

Environmental and Pollution Spatial Data Classification with Support Vector Machines and Geostatistics

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

(1) IDIAP Research Institute, Martigny, Switzerland

(2) Institute of Mineralogy and Petrology, University of Lausanne, Switzerland

Abstract. The work deals with the application of Support Vector Machines (SVM) for environmental and pollution spatial data analysis and modeling. The main attention is paid to classification of spatially distributed data with SVM and comparison with probabilistic mapping using nonparametric geostatistical model (indicator kriging). SVMs with RBF kernels were used. It is shown that optimal bandwidth of kernel can be chosen by minimizing testing error. Real data on sediments pollution in the Geneva lake are 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 approach for working 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 [1][2]. Geostatistics in general is 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 function 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 Cd contamination. The results are compared with SVM classification.

The present work deals with the development and adaptation of geostatistical methods 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.

2. Support Vector Machines

In the early nineties emerged a new paradigm of learning from data called Support Vector Machines (SVM) [4][5]. At first, it was proposed essentially for classification problems of two classes (dichotomies), but now it has been generalized to regression problems [7] as well as to estimation of probability densities [8].

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 [6] for a complete tutorial on SVMs).

2.1. Principle of SVMs

Consider a dichotomy defined by a set of K couples $\{(\mathbf{x}^k, y^k)\}_{k=1,\dots,K}$ in $R^n \times \{-1,+1\}$, where the data point \mathbf{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 $f(\mathbf{x}^k)$ is of the sign y^k hopefully for any $k=1,\dots,K$ and moreover, for any such k the point \mathbf{x}^k lies as far as possible from the decision surface $f=0$.

For simplicity, let first assume that f is a linear function: $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$. If the dichotomy is linearly separable, there exist a vector \mathbf{w} and a b such that $(\mathbf{w} \cdot \mathbf{x} + b)y^k > 0$ for all k . The hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ does not change by rescaling \mathbf{w} and b , and if $\|\mathbf{w}\|=1$, the distance between a point \mathbf{x} and the hyperplane is given by $|\mathbf{w} \cdot \mathbf{x} + b|$. Thus, if the dichotomy is linearly separable, the pair (\mathbf{w}, b) chosen by the support vector machine is the optimal solution of the following problem:

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

or equivalently:

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

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 \mathbf{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 $(\mathbf{w} \cdot \mathbf{x}^k + b)y^k \geq 1$ are replaced by $(\mathbf{w} \cdot \mathbf{x}^k + b)y^k \geq 1 - \xi^k$. Of course, as few ξ^k as possible should be non zero, thus a second objective is to minimize $\sum_k \xi^k$. The new problem has the form

$$\min \|\mathbf{w}\|^2 + C \sum_k \xi^k \\ \text{under the constraints that } (\mathbf{w} \cdot \mathbf{x}^k + b)y^k \geq 1 - \xi^k \quad \text{and } \xi^k \geq 0 \quad \forall k,$$

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 $\Phi: R^n \rightarrow R^N$ admit a kernel function $K: R^n \times R^n \rightarrow R$ with the property that $K(\mathbf{x}^1, \mathbf{x}^2) = \Phi(\mathbf{x}^1) \cdot \Phi(\mathbf{x}^2)$. Thus, even for mappings Φ so that $N \gg n$, 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

$$\min K(\mathbf{w}, \mathbf{w}) + C \sum_k \xi^k$$

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

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

- *Polynomial kernel:* $K(\mathbf{x}^1, \mathbf{x}^2) = (\mathbf{x}^1 \cdot \mathbf{x}^2 + 1)^p$.
The result of an SVM with polynomial kernel is a polynomial of degree p .
- *Radial Basis Function (RBF) kernel:* $K(\mathbf{x}^1, \mathbf{x}^2) = \exp(-\|\mathbf{x}^1 - \mathbf{x}^2\|^2 / 2\sigma^2)$.
The result of an SVM with RBF kernel is an RBF network where σ^2 is the variance of the RB functions.
- *Hyperbolic tangent kernel:* $K(\mathbf{x}^1, \mathbf{x}^2) = \tanh(\kappa \mathbf{x}^1 \cdot \mathbf{x}^2 - \delta)$.
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. Probabilistic Mapping with Indicator Kriging

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 [1][2]. Indicator is a function $I = \text{Ind}(Z = Z^*) = 1$ 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 \mathbf{x} the conditional cumulative distribution function (ccdf) value or posterior probability: $F(\mathbf{x}; Z^* | (n)) = \text{Prob}\{Z(\mathbf{x}) \leq Z^* | (n)\}$ where the conditioning information consist of n data measurements and $\mathbf{x} = (x_1, x_2)$ 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) Unbiased Linear Estimator (BLUE) of the random function. Each ccdf value can be estimated as linear combination of neighboring indicator data using kriging algorithm [2]: $[F\{I(\mathbf{x}; Z^* | (n))\}]_{\text{IK}} = \sum \lambda_i(\mathbf{x}; Z^*) I(\mathbf{x}_i; Z^*)$, where the weights are given by an ordinary kriging system [1][2].

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

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

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

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

In case of several thresholds cokriging (coestimations) of indicators with analysis and modelling both variograms (autocovariance functions) and cross-variograms (cross-covariance functions) in general should be used.

4. Case study

Data were provided by the CIPEL (International Commission for the Protection of Water of the Lemman Lake, Switzerland). They are of two kind.

The first was chemical analysis of Lemman sediments during the years 1978, 1983 and 1988. These data had not been valorized spatially previously by geostatistical methods.

The second data set is a chemical analysis of water of the Lemman Lake at various depth, in various location, 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 analyzed (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 $\mu\text{g/g}$) 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 is 200.

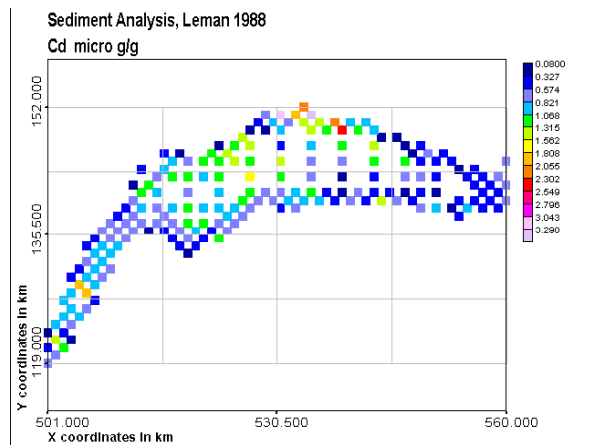
Qualitative and quantitative analysis and description of the monitoring networks and their clustering is an important phase of spatial data analysis. There are different deterministic (describing spatial resolutions), statistical (Morishita diagrams, etc.), and fractal (describing spatial resolution) 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. In the case of Cd measurements clustering is not important.

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). This kind of analysis is of great importance both for the original data analysis and results despite of methods used.

4.1. Cadmium classification

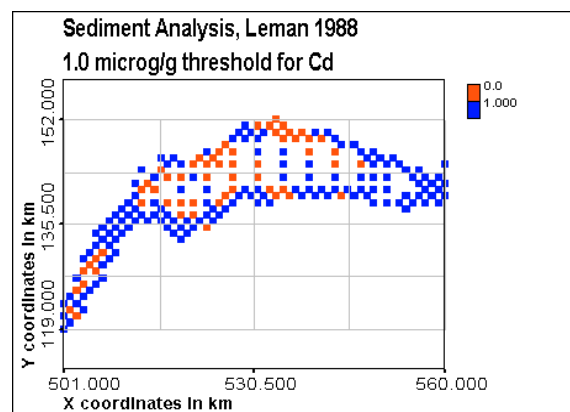
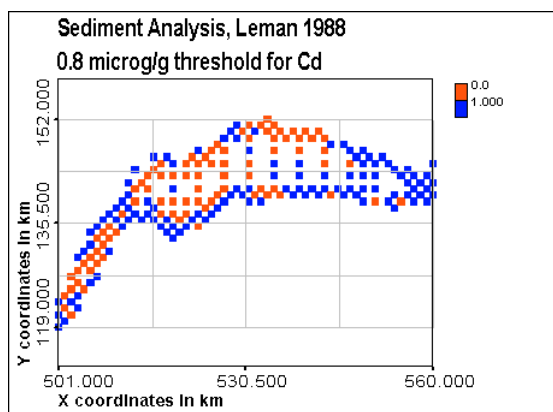
The picture shows the cadmium concentration at each point of measurement, the higher level are reach in the north coast and in the middle of the “Small Lake”, in the southwest. But there is also a large area of medium concentration in the center of the “Great Lake”, and some hot spots on the coasts.

Now, with those observations, we want to make a kind of risk classification of the Leman Lake, in order to know, for a given concentration, which part of it are above this concentration, and which part are below.



Data treatment

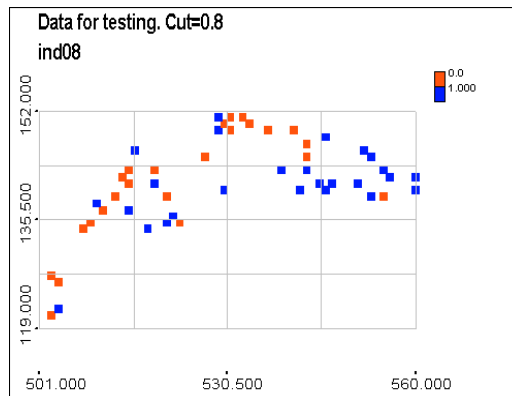
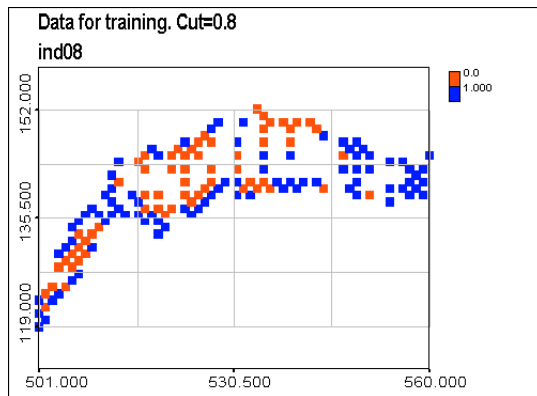
Two thresholds for our case study: the concentration of 0.8 $\mu\text{g/g}$ and of 1.0 $\mu\text{g/g}$. This work has been made with Geostat Office, a data analysis and treatment software for the spatial data analysis and modeling [8].



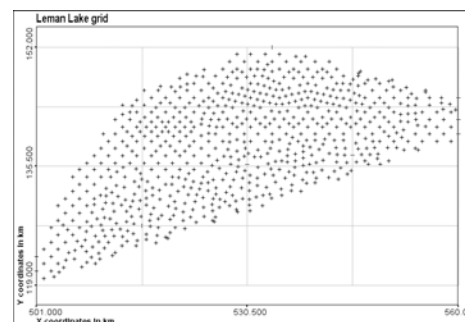
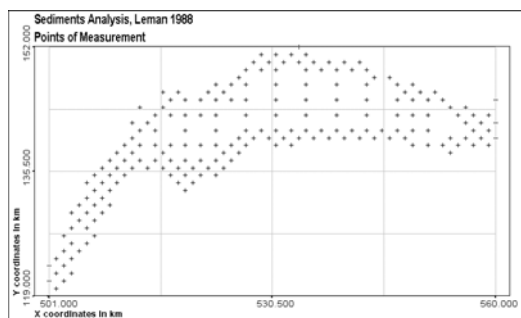
The choice of a 0.8 threshold (very near of the mean value of the data set) gives a quite large area of matching points. In comparison, the choice of the 1.0 threshold concentrates the information much more on the location of very high concentration areas. The future classification will then result in two patterns, different enough to conclude on classifiers' efficiency. Let us note, that both cases are nonlinear 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 see the efficiency of the classifier.

This job has been made with another software, called Netman [8], that permit 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 data network, as shown below for the 0.8 threshold. In general, Netman is a powerful collection of tools for the quantitative and qualitative monitoring networks description, analysis and modeling.



For the geographical grid, Netman generated a network of 720 points based on the initial position of the data point. There are some points outside of the lake borders, but we will see that it does not pose any problems, except for visualization. It should be noted that final results should be prepared 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 object: points, lines and polygons.



4.2. 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 optimizer. At this time, we are testing the efficiency of the RHUL (Royal Holloway University of London, United Kingdoms) software on SVM, in order to compare the results.

Quality of results criteria

The construction of a classifier is decided like this. First of all, we are choosing a kernel type (we tried three of them). Then, we are choosing the specific parameters of this kernel. After this, we are calculating the support vectors' coefficients with the optimizer, thanks to the training data. And then, we are testing the efficiency of those coefficients and kernel's parameters on the testing data.

At the end of this work, we obtain two error ratios, given by:

$$\frac{\text{Number_of_misclassified_data}}{\text{Total_number_of_data}}$$

One of them is specific to the training data and the other to the testing data. Our objective is to minimize both, of course. So, we are usually talking of “good results” when the training and testing errors are both below 10% and 20 % respectively.

Choice of the kernel

The choice of the kernel is a crucial issue in the SV method. With the polynomial kernel, we obtained some good results with a degree of 9 for the function. 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. the coasts of the lake). This is also a well-known problem with polynomial regression on one-dimensional function.

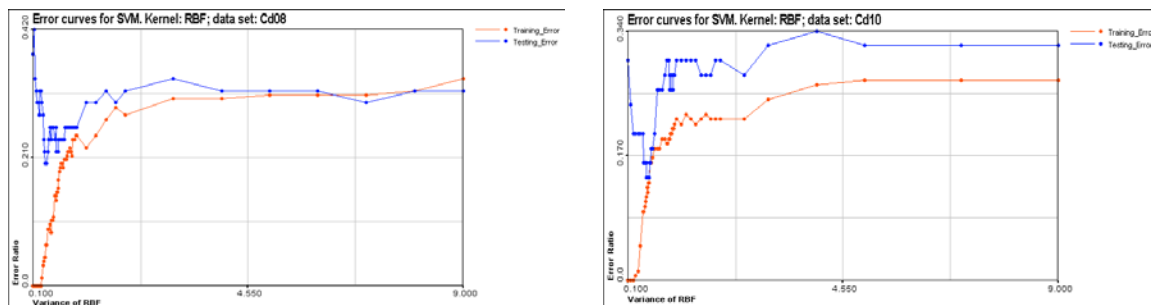
The case of the hyperbolic tangent kernel is very different. In fact, this kernel is using two parameters that are very difficult to optimize with our data

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

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, we tried to generate error curves.

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



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 quasi-null. This is the overfitting part of our 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.

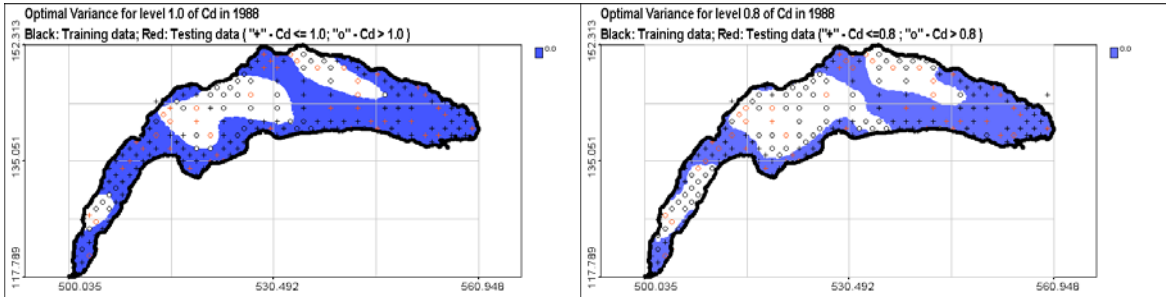
At the end, the two curves are reaching a plateau at a high error value. In this part, nothing can be decided because the classification does not work at all.

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

Results of classification

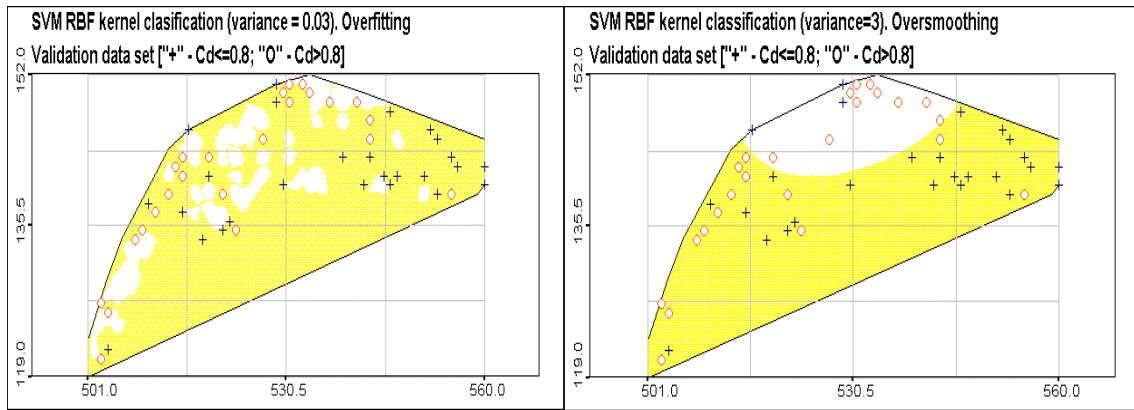
By using different bandwidths variability of the classification, results can be demonstrated: from overfitting at small bandwidth to oversmoothing with rather high values of the kernel bandwidth. The RBF kernel delivers, after using the SVM calculated on the grid, over-fitting and over-smoothing effects, to compare with the optimal parameter result.

All the point outside of the lake borders are set to 1, meaning that they are not considered to be high concentration points. This seems to be obvious, but with polynomial kernel, many points outside of the lake where set to 0, that can not be accepted.



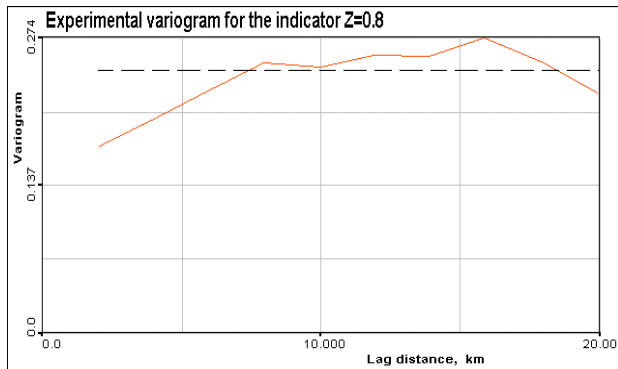
As for the shape of the classification, the comparison of the two optimal classifications for the two thresholds gives a precision on the level of contamination, improving our “risk mapping”.

For the comparison two results with suboptimal (overfitting, oversmoothing) bandwidths are presented below for the level 0.8. Validation data are represented as well.



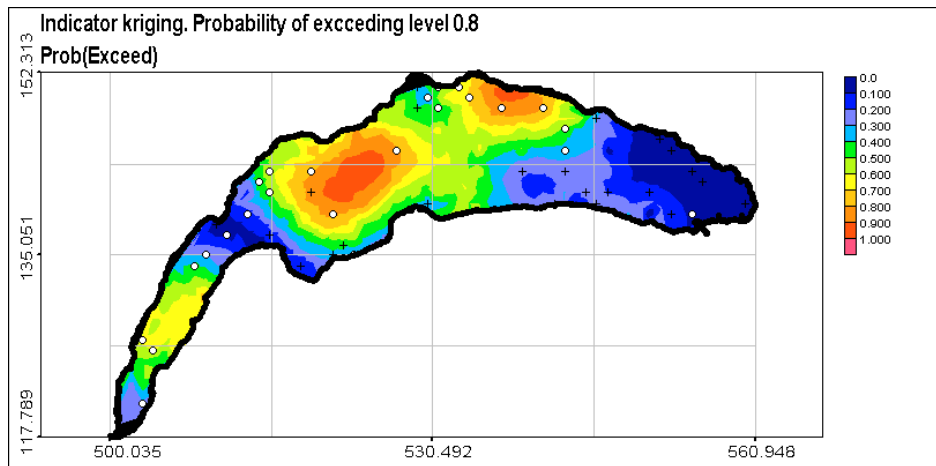
4.3. Indicator kriging

Experimental (based on original data) omnidirectional variogram for the Cd indicator ($Cd=0.8 \mu\text{g/g}$) is presented in figure. It should be noted, that nugget (behavior of the variogram near the origin) is rather high. It means that there is a high stochastic component and small scale variations are important. After theoretical modeling of the variogram (fitting experimental variogram to the known theoretical variogram models described by simple formulas [1]) it is used in the indicator kriging equations.

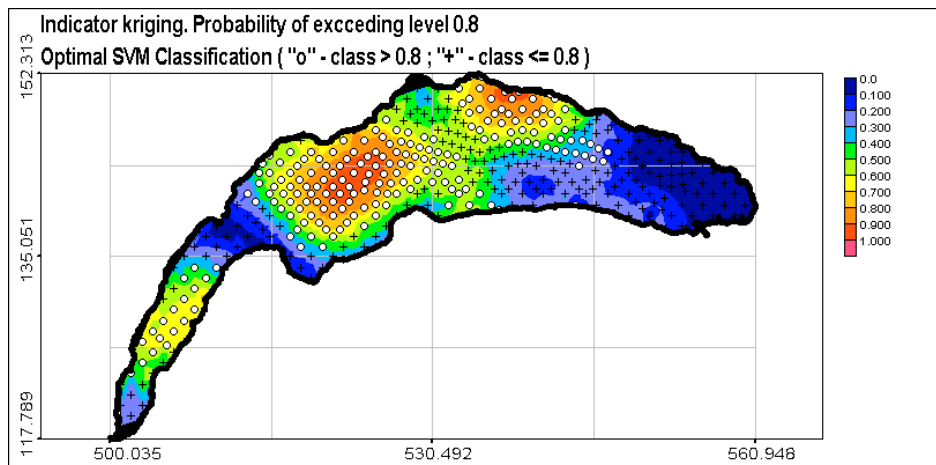


The outputs of the indicator kriging are treated as a ccdf value or posterior probabilities: in our case the outputs are probabilities that Cd

concentrations exceed level $0.8 \mu\text{g/g}$. The result of indicator kriging is presented in Figure (IK estimates). Validation data are also plotted on the same figure. Some data points are really difficult for the classification.



Results of indicator kriging along with optimal SVM classification are presented in figure below. With the exception of some details, results seem to be in a very good agreement.



At this point it should be noted that we used probabilistic mapping with indicator kriging as the geostatistical methods for the comparisons despite there are other real spatial classifiers. In this case 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. At present this work is under progress.

5. Conclusions and discussions.

The first and preliminary results of the SVM application for spatial data classification are promising. 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 of developing data-driven automatic selection of the optimal bandwidth

parameters based on a clear criteria (cross-validation?, jackknife ?, bootstrap?, variography?). Should bandwidth be spatially adaptive?

In general, the problem of multiclass classification is more important for the real environmental and pollution decision making. Usually several thresholds are important. The basic geostatistical problems with multiclass classification with indicator approach deal with different spatial correlations for different indicators, possible cross-correlations between classes, spatial nonstationarity.

It should be noted, that both approaches give smooth outputs of classification. In the present case study and with a given resolution boundary of the indicator kriging outputs is more variable.

Spatial uncertainty and variability of indicators can be described and modeled 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.

These and many others related problems are the tasks for the research in the nearest future.

6. Acknowledgments.

The work was supported in part by Swiss National Research Foundation (Cartann project) and by INTAS project 1957.

7. References

- [1] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.
- [2] C.V.Deutsch and A.G. Journel. *GSLIB. Geostatistical Software Library and User's Guide*. Oxford University Press, New York, 1997.
- [3] B.E.Boser, I.M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers, In *Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, 1992. ACM.
- [4] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20: 273–297, 1995.
- [5] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
- [6] C.J.C Burges. A tutorial on Support Vector Machines for patterns recognition. *To appear in Data Mining and Knowledge Discovery*.
- [7] A.J.Smola and B. Schölkopf. A tutorial on Support Vector Regression. *NeuroCOLT2 Technical Report Series, NC2-TR-1998-030*. October, 1998.
- [8] J. Weston, A. Gammerman, M. Stitson, V. Vapnik, V. Vovk, C. Watkins. Density Estimation using Support Vector Machines. *Technical Report, Csd-TR-97-23*. February 1998.
- [9] M. Kanevski, 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.