IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

ENVIRONMENTAL SPATIAL DATA CLASSIFICATION WITH SUPPORT VECTOR MACHINES

Mikhail Kanevski[1]
Michel Maignan[2]
Nicolas Gilardi[1 2]
Eddy Mayoraz[1]

IDIAP-RR-99-07

May 1999

Accepted for ACAI 99 workshop

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] IDIAP – Dalle Molle Institute of Perceptual Artificial Intelligence, CP 592, 1920 Martigny, Switzerland
[2] Institute of Mineralogy and Petrology, University of Lausanne, BFSH2, 1015 Lausanne, Switzerland

Environmental Spatial Data Classification with Support Vector Machines[1]

M. Kanevski (1), N. Gilardi (1,2), M. Maignan (2), E. Mayoraz (1)

(1) IDIAP, Case Postale 592, 1920 Martigny, Switzerland. Gilardi@idiap.ch,
(2) Institute of Mineralogy and Petrology, University of Lausanne, BFSH2, 1015 Lausanne, Switzerland

Abstract. The report deals with a first application of Support Vector Machines to the environmental spatial data classification. The simplest problem of classification is considered: using original data develop a model for the classification of the regions to be below or above some predefined level of contamination. Thus, we pose a problem as a pattern recognition task.

The report presents 1) short description of Support Vector Machines (SVM) and 2) application of the SVM for spatial (environmental and pollution) data analysis and modelling. SVM are based on the developments of V. Vapnik's Statistical Learning Theory [1]. The ideas of SVM are very attractive both for research and applications. It was shown that they are efficient and work well in many applications. In the present study SVM were applied to the real case studies with spatial data and compared with geostatistical methods like indicator kriging. SVMs with different kernels were applied (radial basis functions - RBF, polynomial kernels, hyperbolic tangents). The basic results have been obtained with local RBF kernels. It was shown that optimal bandwidth of kernel can be chosen by minimising testing error. Real data on sediments pollution in the Geneva Lake were used.

1. INTRODUCTION

Environmental and pollution data are usually spatially distributed and time dependent. At present there are many monitoring networks collecting data from local to global geographical scales. The quality and quantity of information can be different depending on the tools used, monitoring networks design, etc.

In recent years there has been an explosive growth in development of adaptive methods (dependent on the quality and quantity of information and available knowledge) for learning from data and for working with data. Geostatistics (statistics for spatial data) is one of the well-established approaches for working

[1] *The report is an extended version of the paper accepted for the ACAI'99 workshop.*

with spatially distributed data. There is a wide range of geostatistical methods for the multivariate spatial data mapping and predictions, local probability density function estimations (probabilistic/risk mapping), conditional stochastic simulations/cosimulations (generation of equiprobable realizations of the spatial random function) etc (Deutsch and Journel, 1997; Goovaerts, 1997). Geostatistics, in general, is a model-dependent approach based on the exploratory analysis and modeling of spatial correlation structures. Another data-driven model-(semi)free, contemporary approach is based on statistical learning theory, including supervised and unsupervised artificial neural networks, support vector machines etc. In this case learning method is an algorithm that predicts unknown mapping (classification, regression, density estimation) between inputs and outputs from the available data and a priori knowledge.

Support Vector Machine is used as a universal constructive learning procedure based on the statistical learning theory developed by V. Vapnik (Vapnik, 1995). Recently several research groups have shown excellent performance of SVMs on different problems of classification and regression.

Non-parametric geostatistical model - indicator kriging, is used for the probabilistic mapping of cadmium Geneva Lake sediment contamination. The results are compared with SVM classification.

The present work deals with the development and adaptation of geostatistical method and SVM for the classification of spatial data. The problem is to classify spatially distributed data into regions below and above of some predefined levels of contamination.

All geostatistical part of the work and pre and post-processing of data were carried out with the help of Geostat Office software [Kanevski et al. 1999].

2. SUPPORT VECTOR MACHINES

In the early nineties emerged a new paradigm of learning from data called Support Vector Machines (SVM) (Boser et al., 1992; Cortes and Vapnik, 1995; Vapnik 1995). At first, it was proposed essentially for classification problems of two classes (dichotomies), but now it has been generalised to regression problems (Smola and Scholkopf, 1998) as well as to estimation of probability densities (Weston et al., 1998).

This method has the advantage to place into a same framework some of the most widely used models such as linear and polynomial discriminating surfaces; feedforward neural networks or networks composed of radial basis functions. The strength of the method is that it attempts to minimize simultaneously the empirical risk of error (estimation of the error on the training data) and the structural risk (complexity of the model). By opposition to the Bayesian methods based on a modeling of the probability densities of each class, SVMs are focusing on the marginal data and not on statistics such as means and variances.

In the present work SVMs are used for dichotomies, the next section briefly presents the application of SVMs to such problems (see (Burges, 1998) for a complete tutorial on SVMs). A presentation of the ideas of Statistical Learning Theory can be found in the book [Cherkassky and Mulier, 1998]. SVR Support Vector Regression theory and applications are presented in the tutorial [Smola and Scholkopf 1998].

Principles of SVMs

Consider a dichotomy defined by a set of K couples $\{(\boldsymbol{x}^k, y^k)\}_{k=1,\ldots,K}$ in $R^n \times \{-1,+1\}$, where the data point $\boldsymbol{x}^k$ has to be classified as positive (respectively negative) if $y^k=+1$ (resp. $y^k=-1$). In our application, the input space is R^2 where the two dimensions are the spatial coordinates of the points with measurements. The SVM is implementing a function f from R^n into R with the property that $f(\boldsymbol{x}^k)$ is of the sign y^k hopefully for any $k=1,\ldots,K$ and moreover, for any such k the point $\boldsymbol{x}^k$ lies as far as possible from the decision surface $f=0$.

For simplicity, let first assume that f is a linear function: $f(\boldsymbol{x}) = \boldsymbol{w}\cdot\boldsymbol{x} + b$. If the dichotomy is linearly separable, there exists a vector $\boldsymbol{w}$ and a b such that $(\boldsymbol{w}\cdot\boldsymbol{x}+b)y^k>0$ for all k. The hyperplane $\boldsymbol{w}\cdot\boldsymbol{x}+b=0$ does not change by rescaling $\boldsymbol{w}$ and b, and if $\|\boldsymbol{w}\|=1$, the distance between a point $\boldsymbol{x}$ and the hyperplane

is given by $|\boldsymbol{w}\cdot\boldsymbol{x} + b|$. Thus, if the dichotomy is linearly separable, the pair $(\boldsymbol{w},b)$ chosen by the support vector machine is the optimal solution of the following problem:

$$\max \delta \text{ under the constraints that } (\boldsymbol{w}\cdot\boldsymbol{x}^k + b)y^k \geq \delta \ \forall k \text{ and } \|\boldsymbol{w}\|=1,$$

or equivalently:

$$\min \|\boldsymbol{w}\|^2 \text{ under the constraints that } (\boldsymbol{w}\cdot\boldsymbol{x}^k + b)y^k \geq 1 \ \forall k.$$

This problem is a quadratic program (quadratic objective function and linear constraints) and it can be solved by standard packages (in practice, its dual form is solved instead). The data points $\boldsymbol{x}^k$ for which the inequality constraints are satisfied as equalities at the optimal solution are called the *support vectors* and they alone determine the optimal solution.

This problem has no solution if the dichotomy is not linearly separable. To handle this case, non negative slack variables ξ^k are introduced for each data and the former constraints $(\boldsymbol{w}\cdot\boldsymbol{x}^k + b)y^k \geq 1$ are replaced by $(\boldsymbol{w}\cdot\boldsymbol{x}^k + b)y^k \geq 1- \xi^k$. Of course, as few ξ^k as possible should be non zero, thus a second objective is to minimize $\Sigma_k\ \xi^k$. The new problem has the form

$$\min \|\boldsymbol{w}\|^2 + C\Sigma_k\ \xi^k$$

$$\text{under the constraints that } (\boldsymbol{w}\cdot\boldsymbol{x}^k + b)y^k \geq 1- \xi^k \quad \text{and} \quad \xi^k \geq 0 \ \forall k,$$

where C is a constant weighting the second criterion with respect to the first one. This is once again a quadratic program that can be solved by standard algorithms.

Using the property that the resolution of this quadratic program requires essentially only the computation of scalar products of vectors in R^n, the theory of support vector machines extend to non-linear discriminators f in a very elegant way using the so called *kernel functions*. Some mappings Φ: $R^n \rightarrow R^N$ admit a kernel function K: $R^n \times R^n \rightarrow R$ with the property that $K(\boldsymbol{x}^1,\boldsymbol{x}^2) = \Phi(\boldsymbol{x}^1)\cdot\Phi(\boldsymbol{x}^2)$. Thus, even for mappings Φ so that $N>>n$, the scalar products into R^N of images through Φ can be computed very efficiently using the kernel function. Given such a pair (Φ,K), a discriminant function f linear into R^N but non-linear into R^n can be constructed following the same idea as above by resolving the problem

$$\min K(\boldsymbol{w},\boldsymbol{w}) + C\Sigma_k\ \xi^k$$

$$\text{under the constraints that } (K(\boldsymbol{w},\boldsymbol{x}^k) + b)y^k \geq 1- \xi^k \quad \text{and} \quad \xi^k \geq 0 \ \forall k,$$

which has a dual form simpler to solve. Among all the known kernel functions, the following three are the most widely used:

- *Polynomial kernel*: $K(\boldsymbol{x}^1, \boldsymbol{x}^2) = (\boldsymbol{x}^1\cdot\boldsymbol{x}^2 +1)^p$.
  The result of an SVM with polynomial kernel is a polynomial of degree p.
- *Radial Basis Function (RBF) kernel*: $K(\boldsymbol{x}^1, \boldsymbol{x}^2)=\exp(-\|\boldsymbol{x}^1 -\boldsymbol{x}^2\|^2/2\sigma^2)$.
  The result of an SVM with RBF kernel is an RBF network where σ^2 is the variance of the RB functions (bandwidth).
- *Hyperbolic tangent kernel* $K(\boldsymbol{x}^1, \boldsymbol{x}^2)=\tanh(\kappa\boldsymbol{x}^1\cdot\boldsymbol{x}^2-\delta)$.
  The result of an SVM with such a kernel corresponds to a one hidden layer neural network with hyperbolic tangents as transfer functions of the hidden units and no transfer function for the output units.

3. GEOSTATISTICS. PROBABILISTIC MAPPING WITH INDICATOR KRIGING

There are different geostatistical approaches for the spatial data classification [Goovaerts 1997]. Most of them are based on 1) development of class probability distribution functions and 2) classification with some decision rules. In the present paper indicator kriging is used for the mapping of probability of exceeding selected levels. Probabilistic treatment of the results gives some flexibility in the interpretation of the results and comparison with Support Vector Machines classification.

Indicator kriging is a well-developed geostatistical model for the probabilistic mapping – mapping of local conditional probability distribution function (cpdf) based on available data and knowledge (Deutsch and Journel, 1997; Goovaerts, 1997). Indicator is a function I=Ind(Z;Z*)=1 if $Z \leq Z^*$ and =0 if $Z > Z^*$. Indicator coding allows different types of information to be processed together, regardless of their origins. The objective is to evaluate at any location $\boldsymbol{x}$ the conditional cumulative distribution function (ccdf) value or posterior probability: $F(\boldsymbol{x};Z^* |(n))$=Prob$\{Z(\boldsymbol{x}) \leq Z^* |(n)\}$ where the conditioning information consist of n data measurements and $\boldsymbol{x}=(x_1,x_2)$ in a 2 dimensional case. After an indicator transformation, geostatistical model kriging is applied for the indicators.

Kriging is a Best (minimizing variance of the estimates) Linear Unbiased Estimator (BLUE) of the random function. Each ccdf value can be estimated as a linear combination of neighboring indicator data using kriging algorithm (Goovaets, 1997): $[F(\boldsymbol{x}; Z^*|(n))]_{IK} = \Sigma\lambda_i(\boldsymbol{x};Z^*)I(\boldsymbol{x};Z^*)$, where the weights are given by an ordinary kriging system (Deutsch and Journel, 1997; Goovaets, 1997).

$$\Sigma_{j=1} \{\lambda_j(\boldsymbol{x};Z^*)\gamma(\boldsymbol{x}^i-\boldsymbol{x}^j)\} - \mu(\boldsymbol{x};Z^*) = \gamma(\boldsymbol{x}^i-\boldsymbol{x};Z^*)$$

$$\Sigma_{j=1} \{\lambda_j(\boldsymbol{x};Z^*)\}=1, \quad i=1,\dots n$$

The reconstruction of the entire ccdf can be performed by estimating several thresholds/indicators.

In case of second order stationarity spatial correlation function variogram $\gamma(\boldsymbol{h}) = E\{I(\boldsymbol{x};Z^*) - I(\boldsymbol{x}+\boldsymbol{h};Z^*)\}$ depends only on separation vector ($\boldsymbol{h}$) between points and can be estimated by using transformed data (indicators).

In case of several thresholds co-kriging (co-estimations) of indicators with analysis and modelling of both variograms (autocovariance functions) and cross-variograms (cross-covariance functions) in general should be used. In this case it can be compared with multi-class SVM classification [Weston and Watson 1998]. Indicator co-kriging allows to respect the histogram of data, but the indicator variograms for thresholds not close to median are difficult to infer. Moreover, the smoothing effect due to minimising of the variance is large.

4. CASE STUDY

Description of Data

Data were provided by the CIPEL (International Commission for the Protection of Water of the Geneva Lake, Lausanne). They are of two kinds. The first data set is a chemical analysis of sediments during the years 1978, 1983 and 1988. These data had not been previously valorized spatially with geostatistical methods. The second data set is a chemical analysis of water of the Geneva Lake at various depth, at various locations, and from 1957 to 1994 for the longest period.

For this case study, we focus on sediment data of the year 1988. Those data contain a list of chemical elements (heavy metals, and organic molecules) detected during the analysis, and also some information about the various kinds of sediment analysed (diameter of the grain).

The Cadmium concentration was used, as reported on the sediment data set of 1988. Thus, univariate (only one variable) spatial classification and mapping of the Cd concentration (measured in μg/g) is of the main interest. The basic batch statistical parameters of the data are following:

min = 8.e-02; **Q 1/4** = 5.1e-01; **median** = 7.3e-01; **Q 3/4** = 1.030e+00; **max** = 3.290e+00; **mean value** = 8.16e-01; **variance** = 2.2e-01; **sigma** = 4.71e-01; **skewness** = 1.950e+00; **kurtosis** = 6.88e+00.

Total number of measurements equals 200.

Cd histogram along with histogram cloud are presented in Figure 1. The histogram cloud represents the density of measurement points versus measurement values. Histogram is a summary statistics of the histogram cloud after dividing the measurement values into several bins.

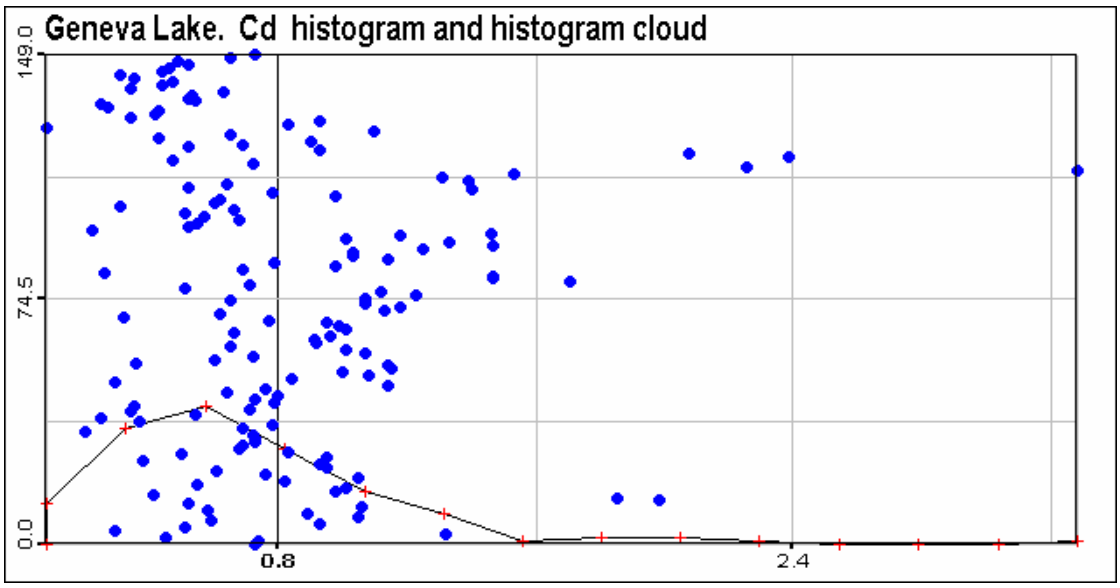

Figure 1. Histogram and histogram cloud of original Cd data.

Qualitative and quantitative analysis and description of the monitoring networks and their clustering is an important phase of spatial data analysis. There are different deterministic (describing spatial resolutions), statistical (Morishita diagrams, etc.), and fractal (describing dimensional resolution) measures for the monitoring network analysis. In general, in order to work with representative data sets, different declustering procedures (random declustering, cell declustering, Voronoi polygons, kriging weights) applied to the original raw data should be used.

The second important step in the geostatistical spatial data analysis deals with comprehensive exploratory description and modeling of spatial continuity using spatial correlation functions: variograms, covariance functions, madograms, rodograms, etc. In the present work it was performed on indicator transformed data (see below). The variography is of great importance both for the original data analysis and the results despite of the methods used.

The figure 2 shows the cadmium concentration at each point of measurement, the higher level are reached in the north coast and in the middle of the “Small Lake”, in the southwest. But there is also a large area of medium concentration in the center of the “Great Lake”, and some hot spots near the coasts.

Figure 2 was prepared with linear interpolation algorithm (Delauney triangulation) and can be considered as a visualisation of the raw data.

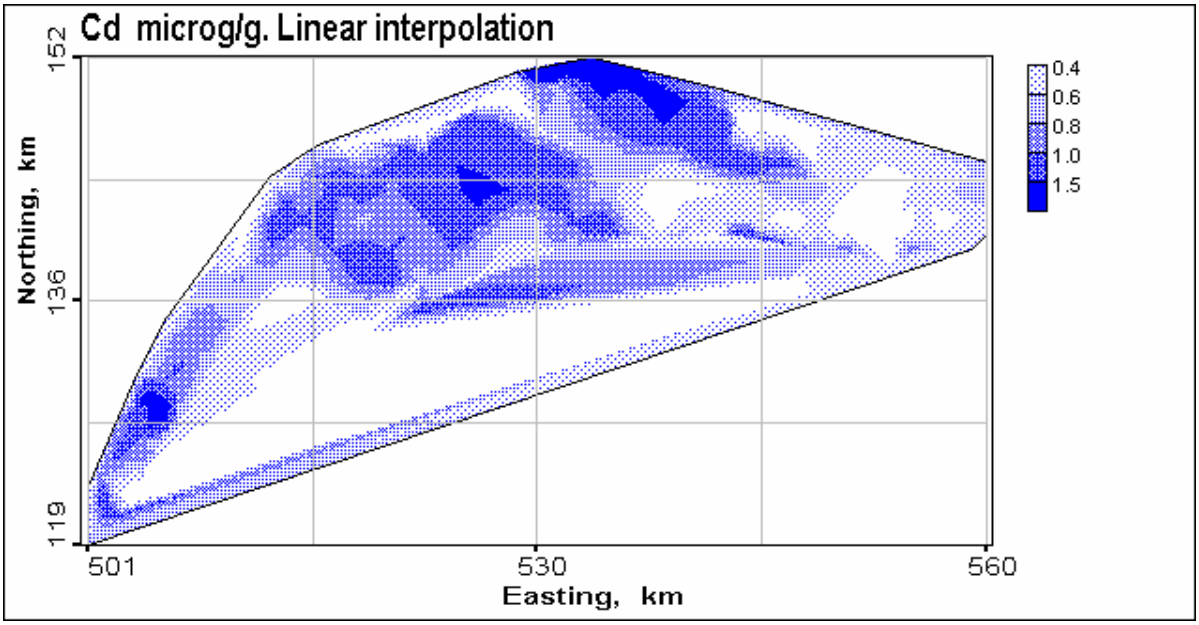

Figure 2. Linear interpolation of original data (visualisation of data).

Data Pre-processing

Original data were split into training (150 measurement points) and testing (50 measurement points) data sets. Two different techniques have been applied: 1) random splitting, when 50 data points are randomly selected from the original file: 2) spatial declustering with random selection. In the latter case region is covered by regular grids and from each cell one data is randomly selected. Selected data set is more homogeneous, the rest data set is more clustered. Number of extracted data depends on the regular grid. This procedure usually gives more representative data sets. One of the example used for the study is presented in Figure 3. Netman module of the Geostat Office was used for the data analysis and spitting.

In general data splitting should be performed many times which helps in understanding of algorithms stability and effect of clustering of the monitoring network .

In the present paper we are trying to solve one of the important problems of decision-oriented mapping with pattern recognition technique. The posing of the problem is following: using available data draw a decision boundary for the indicator function (indicator is a binary function equals 0 if data are above threshold or 1 otherwise). Thus, we are interesting in drawing the boundaries separating data into the regions above and below some predefined level. Usually this level is an intervention or countermeasure one for pollution.

For the present research study experiments were performed mainly with two thresholds: C1=0.8 μg/g; and C2=1.0 μg/g.

The choice of a 0.8 threshold (very close to the mean value of the data) gives a quite large area of matching points see histogram cloud. In comparison, the choice of the 1.0 threshold concentrates the information much more on the location of higher concentration areas. The future classification will then result in two patterns, different so to conclude on classifiers' efficiency. Let us note that both cases are non-linear classification problems.

The next step consists in preparing data for training and testing the classifier, and also in creating a geographical grid in order to perform spatial classifications.

This work has been done with a Netman module of Geostat Office [Kanevski e al. 1999]. Netman permits to split the data set of 200 values in one training data set of 150 values, and one validation data set of 50 values, without destroying the spatial structure of the monitoring network, as shown in figures 4 and

5 for the 0.8 threshold. In general, Netman is a powerful collection of tools for the quantitative and qualitative monitoring networks description, analysis and modelling.

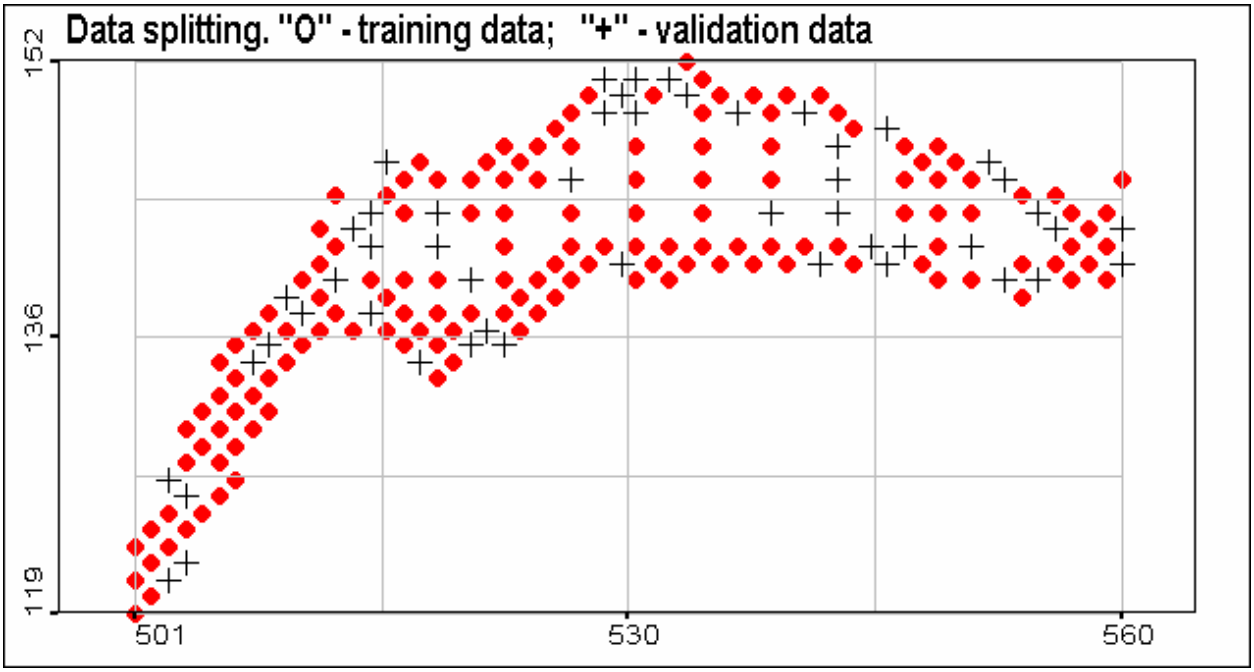

Figure 3. Posplot of data splitting.

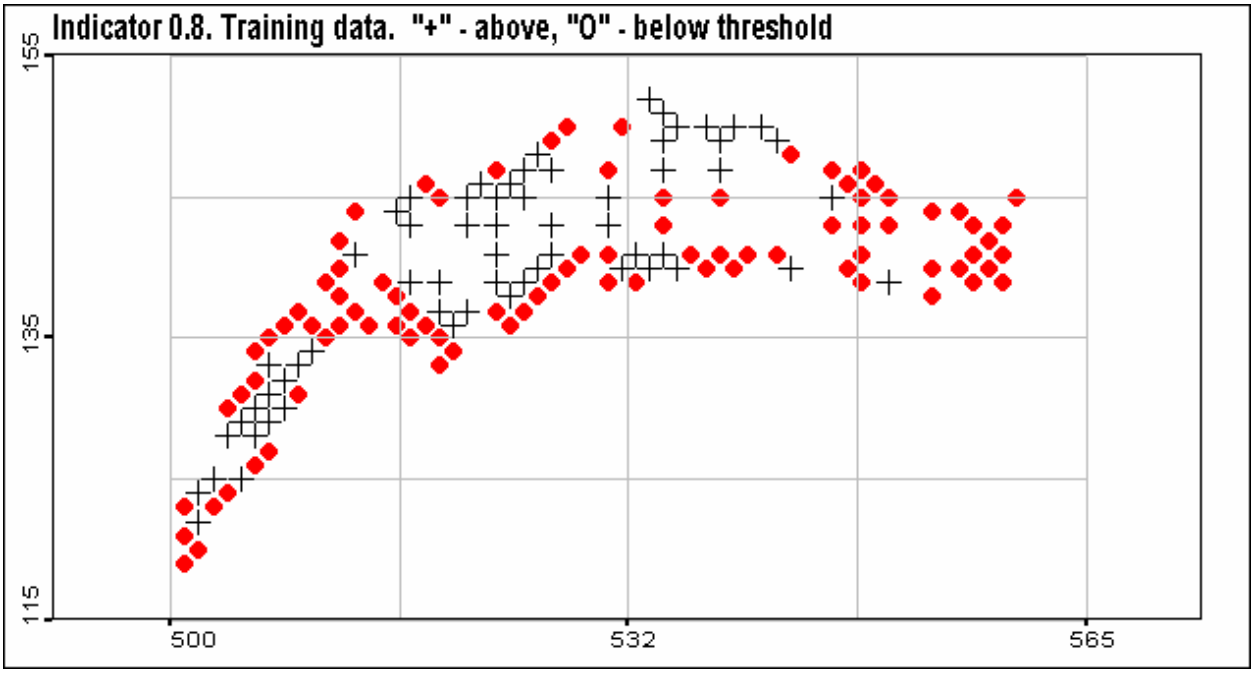

Figure 4. Indicator 0.8 postplot. Training data set.

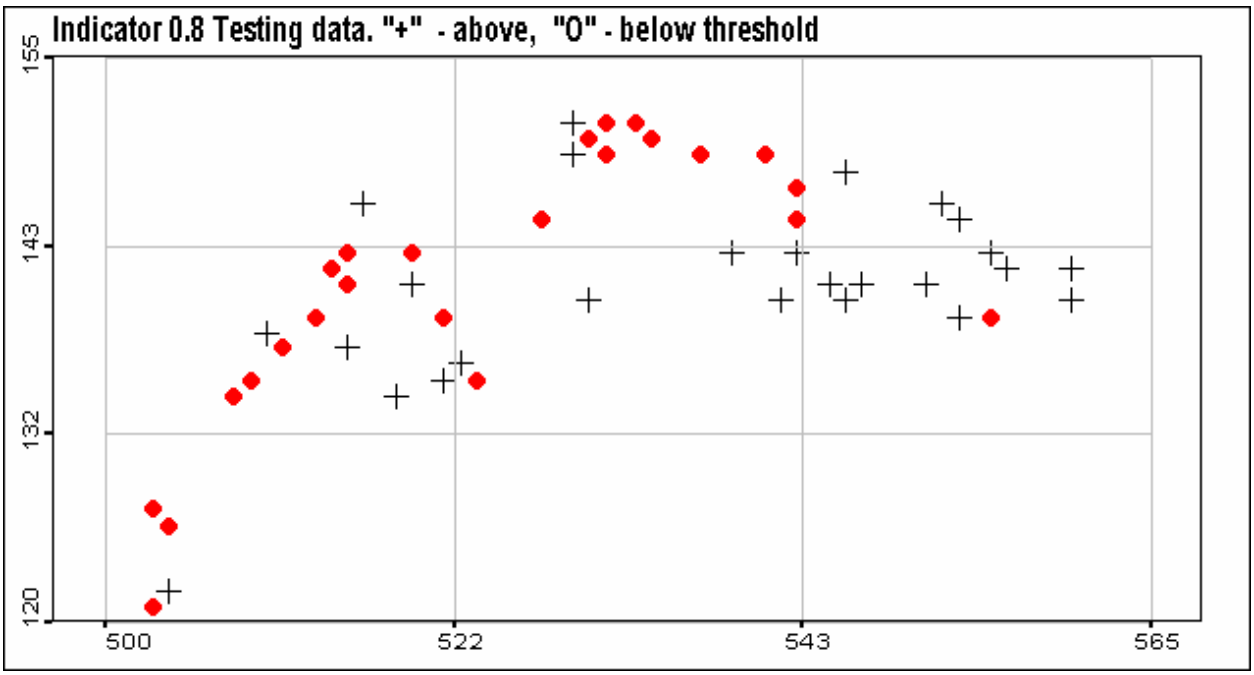

Figure 5. Indicator 0.8 postplot. Testing data.

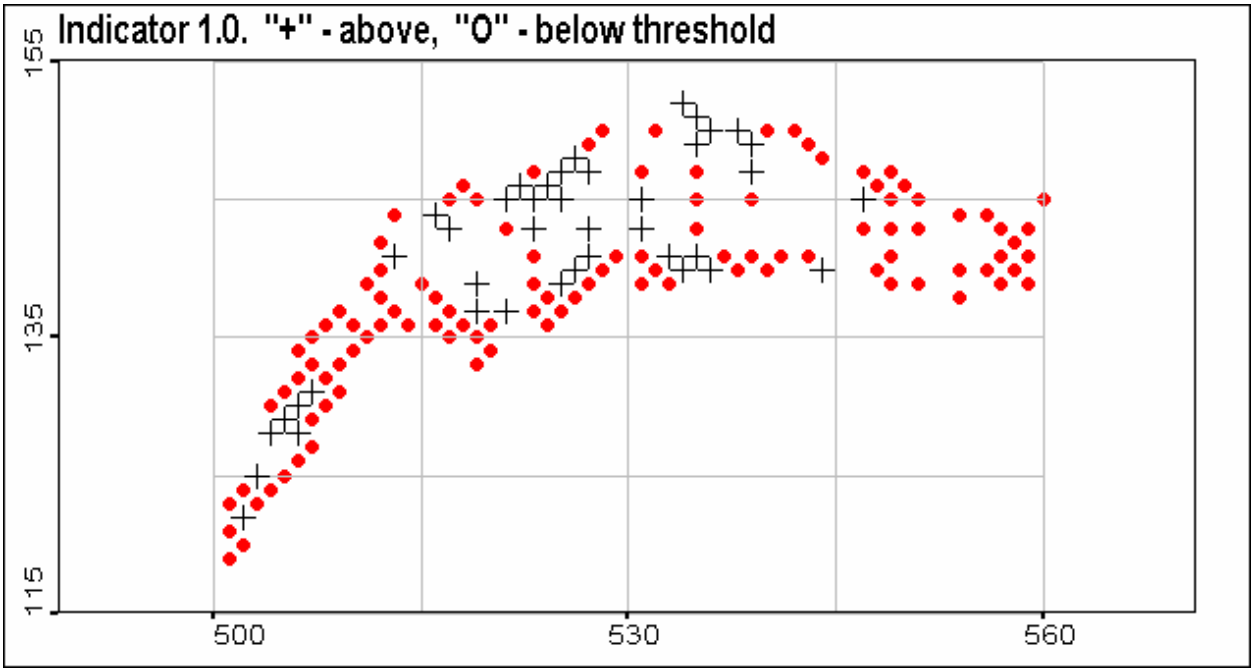

Figure 6. Indicator 1.0 postplot. Training data set.

For the geographical grid, Netman generated a dense network of 720 points (figure 7) based on the initial positions of the data points. There are some points outside of the lake border, but it does not pose any problems, except for visualisation. In general, the final results should be prepared as the decision-

oriented maps and presented with the help of Geographical Information Systems. Geostat Office is able to export to GIS all basic results as the main topological objects: points, lines and polygons.

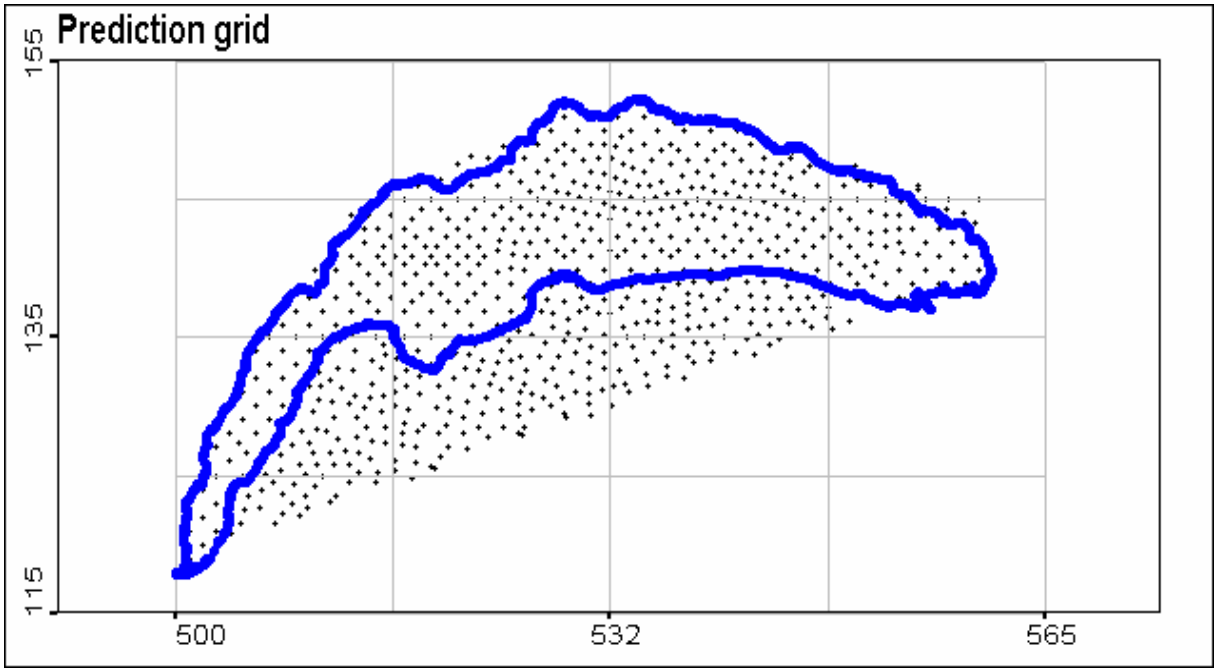

Figure 7. Prediction grid. The coastline of the Leman lake is presented as well.

SVMLab program used in the present study at first linearly transforms spatial co-ordinates (input space) into Zscores by using the following relationsip:

$$\mathbf{X}_{score} = \frac{\mathbf{x} - \mathbf{m}_x}{\sigma_x}$$

where vector $\mathbf{x}=(x,y)$, and σ - is a standard deviation (square root of variance). The parameters for the Lake are following: m_x =526.9; m_y = 139.29; σ_x =16.78; σ_y =7.1. Let us remark, that occasionally relationship between σ_x and σ_y is close to the relationship between ranges of spatial correlation in X and Y directions (see variography below).

Postplot of the training data using transformed co-ordinates is presented in Figure 8. This is an affine transformation of the input space. In fact, working with the unique sigma value of the kernel (bandwidth of the SVM kernel) means anisotropic modelling in the input space. The bandwidth of the SVM kernel for the results below is presented in the transformed co-ordinates. These remains to ascertain that this coordinate reduction does not mesh the anisotropies "naturally" present in the data. This applies specifically to the Leman data where the flows from Valais down to Geneva are a major fact.

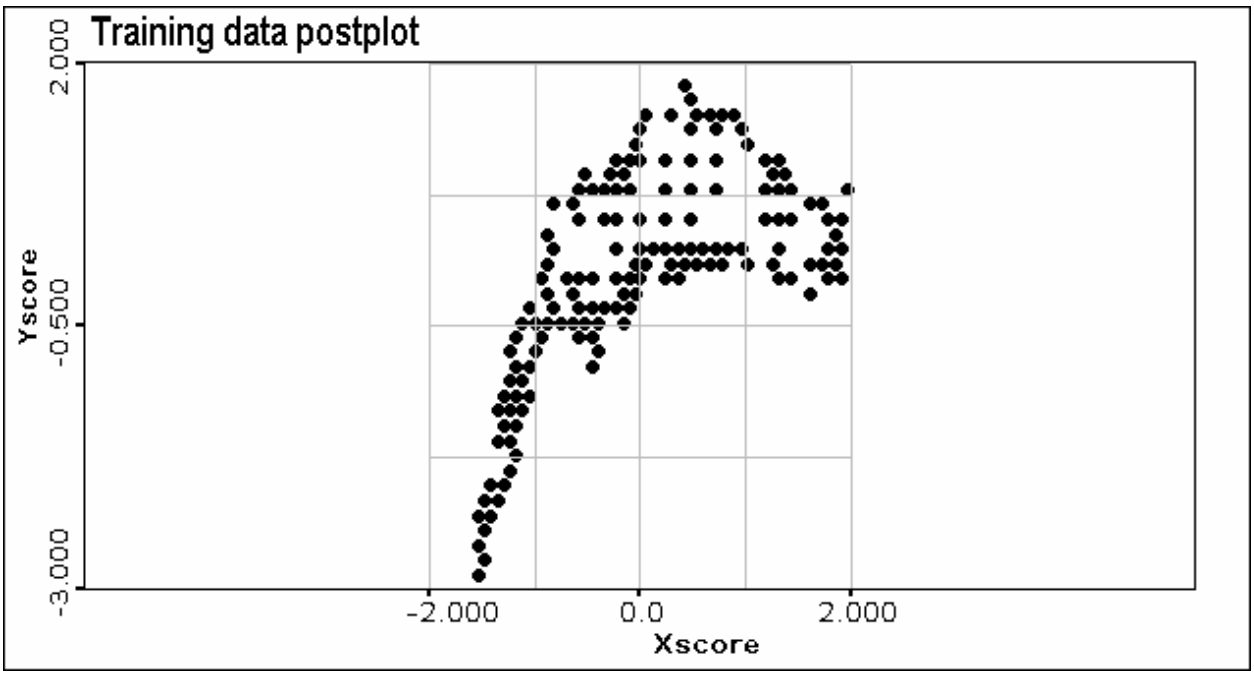

Figure 8. Postplot of training data set using transformed data.

Classification with SVM

Most of our results on SVM were obtained with a classification program made at IDIAP (Dalle Molle Institute of Perceptive Artificial Intelligence, Switzerland). This program is an Octave application using the LOQO free optimiser. At this time, we are testing the efficiency of the RHUL (Royal Holloway University of London, United Kingdom) software on SVM, in order to compare the results.

Quality of the results criteria

The construction of a classifier is decided as follows. First of all, a kernel type is chosen (three basic kernel have been applied). Then, the specific parameters of the kernel are selected. After this, the support vectors' coefficients with the optimiser are calculated according to the training data. Finally, the efficiency of those coefficients and kernel's parameters are estimated using testing data set.

Two error measures were used:

$$\frac{Number\_of\_misclassified\_data}{Total\_number\_of\_data}$$

One of them is specific to the training data and the other one to the testing data. The objective of the training is to minimise both.

Choice of the kernel

The choice of the kernel is a crucial issue in the SVM method. With the polynomial kernel, high quality results with a degree 9 were obtained. But when using the grid in order to *see* the quality of the results, error was growing in a pathetic way at the border of the classes (i.e. near the coasts of the lake). This is also a well-known problem with polynomial regression for one-dimensional function.

The case of the hyperbolic tangent kernel is very different. In fact, this kernel is using two parameters that are difficult to optimise with a given data.

With the RBF kernel, very good results (even better than with polynomial kernel), without strange features at the border of the classes were obtained. In addition, this kernel is very simple to use, as it needs only one parameter (variance of the kernel - bandwidth). The interpretation of the kernel is rather simple as well.

Error curves

In order to make a legitimated choice of the optimal classifiers, and also to understand the variation of the training error and the testing error, error curves were calculated (figures 8 & 9).

Those curves represent the variation of training error and testing error versus the kernel variance parameter.

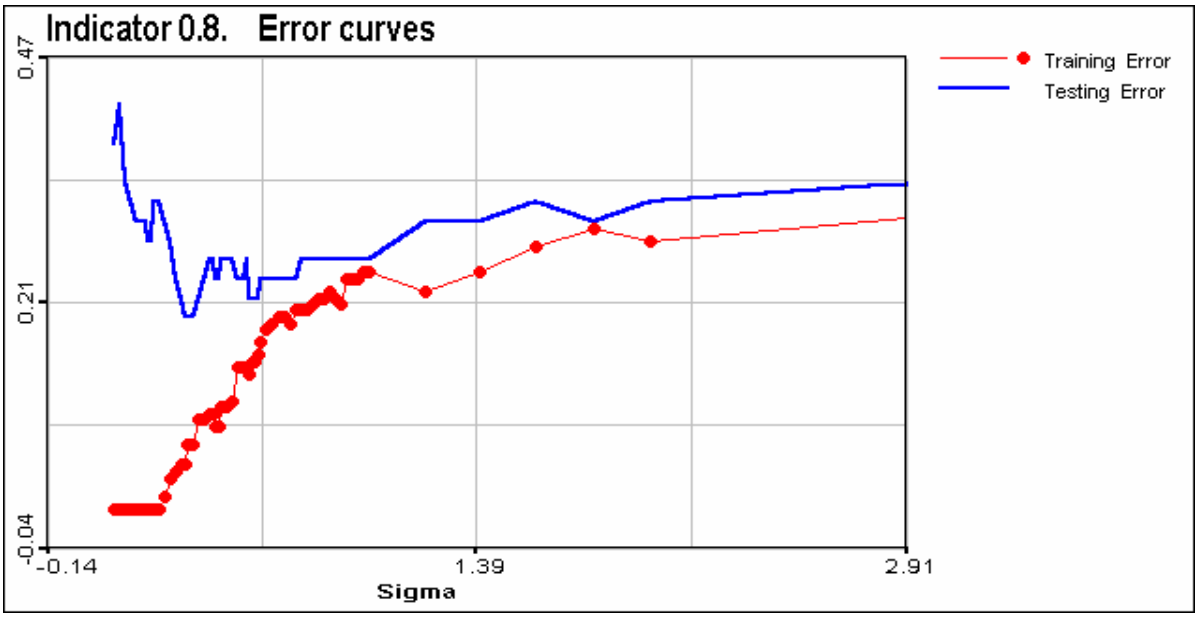

Figure 9. Indicator 0.8. SVM error curves for the training and testing data sets.

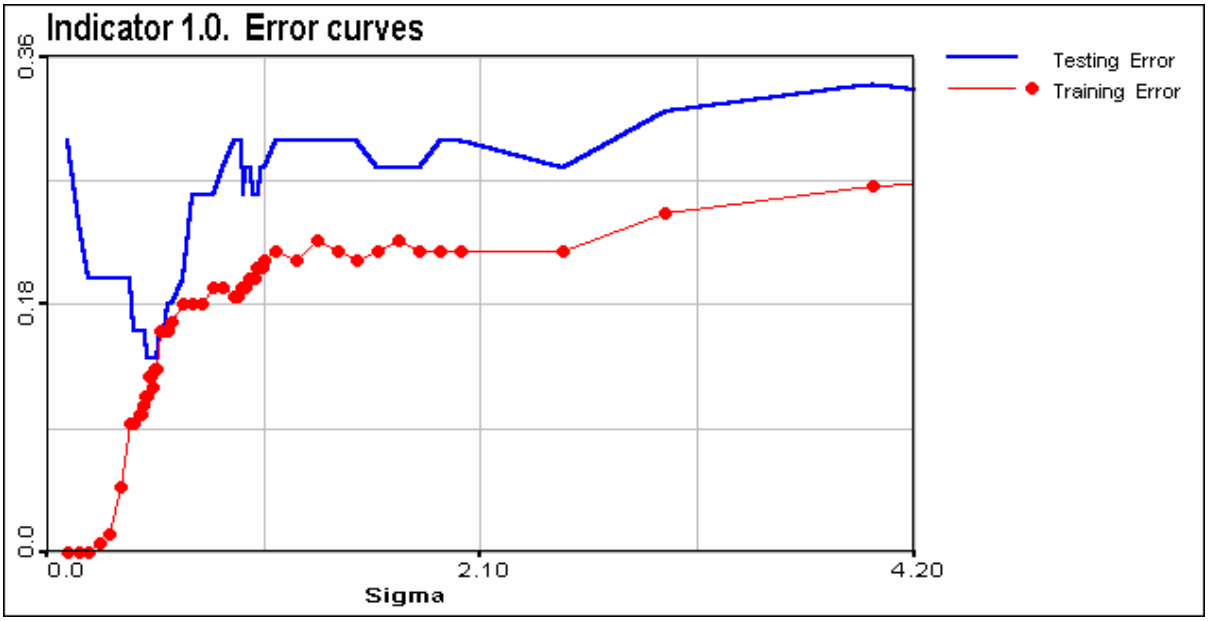

Figure 10. Indicator 1.0. SVM error curves for training and testing data sets.

The error curves for the two thresholds are quite similar and can be divided into three parts. First, testing error is at a very high level while training error is almost zero. This is the over-fitting part of the curves. Second, testing error is falling as fast as training error is rising. But after this decrease, testing error starts to follow the raise of the training error. Classifier's parameter is optimal when testing error is reaching its minimum. At the end, the two curves are reaching a plateau at a high error value. In this part, nothing can be decided because the classification does not work correctly. This is a region of oversmoothing.

After selecting the optimal bandwidth of the kernel, it can be used for the spatial classification (predictions on the dense grid).

Results of the SVM classification

By changing the kernel variance parameter (bandwidth), different results can be demonstrated: from overfitting at small bandwidth, to oversmoothing with rather high values of the kernel bandwidth. The application of the trained SVM for the spatial classification at the grid points is presented in Figures 10-12 for the 0.8 indicator and in Figures 13-15 for the 1.0 indicator. SVM with the optimal kernel variance (minimum testing error) as well as with sub-optimal kernel variances giving rise to over-fitting and over-smoothing effects have been applied for the comparison.

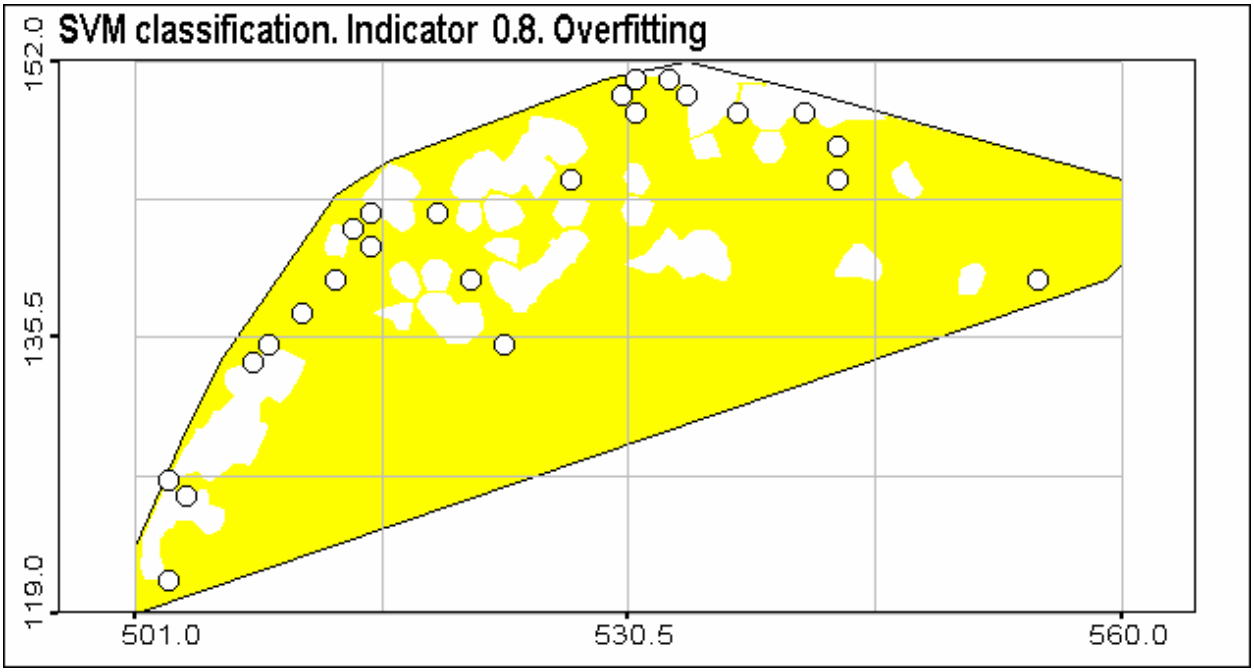

Figure 11. SVM classification. RBF kernel. Over-fitting. White zones correspond to the above threshold value classification. Dots indicate above threshold values for the validation data. (σ=0.03)

The shapes obtained for these 3 classes are currently under investigations by methods of mathematical morphology: relationships of perimeters to surface, areas, etc. Comprehensive structural analysis of the results with geostatistical tools (variography) is under study and will be published with discussions separately.

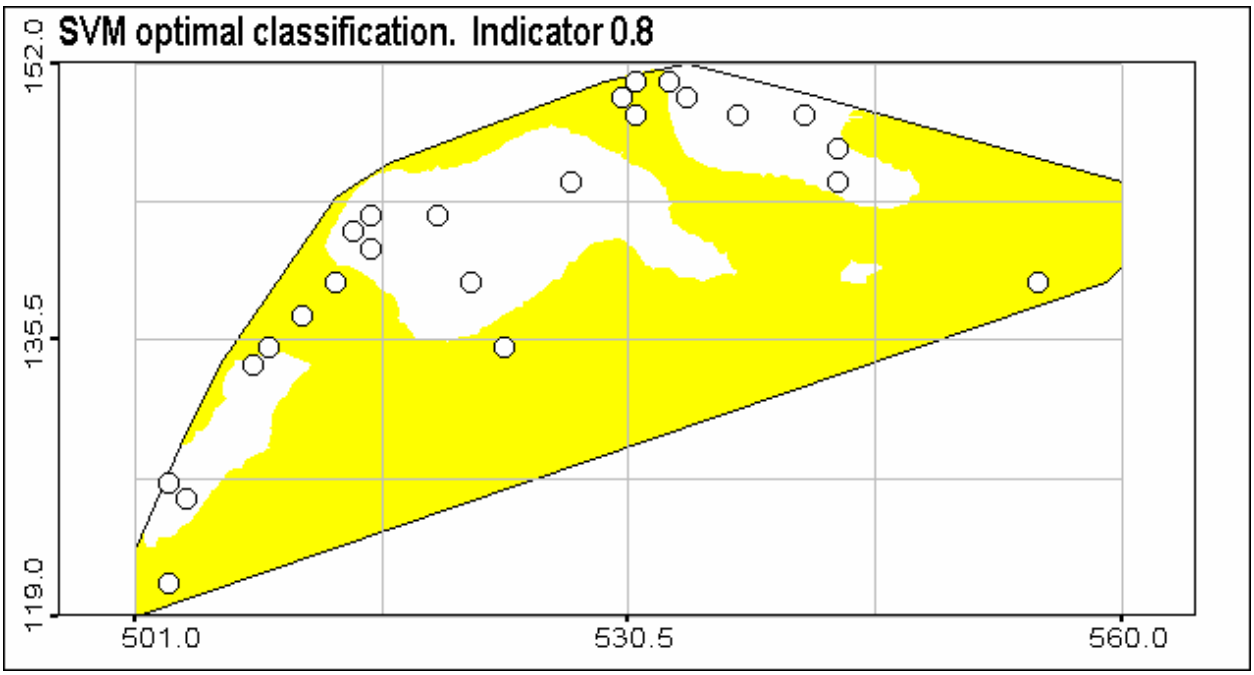

Figure 12. SVM optimal classification. RBF kernel. White zones correspond to the above threshold value classification. Dots indicate above threshold values for the validation data. (σ=0.35)

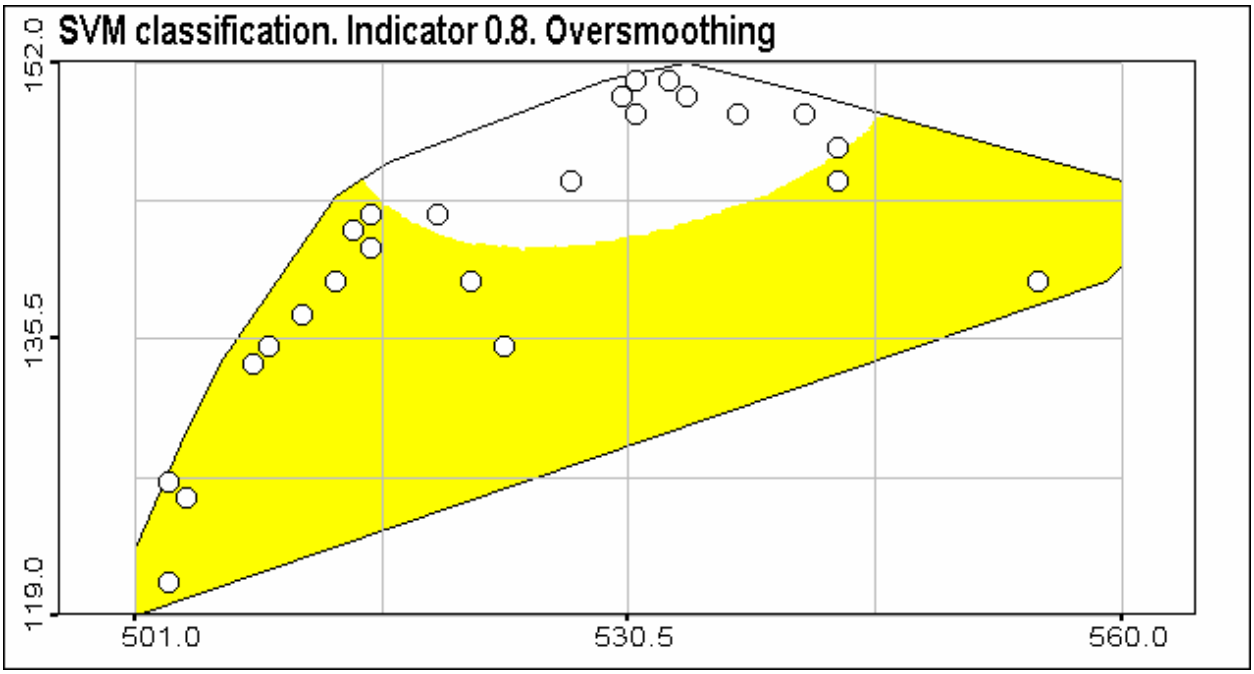

Figure 13. SVM classification. RBF kernel. Oversmoothing. White zonez correspond to the above threshold value classification. Dots indicate above threshold values for the validation data. (σ=3.0)

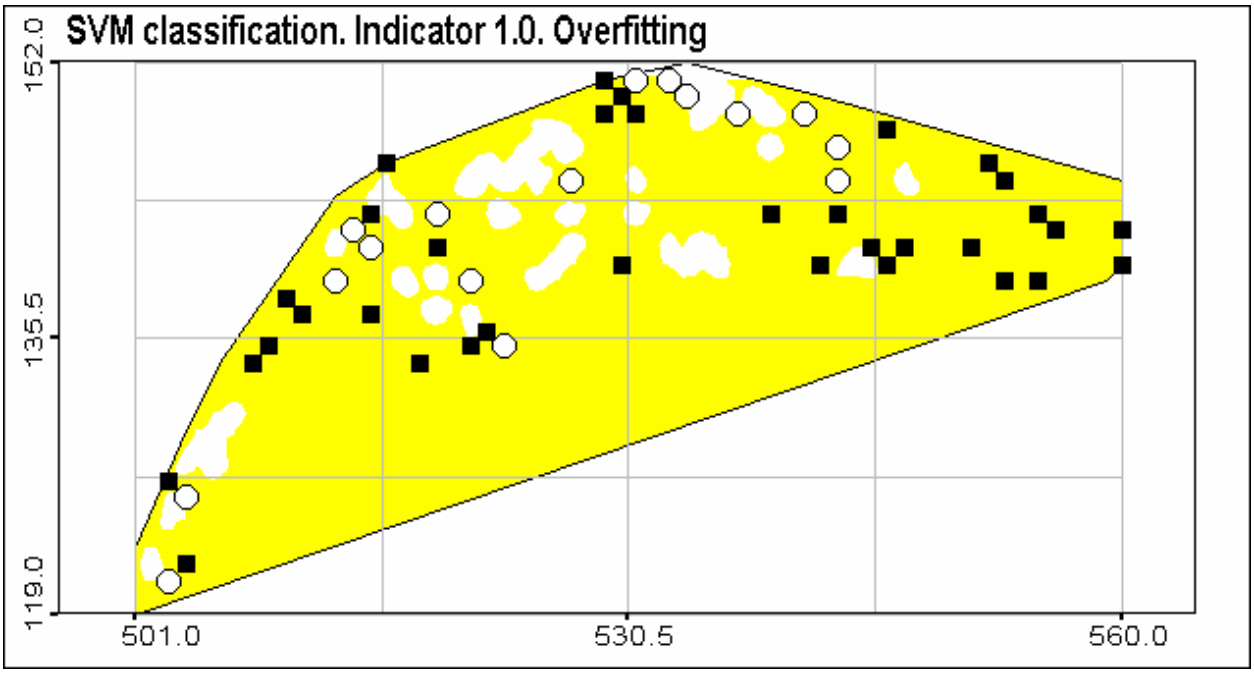

Figure 14. SVM classification. RBF kernel. Overfitting (σ=0.1). White zone corresponds to the above threshold values. Dots indicate above threshold values for the validation data; filled squares – below threshold values.

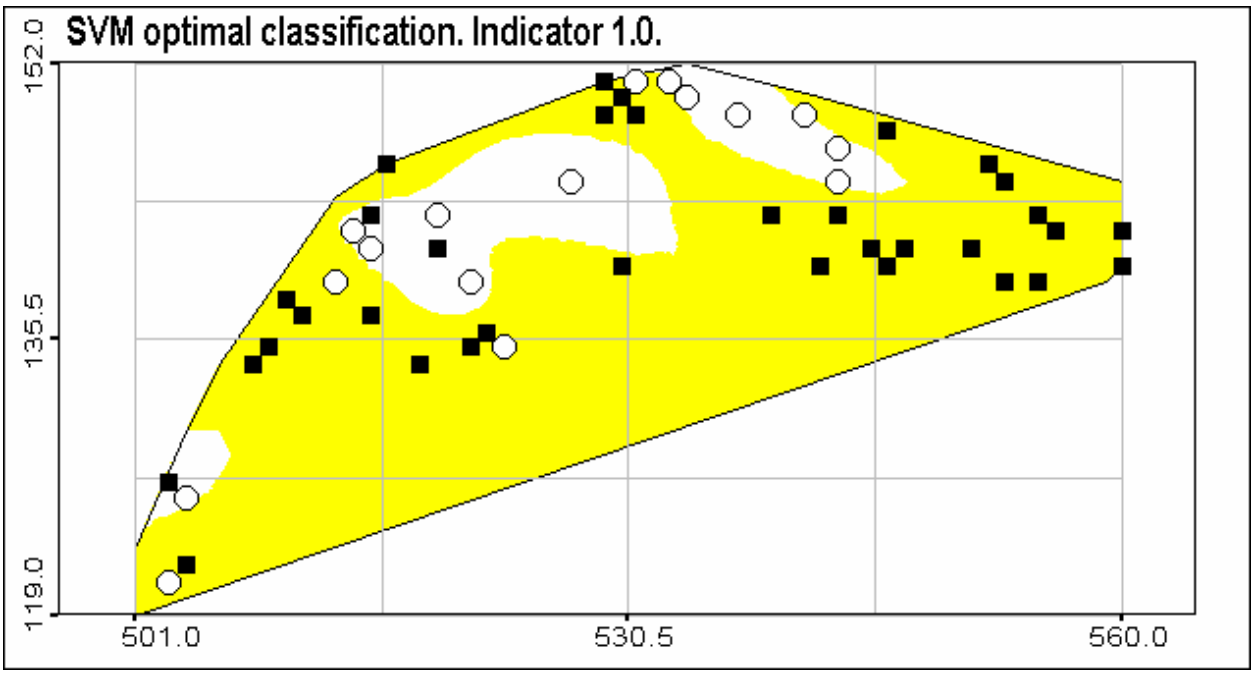

Figure 15. SVM optimal classification. RBF kernel. White zone corresponds to the above threshold values. Dots indicate above threshold values for the validation data; filled squares – below threshold values. (σ=0.5)

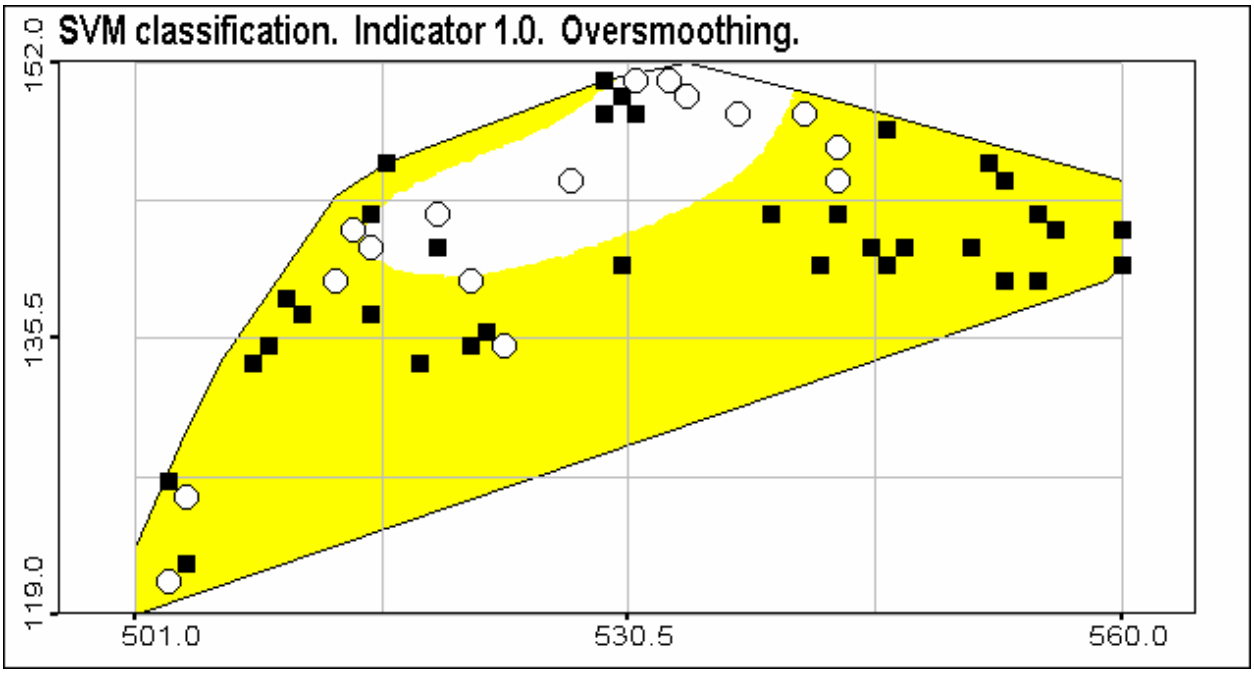

Figure 16. SVM classification. RBF kernel. Over-smoothing. White zone corresponds to the above threshold values. Dots indicate above threshold values for the validation data; filled squares – below threshold values. (σ=10.0)

Indicator Kriging

Experimental variography

Geostatistics is based on a description and modeling of spatial continuity using original data (hard data) and knowledge about phenomena under study (soft data). There are several measures describing spatial continuity (spatial correlation structures). The most widely used measures in geostatistics are covariance funtion and variogram/semivariogram. The basic theoretical formulas as well as empirical estimates of the corresponding functions are presented below.

1. Covariance function. Theoretical formula

$$C(\mathbf{x},\mathbf{h}) = E\{(Z(\mathbf{x})-m(\mathbf{x}))(Z(\mathbf{x}+\mathbf{h})-m(\mathbf{x}+\mathbf{h}))\}$$

Covariance function. Empirical estimate (under the hypotheses of second-order stationarity $C()=C(\mathbf{h})$.

$$C(h) = \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})} Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h}) - m_{-h}m_{+h}$$

$$m_{-\mathbf{h}} = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})} Z(\mathbf{x}_i)$$

$$m_{+\mathbf{h}} = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})} Z(\mathbf{x}+\mathbf{h})$$

2. In multivariate case cross-covariance function is considered as well. Theoretical formula

$$C_{ij}(\mathbf{x},\mathbf{h}) = E\{(Z_i(\mathbf{x})-m_i(\mathbf{x}))(Z_j(\mathbf{x}+\mathbf{h})-m_j(\mathbf{x}+\mathbf{h}))\}$$

3. Semivariogram/variogram (the basic tool of the spatial structural analysis - variography). Theoretical formula (under the intrinsic hypotheses)

$$\gamma(\mathbf{x},\mathbf{h}) = \tfrac{1}{2}Var\{Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\} = E\left\{\left(Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\right)^2\right\} = \gamma(\mathbf{h})$$

Empirical estimate of the semivariogram

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left(Z_i(\mathbf{x}) - Z_i(\mathbf{x}+\mathbf{h})\right)^2$$

In the two dimensional space vector $\mathbf{x}=(x,y)$ and vector $\mathbf{h}$ defines a distance and direction between two points in space.

General understanding of spatial continuity described by variograms can be obtained from the variogram analysis of raw data. The variogram rose (presentation of the variograms, calculated in several directions) for the raw data of Cd sediments concentration is presented in Figure 16. Geometrical anisotropy (different ranges of correlation in different directions) is evident. These anisotropies respect the main flow of the river Rhone in the lake.

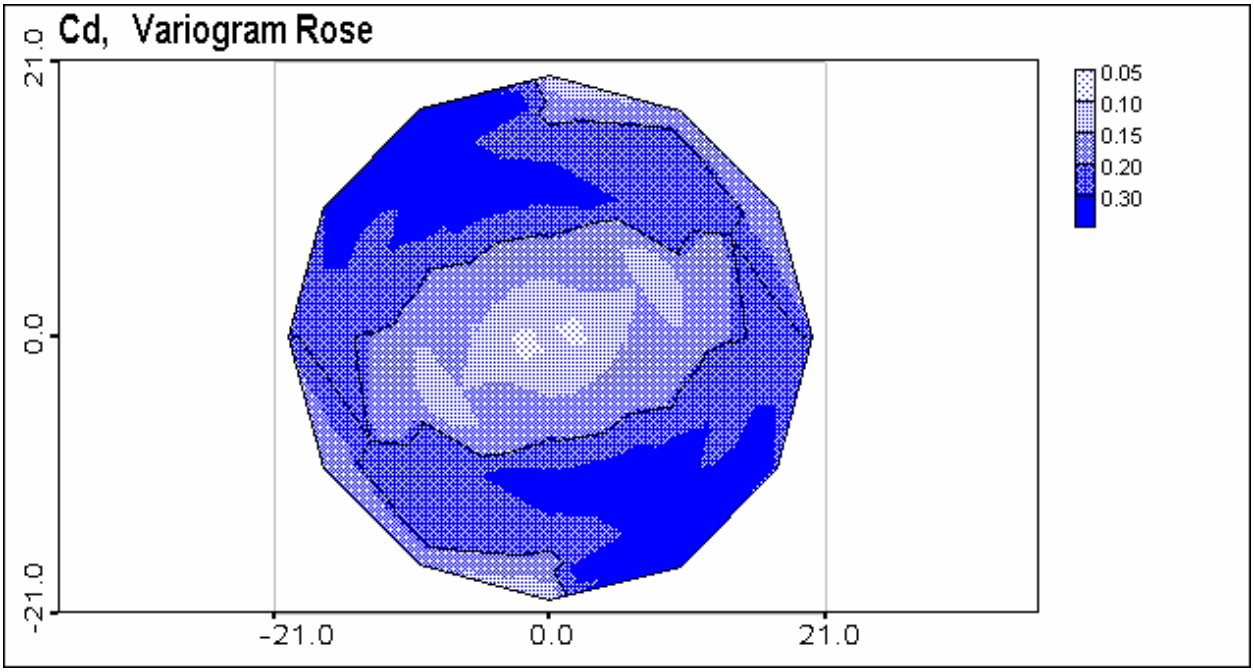

Figure. 17. Experimental variogram rose of Cd sediment concentration.

Anisotropic structure of the spatial correlation of raw data can be observed. This is so-called geometric anisotropy – different ranges of the correlation in different directions. The major axis of correlation is along the 10 degrees calculating from WE direction. Minor axis is in a transverse direction 100 degrees from WE direction.

1 dimensional figures for the anositropic variograms for the raw Cd data are presented in Figure 17. The ratio between major and minor axes is near $R(\|)/R(\perp)\approx 15/7$.

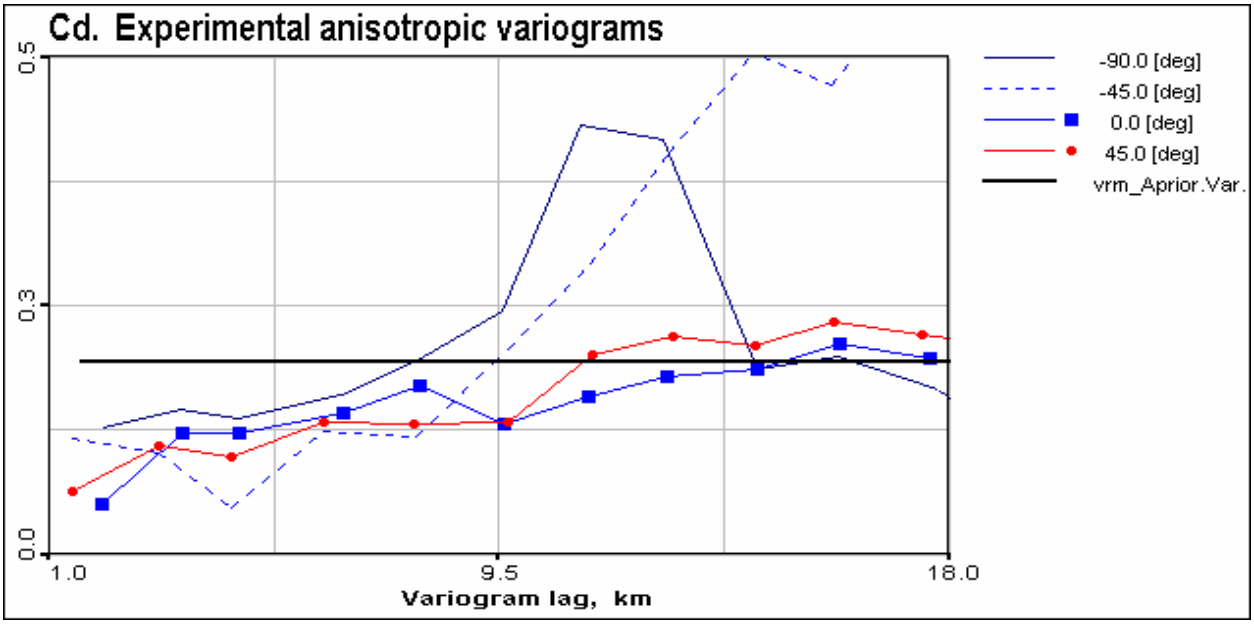

Figure 18. Cd raw data. Experimental anisotropic variograms.

For the indicator kriging is important to calculate and to model indicator variogram. After transformation of the raw data to the indicators comprehensive structural analysis (variography) was carried out.

The experimental variogram rose is presented in the Figure 18. What is different in comparison with the raw data variogram rose? First, angle of the geometric anisotropy was changed. Now, the major axis of the anisotropy is along the direction –15 degrees from the WE direction. Second, the correlation ranges have been decreased. But the ratio between major and minor ranges is almost the same $R(\|)/R(\perp)\approx 11/5$.

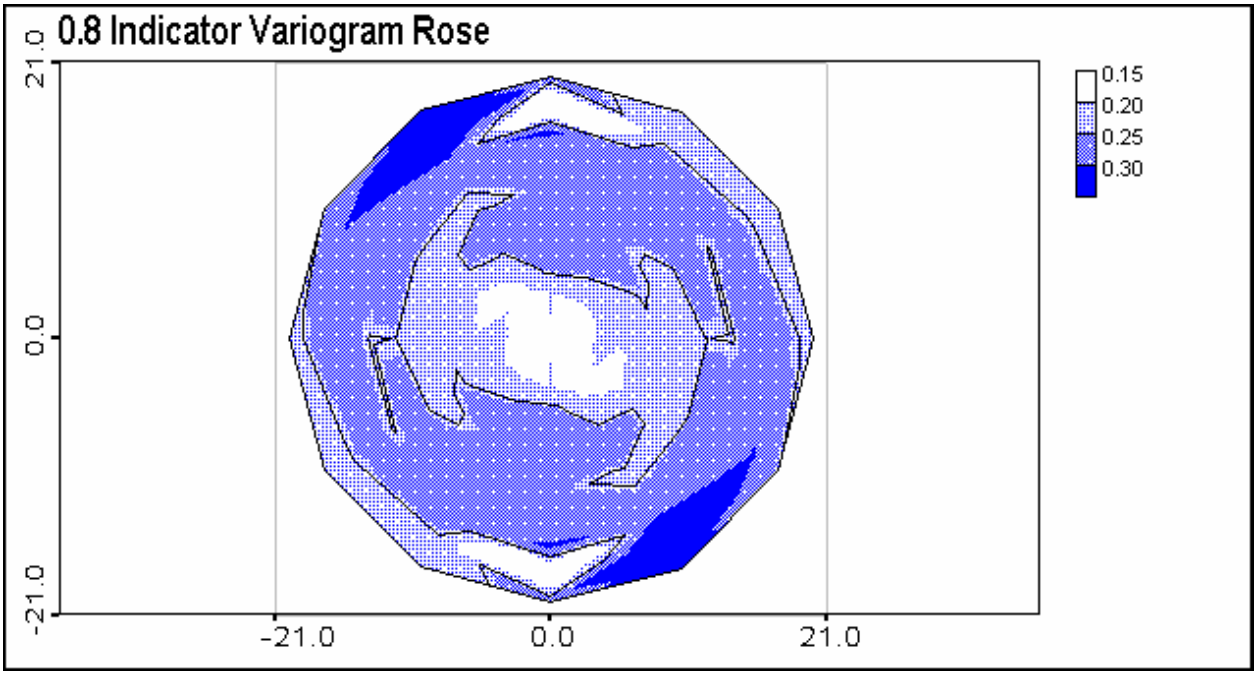

Figure 18. 0.8 indicator Variogram Rose.

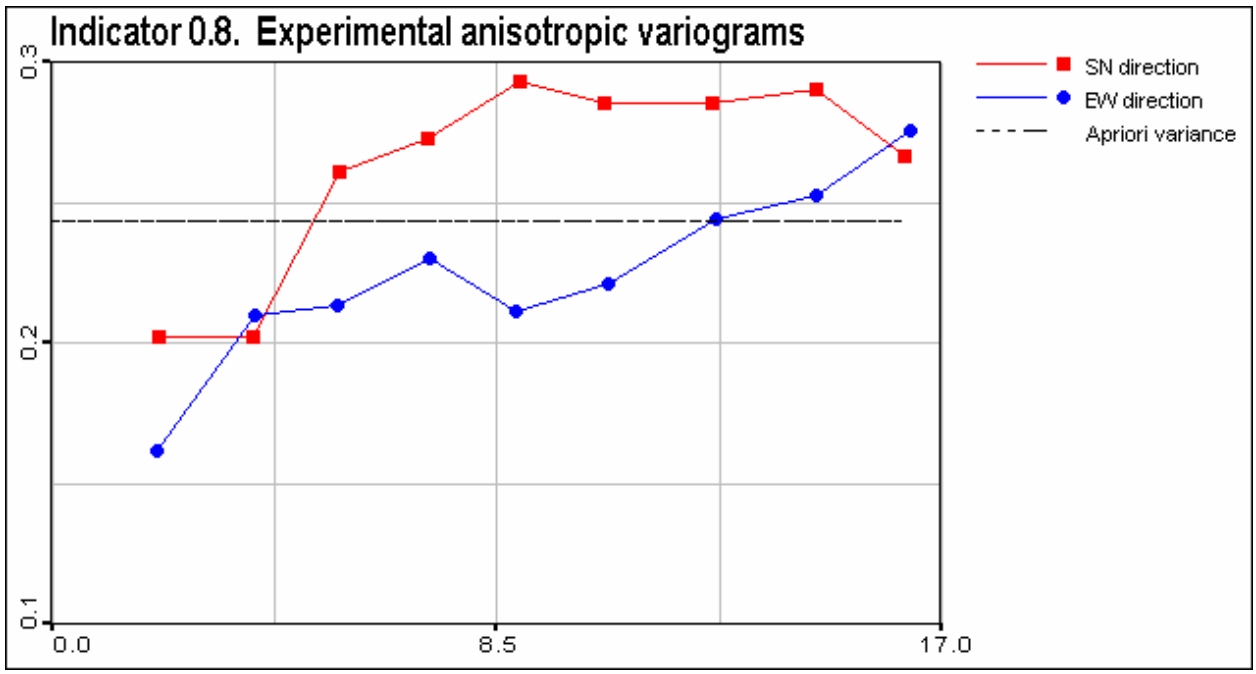

Figure 20. Anisotropic indicator experimental variograms.

One dimensional variograms calculated in two directions (WE and SN) are presented in Figure 19. Rather high nugget effect can be mentioned as well. Nugget effect describes small scale variability. The relationship between nugget and sill (plato) reflects the relations between stochastic and structured components of the data.

Variogram rose for the transformed coordinates was calculated as well (see Figure 20.). It is evident that variogram rose in this case is much more isotropic. It means, that occasionally by using Zscore transformation the object under study was simplified. Actually, it should help SVM to develop model, which would be simpler (less support vectors) than in the case of anisotropic object. The general question is how data pre-processing are important for the support vector machines? Remember, that taking into account inherent properties of the data (symmetries, etc.) should help to improve the results of the SVM classification. This fact was mentioned in the book of V. Vapnik [Vapnik 1995]. Of course, this is possible only when do we have only one (possibly anisotropic) scale of variability. In case of multi-scale variability there is still some possibility to simplify patterns with the help of general affine transformation. The problem of the relationships between spatial correlation structures and SVM classification of environmental data is under extensive study at present and will be published.

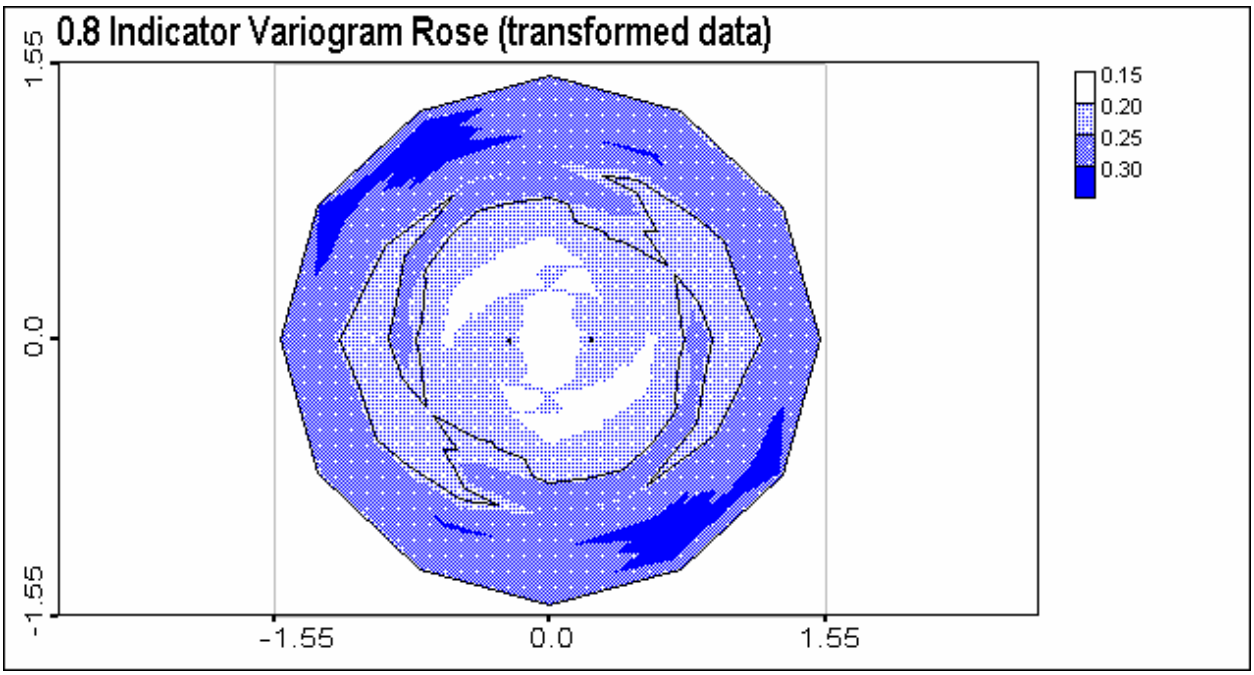

Figure 21 . 0.8 Indicator Variogram Rose with x,y-transformed coordinates.

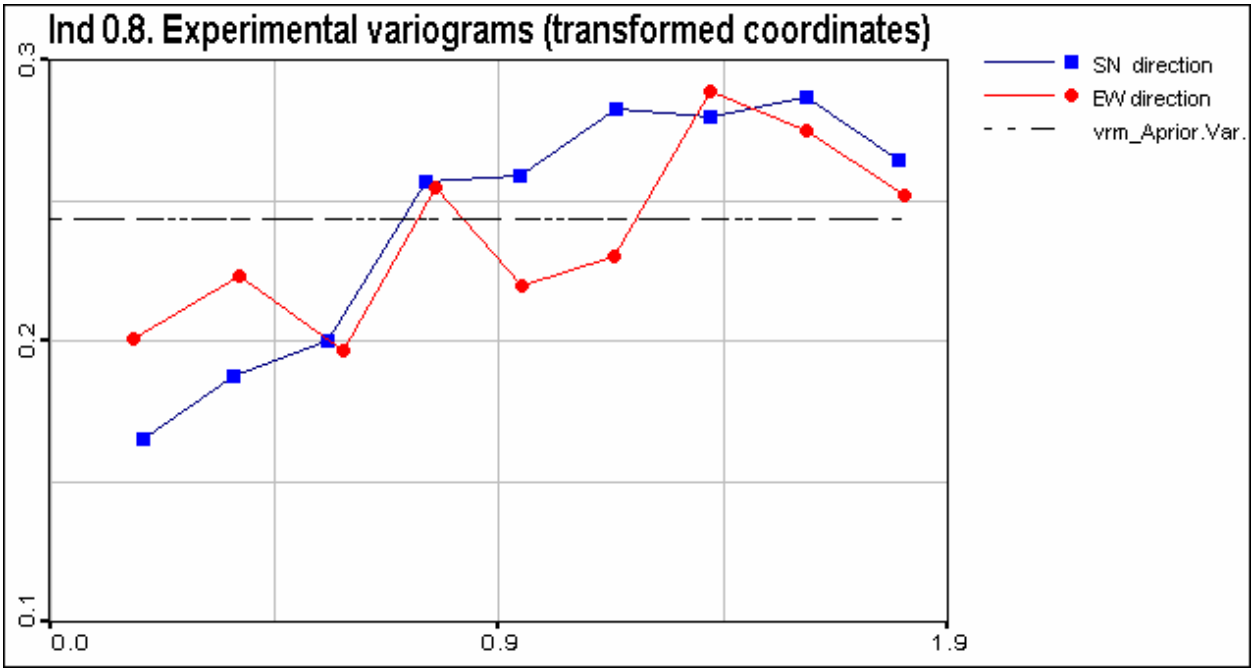

Figure 22. Indicator 0.8. Experimental anisotropic variograms (transformed coordinates).

It should be noted, that variogram in transformed coordinates is more isotropic than in original coordinates. In fact, it was not an objective of this transformation. In fact, both topology of the monitoring networks (input space) and spatial correlation structures (output space = phenomena) should be taken into account. Probably, preprocessing of the data by affine transformation and taking into account anisotropic structure of the phenomena can help SVM in classification. This investigation is under progress.

Indicator variogram modelling

Indicator variogram is a variogram calculated for the indicators.

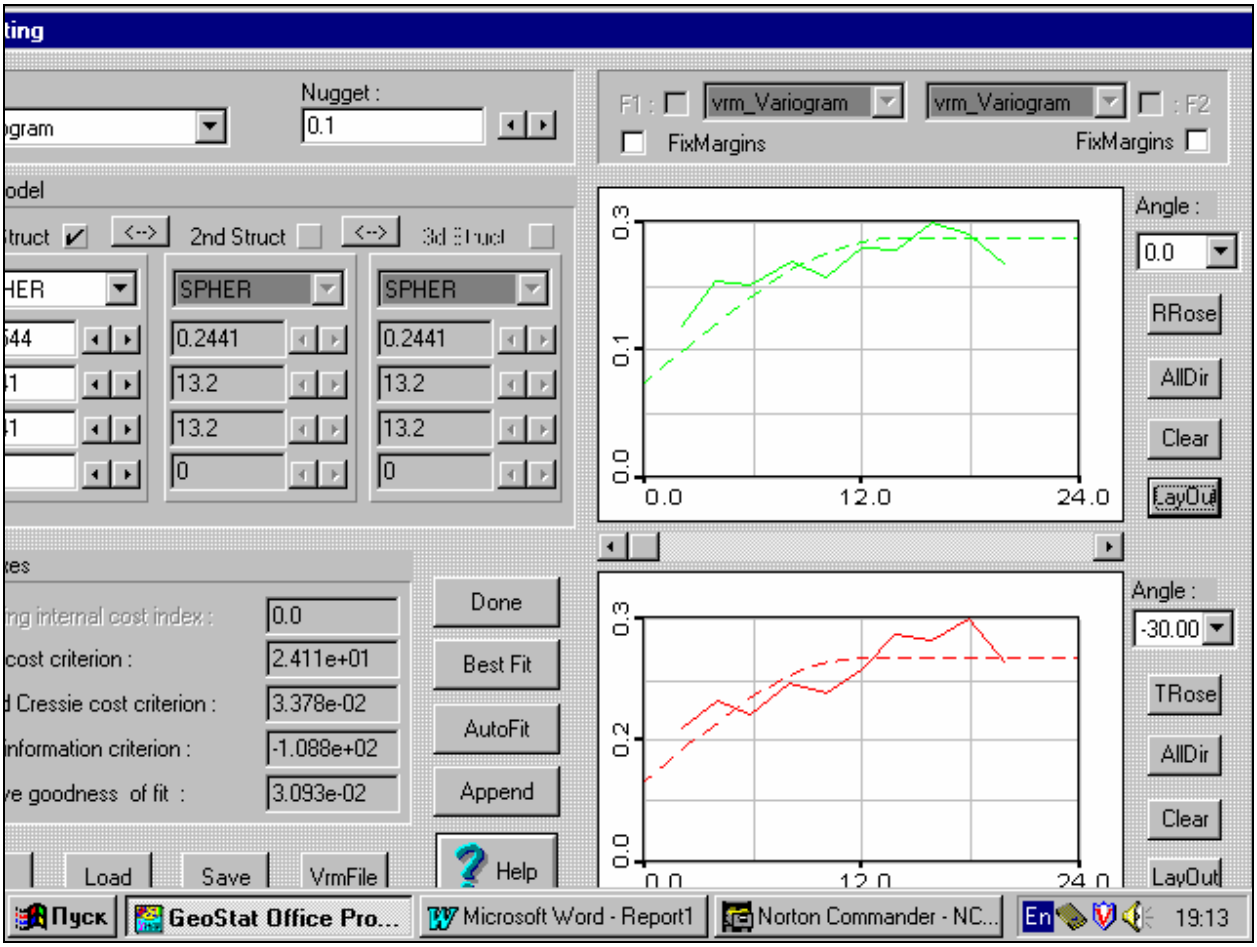

Figure 23. Anisotropic 0.8 indicator variogram fitting with the Geostat Office.

With the help of Geostat Office software anisotropic variograms have been modeled. Modelling means fitting of known theoretical variogram structures (simple formulas with few parameters) to the experimental variograms calculated using data [Goovaerts 1997, Journel and Deutsch 1997]. Interactive module of the Geostat Office for the variogram modelling is presented in Figure 22.

Nested variogram consisted of pure nugget effect (constant value) and spherical models with different ranges in different directions described above was used for the indicator kriging. The formula for the spherical model is the following:

$$\gamma_{sph} = cSph(\frac{h}{a}) = \begin{cases} c[1.5(h/a) - 0.5(h/a)^3, & \text{if } h \le a \\ c & \text{if } h \ge a \end{cases}$$

where a defines a range of correlation and c is the so called sill (Plato).

Indicator kriging. Results

Let us consider some results of the indicator kriging. First, remind that nugget value (behaviour of the variogram near the origin) is rather high. It means that there is a high stochastic component and small-scale variations are important.

The outputs of the indicator kriging are treated as a ccdf value or posterior probabilities. In the present case the outputs are probabilities that Cd concentrations does not exceed level 0.8 μg/g. The result of indicator kriging is presented in figure 24 (IK estimates). For the convenience results for the probability of exceeding level 0.8 are presented [probability of exceeding = (1-probability of not exceeding)] validation data are also plotted. Some data points are really difficult to classify correctly.

Results of indicator kriging along with optimal SVM classification are presented in figure 25. With the exception of some details, results seem to be in a good agreement with each other and with validation data.

At this point it should be noted that probabilistic mapping with indicator kriging as the geostatistical methods for the comparisons was used instead of other spatial classifiers. In this study it was important to understand spatial uncertainty and variability of data and models. Actually, indicator simulations should give much better quantification of the spatial pattern uncertainty and variability. This work is currently under progress.

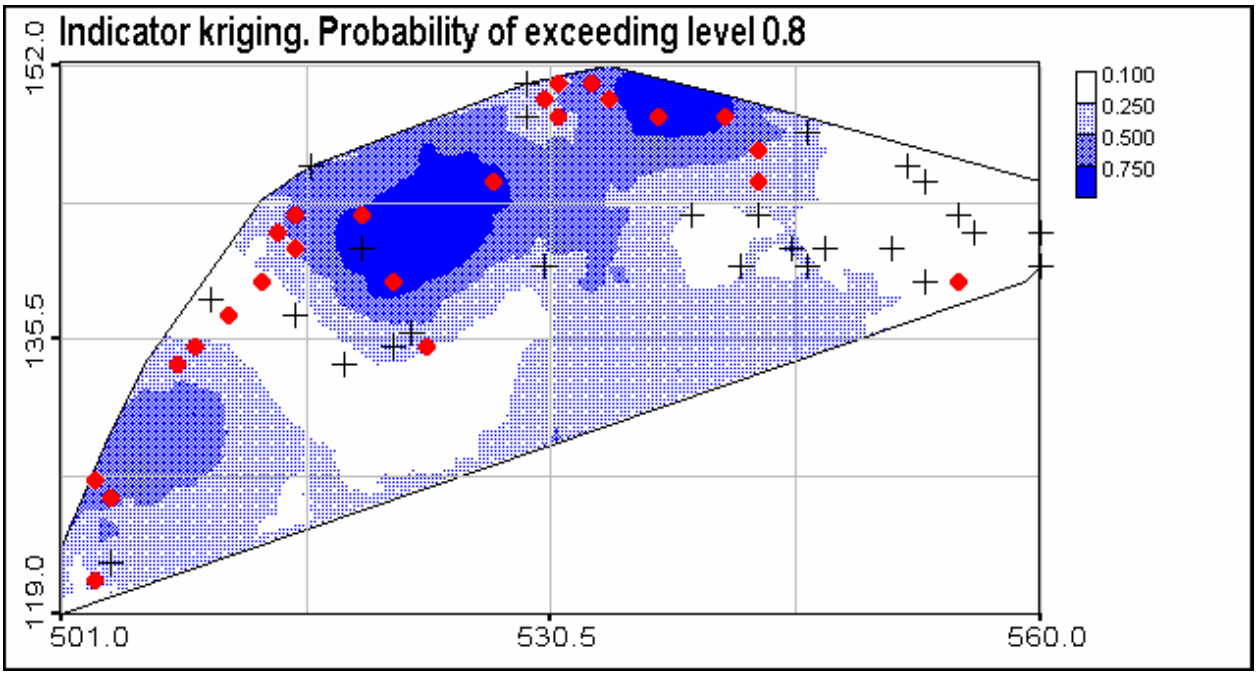

Figure 24. Indicator kriging. Probability of exceeding level 0.8. Validation data are presented as well.

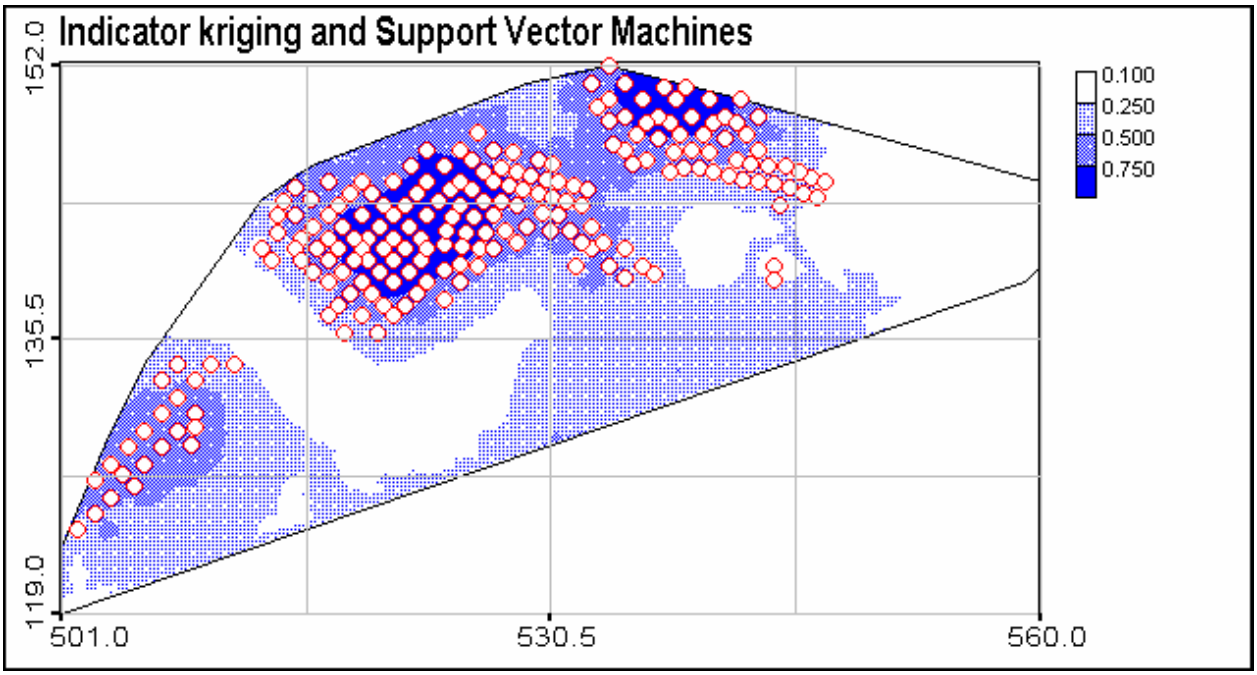

Figure 25. Indicator kriging and Support Vector Machines.

5. CONCLUSIONS AND DISCUSSION

The first and preliminary results of the SVM application for spatial data classification are promising. It was shown that the quality and quantity of the information extracted from data can be controlled by changing kernel parameters and using testing data sets. An important problem for the future research consist in developing data-driven automatic selection of the optimal bandwidth parameters.

In general, the problem of joint multi-class classification is more important for the real environmental and pollution decision-making. Usually several thresholds are important. There are some developments for this case both in statistical learning theory and geostatistics which are interesting to be studied.

Spatial uncertainty and variability of indicators can be described and modelled with the help of conditional stochastic simulations. It seems that direct estimation of the local conditional distribution function (probabilistic treatment of the SVM's outputs) can improve both data treatment and interpretation of the results.

An important problem deals with the number of support vectors and the quality of classification. There is some indication that near the optimum the number of support vectors is minimal. It should be noted, that expected testing error is bounded by the ratio of the expected number of support vectors to the number of training data [Vapnik 1995]. This can be additional criteria for the selection of SVM parameters (e.g. bandwidth).

Another important question which is under study concerns pre-processing of raw data and its influence on the results. Usually data driven approaches highly depends on data pre-processing.

Finally, there are some question related to the selection of kernel function and hyperparameters.

6. ACKNOWLEDGMENTS

The work was supported in part by Swiss National Research Foundation (Cartann project: FN 2100-054115.98) and by European INTAS grant 96-1957. The work was carried out during M. Kanevski's stay at IDIAP as a visiting Professor. The authors thank to Geostat Office group (IBRAE, Moscow) for the access to the Geostat Office software.

7. REFERENCES

Boser B.E., I.M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers, In *Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, 1992. ACM.

Burges C.J.C. A tutorial on Support Vector Machines for patterns recognition. *To appear in Data Mining and Knowledge Discovery*. 1998.

Cortes C. and V. Vapnik. Support vector networks. *Machine Learning*, 20: 273–297, 1995.

Cherkassky V and F. Mulier. *Learning from data*. John Wiley & Sons, Inc. N.Y. 441 p. 1998.

Cristianini N., Campbell C., J. Shawe-Taylor NeuroCOLT2 Technical Report Series, NC2-TR-1998-017. 12 p. 1998

Deutsch C.V. and A.G. Journel. *GSLIB. Geostatistical Software Library and User's Guide*. Oxford University Press, New York, 1997.

Goovaerts P.. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.

Kanevski M., V. Demyanov, S. Chernov, E. Savelieva, A. Serov, V. Timonin, M. Maignan. Geostat Office for environmental and pollution data analysis. *Mathematische Geologie,* Dresden, April 1999.

Smola A.J. and B. Schölkopf. A tutorial on Support Vector Regression. *NeuroCOLT2 Technical Report Series, NC2-TR-1998-030.* October, 1998.

Vapnik V.. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

Weston J., A. Gammerman, M. Stitson, V. Vapnik, V. Vovk, C. Watkins. Density Estimation using Support Vector Machines. *Technical Report, Csd-TR-97-23*. February 1998.

Weston J, Watkins C. Multi-class Support Vector Machines. Technical Report CSD-TR-98-04, 9p. 1998.