work aims at estimating a dependence between Z and the geographical co-ordinates based on empirical data (samples) $S_n=(x_i,y_i,Z_i,\varepsilon_i)$, i = 1,…n, where

x_i,y_i, - are the geographical co-ordinates of samples

Z_i - are observed or measured quantities. It is assumed to be the realization of a random variable Z_i with an unknown probability distribution $P_{x,y}(Z)$.

ε_i - is the measurement accuracy for the observation Z_i

n denotes the sample size

A. *2.2 Prediction problem*

Assuming f is a prediction function (i.e. a function used to predict the value of Z knowing the geographical co-ordinates), we define the cost of choosing this particular function for a given decision process. First, for a given observation (x,y,Z) we define the ε-insensitive cost function:

$$C\{(x,y),Z,\varepsilon,f\}=\begin{cases}|f(x,y)-Z|-\varepsilon & \text{if } |f(x,y)-Z|>\varepsilon \\ 0 & \text{otherwise}\end{cases} \quad (20)$$

where ε characterizes some acceptable error.

Now, for all possible observations we define the global or generalisation error also known as the integrated prediction error IPE:

$$IPE(f)=\int E_Z(C((x,y,)Z,\varepsilon,f))\omega(x,y)dxdy \quad (21)$$

where ω(x,y) is some economical measure, indicating the relative importance of a mistake at point (x,y). In case of non-homogeneous monitoring networks this function can take into account spatial clustering. Usually ω(x,y) = 1, so that all positions ar/e assumed to be equally important.

B. *2.3 Empirical and Structural Risk Minimization*

1) *2.3.1 Function Modeling*

Let us assume that solution is a function that can be decomposed into two different components: a

trend plus a remaining random process.

$$\hat{f}(x,y) = \sum_{k=1}^{m} w_k \varphi_k(x,y) + \sum_{j=1}^{J} \beta_j K_j(x,y) \qquad (22)$$

where $K_j(x,y)$ is a basis of the trend component and φ_k, k=1,..m is an orthonormal basis of the remaining part (note that m can be infinity).

The complexity of the solution can be tuned through $||w||^2 = \sum_{k=1..m} w_k^2$ [1]. Thus, a relevant strategy to minimise IPE is to minimize the empirical error together with maintaining $||w||^2$ small. This can be obtained by minimising the following cost function:

$$\begin{cases} \text{minimize} \quad \frac{1}{2}||w||^2 \\ \text{subject to} \quad |f(x_i,y_i) - Z_i| \le \varepsilon_i \text{ , for i = 1,...n} \end{cases}$$

When data lie outside of this epsilon tube due to noise or outliers making these constraints too strong and impossible to fulfil, Vapnik suggested to introduce slack variables ξ_i , ξ_i^* . These variables measure the distance between the observation and the ε tube.

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

Following the ideas as in the case of SVM classification we arrive at the following optimization problem:

$$\text{minimise} \quad \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \qquad (23)$$

$$\text{subject to} \begin{cases} f(x_i,y_i) - Z_i - \varepsilon_i \le \xi_i \\ -f(x_i,y_i) + Z_i - \varepsilon_i \le \xi_i^* \\ \xi, \xi_i^* \ge 0 \qquad \text{for i = 1,...n} \end{cases}$$

2) 2.3.2 Dual formulation

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

$$L(w,\xi,\xi^*\alpha) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(Z_i - f(x_i,y_i) + \varepsilon_i + \xi_i) - \sum_{i=1}^{n}\alpha_i^*(f(x_i,y_i) - Z_i + \varepsilon_i + \xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

where $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ are the Lagrangian multipliers associated with the constraints. They can be roughly interpreted as a measure of the influence of the constraints in the solution. A solution with $\alpha_i = \alpha_i^* = 0$ can be interpreted as "the corresponding data point has no influence on this solution".

Finally, the dual formulation of the problem is as follows:

$$\begin{aligned} \text{maximise} \quad & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i^* - \alpha_i)\left(\sum_{k=1}^{m}\varphi_k(x_i,y_i)\varphi_k(x_j,y_j)\right)(\alpha_j^* - \alpha_j) \\ & -\sum_{i=1}^{n}\varepsilon_i(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n} Z_i(\alpha_i^* - \alpha_i) \\ \text{subject to} \quad & \begin{cases} \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)K_j(x_i,y_i) = 0 \quad \text{for } K_j = 1,...m \\ 0 \le \alpha_i^*, \alpha_i \le C \qquad \text{for i,...n} \end{cases} \end{aligned} \qquad (24)$$

By using kernel trick this problem can be solved without direct modeling in a feature space (the same as in non-linear classification). To do so it is necessary to choose φ_k such that:

$$\sum_{k=1}^{m} \varphi_k(x_i, y_i)\varphi_k(x_j, y_j) = G\big((x_i, y_i),(x_j, y_j)\big)$$

This is the case in reproducing kernel Hilbert space, where G is the reproducing kernel. Functions φ_k are the eigen functions of G. In this case the solution can be formulated in the following form:

$$\hat{f}(x,y) = \sum_{i=1}^{n} v_i G((x,y),(x_i,y_i)) + \sum_{j=1}^{m} \beta_j K_j(x,y)$$

with $v_i = (\alpha_i^* - \alpha_i)$. This solution only depends on the kernel function G. The main results were obtained with Gaussian RBF kernel and $K_j(x,y)=1$.

VI. SVR MAPPING. CASE STUDY

Let us consider mapping of soil pollution by Chernobyl radionuclide Sr90 in the Western part of Briansk region, Russia.

The case study follows the basic methodology applied to the classification in the previous sections. An important development deals with comprehensive analysis of the residuals. In terms of geostatistics useful information to be extracted from data and modeled with SVR is a spatially structured (spatially correlated) information. From this point of view variography of the residuals is a powerful and efficient tool for controlling the performance of SVR mapping.

The variogram rose of training Sr90 data is presented in Figure 13.

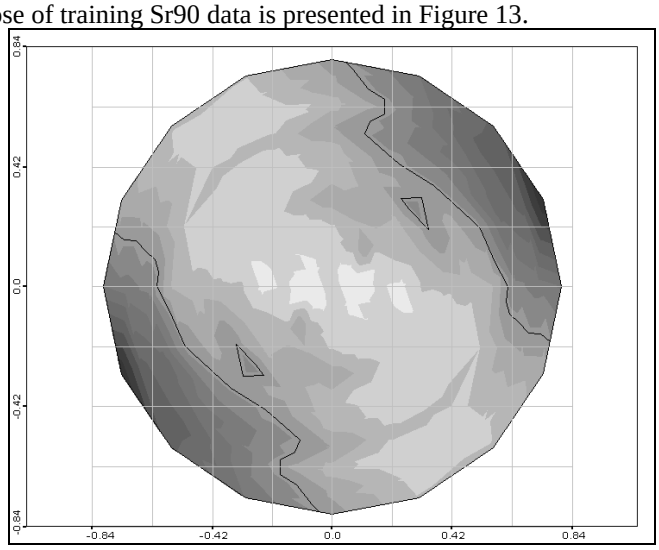

Figure 13. Variogram rose of Sr90 raw data.

In case of regression when Gaussian RBF kernel is fixed there are three hyper-parameters: kernel bandwidth, regularization constant C and ε . Therefore, an error cube has to be estimated and analyzed to find optimal SVR parameters. Some ideas on the selection of hyper-parameters are discussed in [7]. Training and testing error surface are presented in Figures 14, 15.

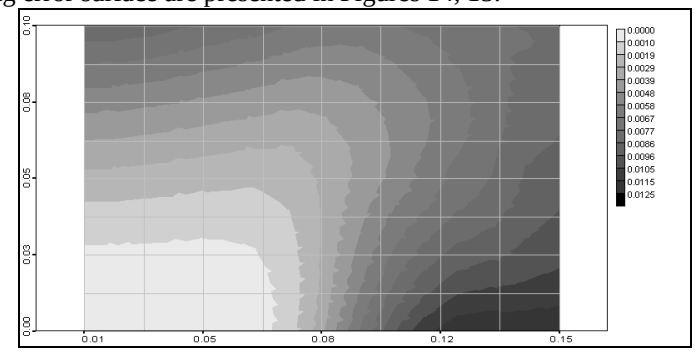

Figure 14. Training error surface, C= 1000. Axes correspond to X – kernel bandwidth; Y - ε parameter.

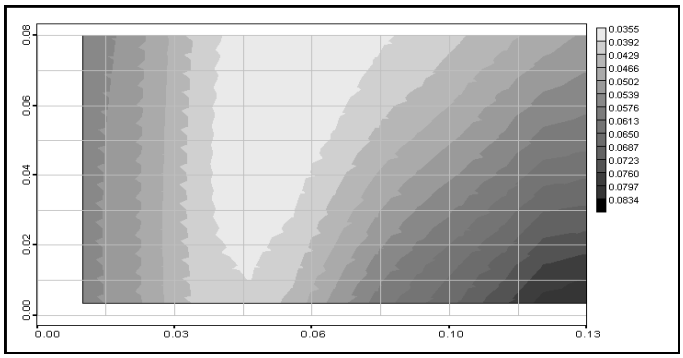

Figure 15. Testing error surface, C= 1000. Axes correspond to X – kernel bandwidth; Y - ε parameter.

Normalized number of Support Vectors is presented in Figure 16. The number of Support Vectors is monotonically decreasing with parameter ε. Let us note, that the largest reasonable order of ε corresponds to the standard deviation of data.

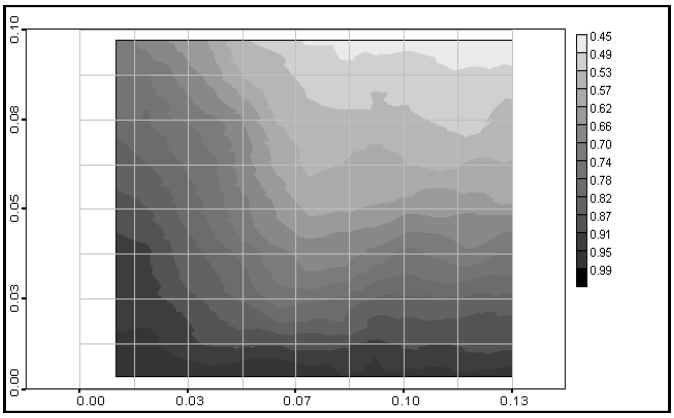

Figure 15. Normalized number of Support Vectors. C= 1000. Axes correspond to X – kernel bandwidth; Y - ε parameter.

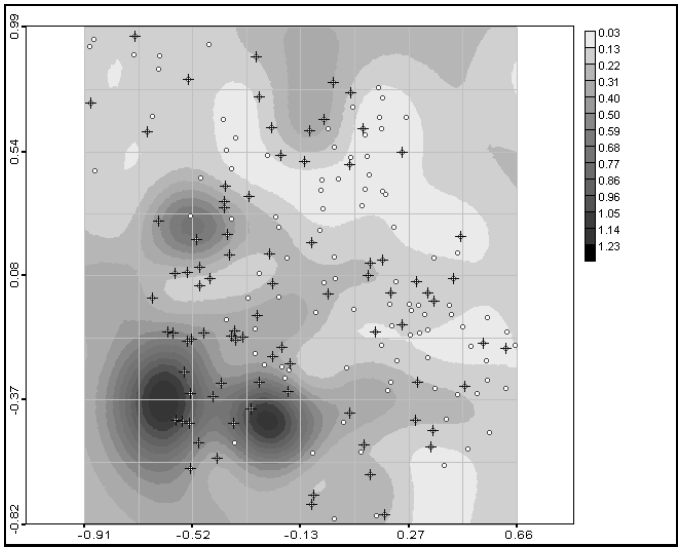

Figure 16. SVR mapping of SR90. Variogram of the training residuals of the model is pure nugget effect corresponding to the nugget of raw data. “o” - training data, “+” – Support Vectors.

The Sr90 concentration varies between 0 and 1.4 Ci/km^2 and the number of training data is 200. Regularization C parameter does not significantly influence error curves when C>1000. At optimal kernel bandwidth training error curves does not change below some value of ε parameter which more or less corresponds to the square root of nugget in original data, and then increases significantly. At fixed kernel bandwidth the number of Support Vectors monotonically decreases (Figure 15). Some discussions on error curves behavior can be found in [7].

VII. 5. CONCLUSIONS

The problem of spatial data analysis and modeling with Support Vector Machines was considered. Both binary and multi-class classification problems were studied.

Multi-class problem was investigated using real data on soil types. Several models generalizing binary class SVC were applied. It was found that simple one-to-rest model gives satisfactory results. There are still some open questions related to the selection of kernel types, local adaptation of SVC and SVR, importance of data preprocessing, etc.

Spatial data mapping with SVR is an efficient nonlinear and robust approach able to extract spatially structured information using raw data. High flexibility of SVR controlled by tuning hyper-parameters can be efficiently used to model non-linear trends as well. Important and rather opened questions deal with multivariate spatial predictions, when the quantity and quality of data for correlated variables is different – the problem of spatial co-estimations; robustness of the solution, direct adaptation and implementation of geostatistical tools into SVC/SVR, understanding of the influence of data clustering (preferential sampling).

ACKNOWLEDGEMENTS

The work was supported in part by European INTAS grants 97-31726, 99-00099 and CARTANN Swiss FNRS grant.

REFERENCES

[1] Vapnik V. Statistical Learning Theory. John Wiley & Sons, 1998.
[2] Cristianini N. and Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000 189 pp.
[3] Burgess C. A tutorial on Support Vector Machines for pattern recognition. Data mining and knowledge discovery, 1998.
[4] Cherkassky V and F. Mulier. Learning from data. Wiley Interscience, N.Y. 1998, 441 p.
[5] Kanevski M., N. Gilardi, M. Maignan, E. Mayoraz. Environmental Spatial Data Classification with Support Vector Machines. IDIAP Research Report. IDIAP-RR-99-07, 24 p., 1999a. (www.idiap.ch)
[6] N Gilardi, M Kanevski, E Mayoraz, M Maignan. Spatial Data Classification with Support Vector Machines. Accepted for Geostat 2000 congress. South Africa, April 2000.
[7] M. Kanevski, S. Canu. Spatial Data Mapping with Support Vector Regression. IDIAP Reasearch Report; RR-00-09.
[8] Atkinson P. M., and Lewis P. Geostatistical classification for remote sensing: an introduction. Computers and Geosiences, vol. 26 pp. 361-371, 2000.
[9] Deutsch C.V. and A.G. Journel. GSLIB. Geostatistical Software Library and User's Guide. Oxford University Press, New York, 1997.
[10] Kanevski M, V. Demyanov, S. Chernov, E. Savelieva, A. Serov, V. Timonin, M. Maignan. Geostat Office for Environmental and Pollution Spatial Data Analysis. Mathematische Geologie, N3, April 1999, pp. 73-83.
[11] M. Kanevski, N. Koptelova, V. Demyanov. RamisW - Software for Modelling Migration of Radionuclides in Soil. Institute of Nuclear Safety (IBRAE). Preprint IBRAE 97-16, Moscow, 1997, 21 p.
[12] Weston J., Watkins C. Multi-class Support Vector Machines. Technical Report CSD-TR-98-04, 9p, 1998.
[13] E. Mayoraz and E. Alpaydin Support Vector Machine for Multiclass Classification, , IDIAP-RR 98-06, 1998 (www.idiap.ch)
[14] M. Kanevski, R. Arutyunyan, L. Bolshov, V. Demyanov, M. Maignan. Artificial neural networks and spatial estimations of Chernobyl fallout. Geoinformatics, vol. 7, pp. 5-11, 1996.