Российская Академия Наук

РОССИЙСКАЯ АКАДЕМИЯ НАУК

ИНСТИТУТ ПРОБЛЕМ БЕЗОПАСНОГО РАЗВИТИЯ АТОМНОЙ ЭНЕРГЕТИКИ

RUSSIAN ACADEMY OF SCIENCES

NUCLEAR SAFETY INSTITUTE

Препринт ИБРАЭ № IBRAE-2002-18

Preprint IBRAE-2002-18

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

ROBUST MAPPING OF SPATIAL DATA WITH SUPPORT VECTOR REGRESSION

Москва
2002

Moscow
2002

УДК 502.3

Позднухов А., Каневский М., Майгнан М., Каню С. РОБАСТНОЕ КАРТИРОВАНИЕ ПРОСТРАНСТВЕННЫХ ДАННЫХ С ПОМОЩЬЮ РЕГРЕССИИ НА ОПОРНЫХ ВЕКТОРАХ. Препринт № IBRAE-2002-18. Москва: Институт проблем безопасного развития атомной энергетики РАН, 2002. 17 с. — Библиогр.: 14 назв.

Аннотация

В работе описана модель Support Vector Regression (регрессия на опорных векторах), основанная на статистической теории обучения. На примере реальных данных по загрязнению почв Брянской области радионуклидом ^{137}Cs рассмотрена методология применения SVR модели к решению задачи пространственного картирования. Подробно рассмотрены методы настройки параметров модели. Исследовано поведение модели при наличии в данных искусственного шума и выбросов (outliers). Обсуждаются различные способы использования дополнительной информации о точности проведенных измерений для улучшения прогноза путем внесения этой информации в модель SVR. Представлены предварительные результаты моделирования данных по окружающей среде с использованием модификации SVR модели, позволяющей моделировать нестационарные данные при наличии в них пространственных структур различных масштабов.

Pozdnukhov A., Kanevski M., Maignan M., Canu S. ROBUST MAPPING OF SPATIAL DATA WITH SUPPORT VECTOR REGRESSION. Preprint IBRAE-2002-18. Moscow: Nuclear Safety Institute RAS, July 2002. 17 p. — Refs.: 14 items.

Abstract

The paper is devoted to the description of the Support Vector Regression - a model based on the Statistical Learning Theory. The methodology of application of the SVR models to the problem of spatial data prediction mapping is considered on a real case study: soil contamination with ^{137}Cs radionuclide in Briansk region. The procedure of SVR hyper-parameters tuning is considered in details. The cases of noisy data and data with outliers were considered. A way of incorporation the additional information on measurements quality is discussed. Preliminary results on multi-scale SVR modeling are presented.

Robust mapping of spatial data with Support Vector Regression

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

ИНСТИТУТ ПРОБЛЕМ БЕЗОПАСНОГО РАЗВИТИЯ АТОМНОЙ ЭНЕРГЕТИКИ
113191, Москва, ул. Б. Тульская, 52
тел.: (095) 955-22-31, факс: (095) 958-11-51, эл. почта: anp@ibrae.ac.ru

Contents

1 Introduction

Spatial data mapping is of great importance in various fields: environment, health care, economics. A lot of state-of-the-art methods can be used to make predictions based on some empirical data (measurements): deterministic interpolations, methods of geostatistics: the family of kriging estimators [3], machine learning algorithms such as artificial neural networks (ANN) of different architectures, hybrid ANN-geostatistics models [8], etc.

All the methods mentioned above can be used for solving the problem of spatial data mapping. But when dealing with empirical data we can't trust our data completely as it is always corrupted by noise, sometimes by noise of unknown nature. That's one of the reasons why deterministic models can be inconsistent, since they treat the measurements as values of some unknown function that should be interpolated. Kriging estimators treat the measurements as the realization of some spatial random process. But for obtaining the estimation with kriging one has to estimate the spatial structure of data: spatial correlation function or (semi-)variogram, while this task can be complicated if there is not sufficient number of measurements. ANN is a powerful tool, but it is also suffer from the number of reasons. ANNs of a special type – multiplayer perceptrons – are often used as a detrending tool in hybrid (ANN+geostatistics) models [8].

So, we are interested in a method that would be robust to noise in measurements, would deal with the small empirical datasets and has solid mathematical background.

The present paper deals with such model, based on Statistical Learning Theory (SLT) - Support Vector Regression. SLT is a general mathematical framework devoted to the estimating of the dependencies from

empirical data [14]. SLT models for classification - Support Vector Machines - have shown promising results on different machine learning tasks. The results of SVM classification of spatial data are also promising [5,7]. The properties of SVM for regression - Support Vector Regression (SVR) are less studied. First results of the application of SVR for spatial mapping of physical quantities were obtained by the authors in [5] for mapping of medium porosity, and [6,9] for mapping of radioactive contaminants activity. The presented paper is devoted to further understanding of properties of SVR model for spatial data analysis.

We'll present the theory of the method in section 2. Section 3 discusses standard application of SVR for spatial data mapping on the real case study - soil pollution by Cs137 radionuclide. Sections 4 discusses the properties of the model applied to noised data or data with outliers. Section 5 is devoted to the extension of SVR model that allows incorporating some extended information into the standard SVR. Section 6 presents some preliminary results on the multi-scale modeling using SVR. Section 7 concludes the presented paper.

The paper partly follows the terminology used in Machine Learning.

2 Theory of SVR

2.1 Problem statement in Statistical Learning Theory and Loss functions

First we sate general problem of regression estimation as it is presented in the scope of Statistical Leaning Theory. Suppose we are given a set of observations generated from an unknown probability distribution P(**x**, y) $\{(\boldsymbol{x}_1,y_1),\ldots(\boldsymbol{x}_N,y_N)\}$ with $\boldsymbol{x}_i \in \mathbf{R}^n$, $y_i \in \mathbf{R}$ and a class of functions $F=\{f \mid \mathbf{R}^n \to \mathbf{R}\}$. Our task is to find a function *f* from the given class of functions that minimizes a risk functional:

$$R[f] = \int Q(y - f(x), x) dP(x, y) \tag{1}$$

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

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

$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} Q(y - f(x), x) \tag{2}$$

When it is known that the measurements are corrupted with additive normal noise, then minimization of the empirical risk with a quadratic loss function results in a best unbiased estimator of the regression *f* in the selected class F. But when it is only known that noise generating distribution is symmetric, the use of (modulus) linear loss function is preferable, and results in a model from so-called robust regression family [11].

Support Vector Regression model is based on a new type of loss functions, so-called ε-insensitive loss functions. For example linear ε-insensitive loss is defined as

$$Q(y - f(x), x) = \begin{cases} |y - f(x)| - \varepsilon & \text{if } |y - f(x)| > \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Quadratic ε-insensitive and Huber ε-insensitive loss functions are also used in SVR. The loss function that can be useful in applications where underestimations and overestimations are not equivalent is a non-symmetric linear ε-insensitive loss:

$$Q(y - f(x), x) = \begin{cases} a(f(x) - y - \varepsilon_a) & \text{if } (f(x) - y) > \varepsilon_a \\ b(y - f(x) - \varepsilon_b) & (f(x) - y) < \varepsilon_b \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

2.2 Connections with Maximum Likelihood estimators

To understand the presented approach of minimizing the risk, we establish the connection between Maximum Likelihood Estimators and risk minimization [12]. If the data have been generated according to the model

$$P(x,y) = P(y \mid x)P(x) = P(y - f(x))P(x) \tag{5}$$

we can write the likelihood of the estimation:

$$P(x_1, y_1, \ldots x_N, y_N) = \prod_{i=1}^{N} P(y_i - f(x_i))P(x_i) \tag{6}$$

Assuming that for some function Q

$$P(y - f(x)) = e^{-Q(y-f(x))} \tag{7}$$

the likelihood can be written as

$$P(x_1, y_1, \ldots x_N, y_N) = e^{-\sum_{i=1}^{N} Q(y_i - f(x_i))} \cdot \prod_{i=1}^{N} P(x_i) \tag{8}$$

and maximization of likelihood results in minimizing

$$R_{ML} = \sum_{i=1}^{N} Q(y_i - f(x_i)) \tag{9}$$

what is equivalent to empirical risk minimization.

2.3 Linear SVR: fitting the hyper-plane

We start from the estimation of regression function in a class of linear functions F = {*f(**x**)* | *f(**x**)=(**w**,**x**)+b*}. Support Vector Regression is based on the Structural Risk Minimization principle, which results in a penalization of model complexity simultaneously with keeping small empirical risk (training error). The complexity of linear functions can be controlled by the term $\|w\|^2$ [14]. By another side, we have to minimize the empirical risk. If the symmetrical linear ε-insensitive loss is used, than empirical risk minimization is equivalent to adding the slack variables ξ_i , ξ_i^* into the functional with the linear constraints (6). Introducing the trade-off constant C, we arrive at the following optimization problem:

$$\text{minimize } \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{10}$$

$$\text{subject to } \begin{cases} f(x_i) - y_i - \varepsilon \le \xi_i \\ -f(x_i) + y_i - \varepsilon \le \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \quad \text{for i} = 1,\ldots N \end{cases} \tag{11}$$

The slack variables ξ_i , ξ_i^* measure the distance between the observation and the ε tube (see the example for non-linear function r(x) in figure 1). The distance between the observation and the ε and ξ_i , ξ_i^* is illustrated by the following example: imagine you have a great confidence in your measurement process, but the variance of the measured phenomena is large. In this case, ε has to be chosen a priori very small while the slack variables ξ_i, ξ_i^* are optimized and thus can be large. Remember that inside the epsilon tube *([f(x)-ε, f(x)+ε])* loss function is zero.

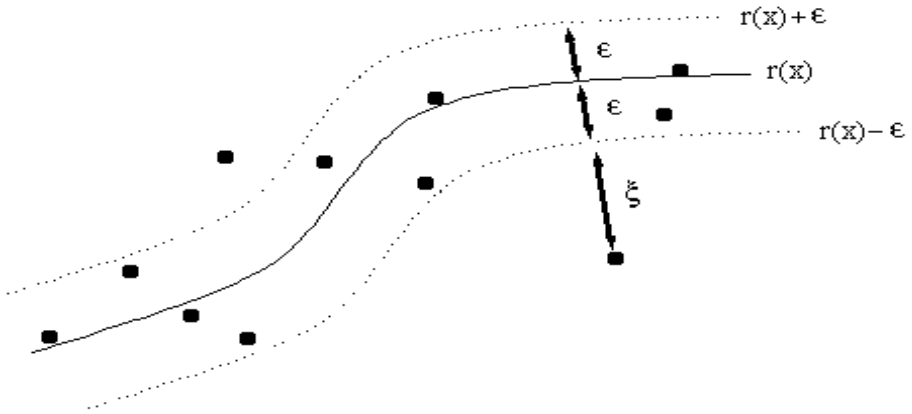

Figure 1. Support vector regression. Explanation of the ε–tube and slack variables

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, $y_i \in [f(x_i)- \varepsilon -\xi_i, f(x_i)+ \varepsilon +\xi_i^*]$.

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

$$L(w,\xi,\xi^*\alpha) = \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}\alpha_i(y_i - f(x_i) + \varepsilon + \xi_i) - \sum_{i=1}^{N}\alpha_i^*(f(x_i) - y_i + \varepsilon + \xi_i^*) - \sum_{i=1}^{N}(\eta_i\xi_i + \eta_i^*\xi_i^*) \tag{12}$$

where $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ are 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". Other points with non-zero α_i or α_i^* are the "support vectors (SVs)" of the problem.

At the minimum the derivative of the Lagrangian equals to zero (Kuhn-Tacker conditions). Thus it can be checked that:

$$\begin{aligned} w &= \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\cdot x_i \\ \eta_i &= C - \alpha_i \quad \text{for i = 1,..., N} \\ \eta_i^* &= C - \alpha_i^* \quad \text{for i = 1,..., N} \end{aligned} \tag{13}$$

These variables can be removed from the original formulation of the minimisation problem to get the dual formulation of the problem:

$$\begin{aligned} \text{maximize } & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) \\ & -\varepsilon\sum_{i=1}^{N}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) \end{aligned} \tag{14}$$

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

This problem is a Quadratic Programming problem hence can be numerically solved by a number of methods. After we get the values α_i and α_i^* we can compute *b* from the constraints of primal problem (5) and make predictions:

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i)(x_i \cdot x) + b \tag{15}$$

2.4 Non-linear SVR. Kernel trick

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

$$\iint K(x', x'') g(x') g(x'') dx' dx'' > 0 \quad \forall g \in L^2(R^n) \tag{16}$$

then it can be expanded in a uniformly converging series

$$K(x', x'') = \sum_j \lambda_j \Phi_j(x') \Phi_j(x'') \tag{17}$$

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

$$\begin{aligned} \text{maximise } & -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \\ & - \varepsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} y_i (\alpha_i^* - \alpha_i) \end{aligned} \tag{18}$$

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

and the prediction is a non-linear regression function:

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) K(x_i, x) + b \tag{19}$$

3 Case Study

In the present section the problem of spatial data mapping (spatial regression) using SVR is considered. We'll consider the mapping of soil pollution by Chernobyl radionuclide Cs137 in western part of Briansk region, Russia.

3.1 General methodology and Data description

The full case study should follow general methodology of spatial data analysis with geostatistics and Machine Learning algorithms: monitoring network analysis, understanding of clustering, exploratory data analysis, exploratory variography, understanding of spatial continuity; data preparation: splitting of data in training testing and validation data subsets, SVR training and testing, selection of the optimal SVR hyper-parameters, understanding the quality of the results by using exploratory analysis (statistics, variography) of the residuals, validation of the results, spatial data mapping with optimal SVR model. First stages are common for all the environmental mapping tasks and will be considered in brief. The examples of full-stage research can be found in [8]. The application of SVR to spatial data mapping is considered in details.

Full data set consists of 684 measurements, Lambert projection co-ordinates are used, pollution values(originally in kBq/sq.m.) were scaled into [0,1] interval. 200 validation data points were selected using

declustering procedure to obtain the representative validation data. Postplot of training (484 measurements) data and the locations of validation points are shown in figure 2, left.

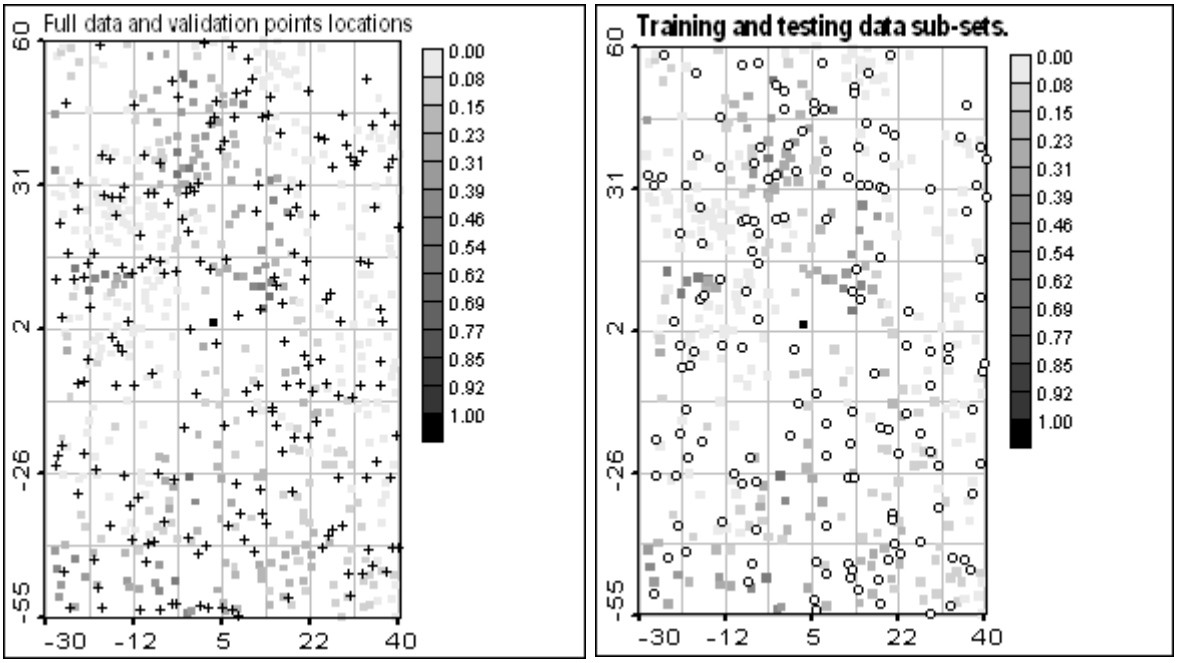

Figure 2. Left postplot: Full data set (boxes) and locations of validation points (crosses); Right postplot: training sub-set (boxes) and testing sub-set (small empty circles)

3.2 Tuning of SVR parameters

There are several hyper-parameters to be selected during SVR tuning: ε, C, and kernel parameters. The kernel choice is an important part of SVR application. Gaussian RBF kernel

$$K(x', x'') = e^{-\frac{\|x'-x''\|^2}{2\sigma^2}}, \quad \text{or} \quad K(x', x'') = e^{-(x'-x'')^T \Sigma^{-1} (x'-x'')}, \tag{20}$$

was found to be the most suitable due to several reasons: it is easy to interpret, it has one free parameter - bandwidth (isotropic kernel) or three free parameters (anisotropic kernel n a 2D input space), it is "local" in a sense that it tends to zero as distance tends to infinity. All these features make the Gaussian RBF kernel attractive for spatial data mapping. Isotropic RBF kernels were used for the present case study.

Let's notice that from SVR theory (see section 2.4) it follows that kernel bandwidth σ is common for all the training points and can't be varied in space, while ε and C can be chosen individually for every training point. We'll show the way how to incorporate additional information into model by setting individual ε_i and C_i to each training point in section 5.

First of all we'll consider the SVR application when ε and C are the same for all the points and discuss general SVR properties for spatial data mapping.

Two wide-used ways to tune the parameters exist: 1) to split data into training-testing subsets and analyse the testing error cube 2) calculate and analyse the leave-one-out error cube. Both ways will be illustrated in the present study. Data splitting into training and testing subset is shown as the postplot in figure 2, right.

Comprehensive search in 3D hyper-parameter space (σ, ε, C) was performed. Some 2D errors surfaces (σ, C) are presented in figures 3. Here and below error surfaces are presented as follows: X-axis - RBF kernel bandwidth σ, Y-axis - logarithm of C parameter. (Note that range of C parameter in leave-one-out error surface differs from the range on the rest figures.) The value of ε parameter was fixed at ε=0.02.

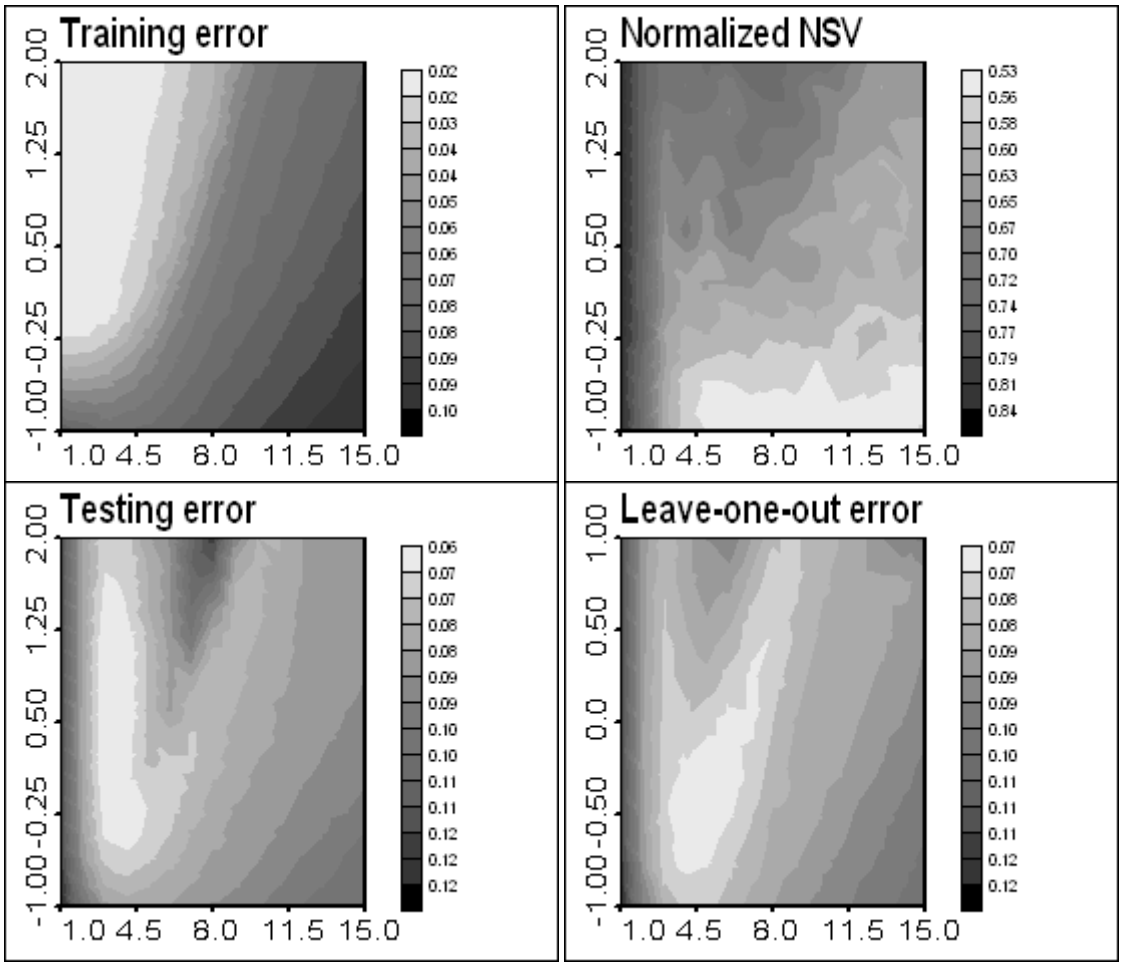

Figure 3. Training error (top left), normalized number of SVs (top right), testing error (bottom left) and leave-one-out error (bottom right) surfaces. X-axis: RBF kernel bandwidth, Y-axis: Logarithm of C parameter

3.3 Influence of the parameters on the solution

Partly basing on the presented error surfaces we consider the influence of the parameters on the solution.

σ - kernel bandwidth. The most evident properties depending on this parameter are: at small σ values - much less than the region's area - the model is close to overfitting, at large σ values - of the order of the region's area - to oversmoothing (or underfitting in ML terminology). The same is clear from SLT side: small σ values lead to too high VC-dimension - high dimension of the feature space - too many features are used for modelling, that leads to overfitting. For large σ values the VC dimension is too low - too few features are used to model the data.

The optimal value of this parameter mainly depends on two characteristics of data: correlation radius and data variability. Another property that was observed is presented below.

Let us fix the testing set and consider several training sets of different sizes (50, 100, 150, 200, 250 points) extracted from the data. The average distance between points is used as a characteristic of the training data monitoring network. Error surfaces for fixed ε=0.05 were obtained for every training set and the best sigma in a sense of minimal testing error was chosen for every set. All the results are averaged over five different training sets of the chosen size. The dependence between "optimal" sigma's and averaged distance is shown in the Figure 4.

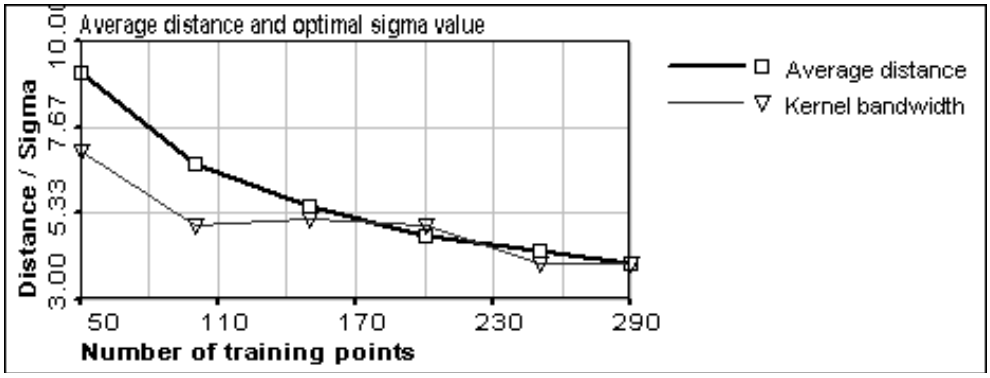

Figure 4. Average distance between training points and "optimal" sigma values

The conclusion one can make is: optimal bandwidth depends of average distance between training points. This dependence is not very evident for the training set of the sizes >150 points, but we can conclude that optimal value of sigma is roughly proportional to the average distance. For small training sets it is evident that we have to use large kernel bandwidths. It is also clear that this dependence vanishes when training set is large enough.

The conclusion above is just a kind of heuristics that can be used for RBF kernel parameter tuning. In general the question of optimal sigma is connected with the complicated question of monitoring network analysis: whether we can describe given/unknown phenomenon with the measurements from given monitoring network.

C - the parameter that defines the trade-off between training error and model complexity, an inverse value is the regularisation constant. In dual formulation C defines the upper bound of the multipliers α_i and α_i^* (see Eq. 18), hence defines the maximal influence the point can exert on the solution. These two considerations allow to conclude: 1) the more noisy the data the less should be the value of C, at the same time the lower the value of C, the smoother the results of prediction mapping 2) C should be not much less then the maximum of the training data to fit the high values well.

ε - the width of the insensitive region of the loss function. This is the parameter that defines the sparseness of the SVR solution - the points that lie inside the ε-tube have zero weights. It is the main parameter that incorporates the information of the measurements' quality. It should be of the same order as the measurements' accuracy or as the square root of nugget in model variogram. It also influence on the smoothness of the mapping: the larger its value, the smoother the result. Let us remind, that nugget incorporates measurement errors and small scale variability.

3.4 Validation of SVR model with chosen parameters

The optimal parameters for the presented case study can be chosen taking into account training, testing and cross-validation errors, and the number of support vectors. Two sets of parameters (hence two models) will be considered: SVR^{test}(C = 4, ε = 0.02, σ = 3) and SVR^{loo}(C = 0.32, ε = 0.02, σ = 4) according to the minimum of testing and cross-validation errors correspondingly. These two models will be compared below.

Table 1. Characteristics of SVR models

	Training RMSE	Testing/LOO RMSE	Normalized Number of SV	Validation RMSE	Correlation coefficient
SVR^{test}	0.019	0.064	0.64	0.078	0.80
SVR^{loo}	0.048	0.071	0.53	**0.064**	**0.86**

Table 1 presents some characteristics of the models. The results on validation data show that SVR^{loo} outperformed SVR^{test}. In [9] it was proposed to tune the parameters basing on the analysis of testing error, and cross-validation error surfaces were not used since testing error calculation is much faster then leave-one-out error calculation. But as it follows from the presented (and several other) case study, it can result in choice of non-optimal parameters. This becomes more clear if we calculate the validation error surface (figure 5). As one can see (compare figures 3 (bottom) and 5), validation error is reproduced by cross-validation error well, what is

not surprising since cross-validation gives a non-biased estimation of the validation error. At the same time, testing error surface just satisfactory reproduces the validation error. Hence if cross-validation calculation is too time consuming, it can be recommended to use k-fold cross-validation or use testing error averaged over several splits.

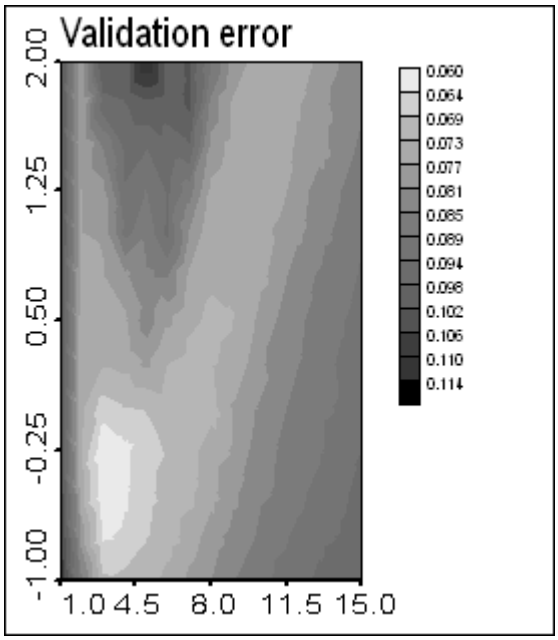

Figure 5. Validation error surface

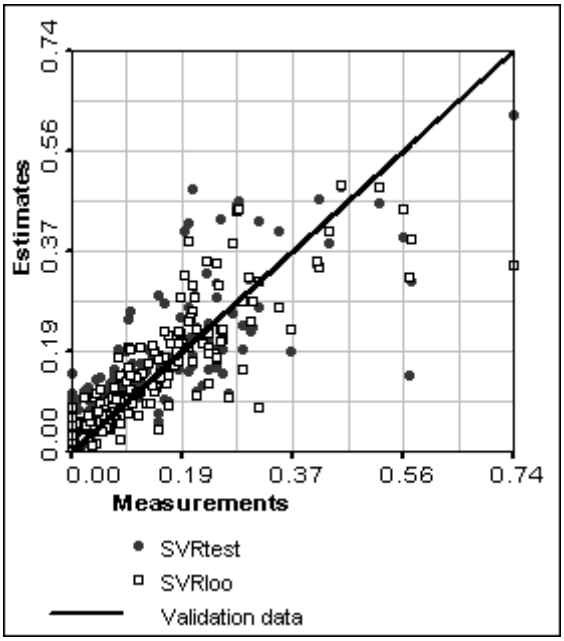

Figure 6. Scatter-plot of validation results

In conclusion the scatter-plot of the validation results is presented in the figure 6.

Another powerful tool to control the quality of the results is the variography of the residuals. Let us consider the omnidirectional variograms of the validation data and the residuals of SVR prediction of validation data for both models (figure 7).

Validation data residuals variograms of both SVR^{loo} and SVR^{test} demonstrate almost pure nugget effect, hence models extracted all spatially structured information. It means good results on validation data.

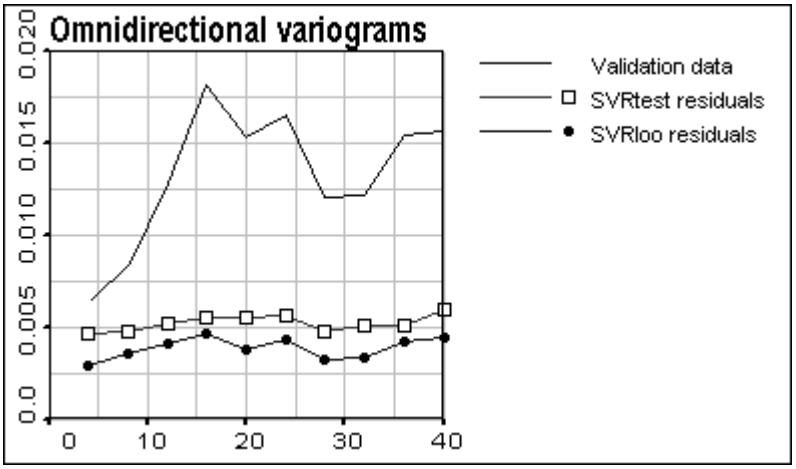

Figure 7. Omnidirectional variogram of validation data and residuals of both models

3.5 SVR prediction mapping

To conclude the part devoted to the application of SVR for the spatial data mapping, we present the map of prediction of Cs137 activity of both models. The spot of high activity in the center is reproduced by SVR^{test} more accurately due to high value of C. SVR^{loo} gives more smoothed result, that nevertheless is better for the rest part

of the mapping area according to validation results. Note that SVR can give negative predictions (empty areas on figures 8).

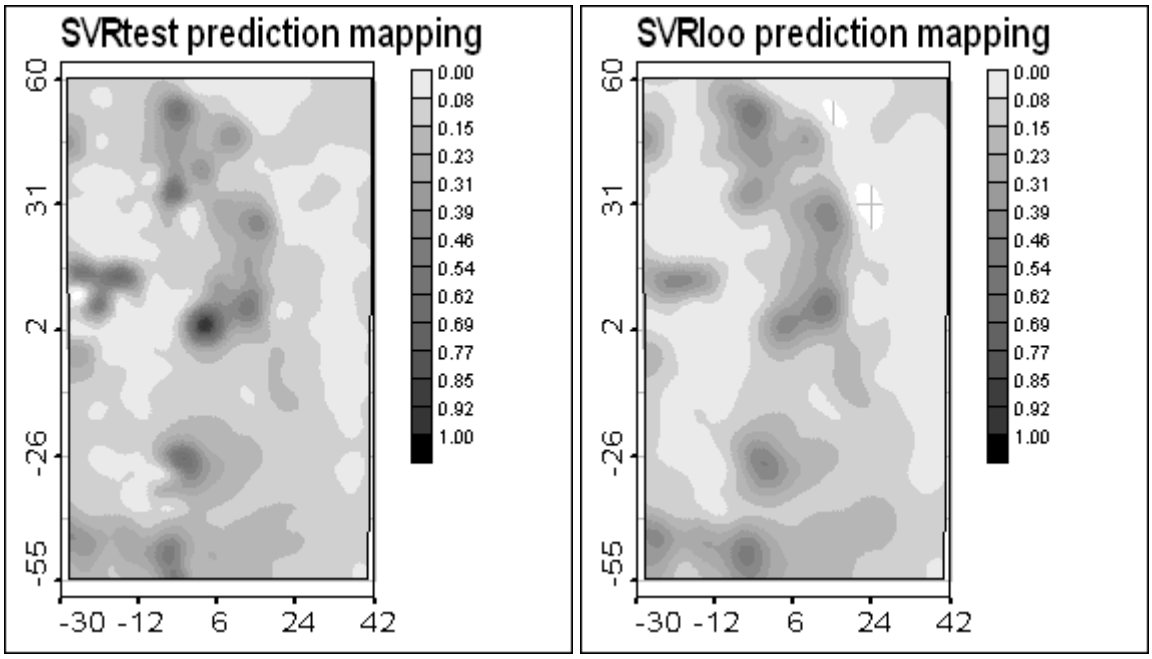

Figure 8. Prediction mapping of SVR^{test}(left) and SVR^{loo}(right) models

4 Robustness of the solution

To understand the robustness of SVR we'll consider two cases: mapping of data with artificially added normal noise and data with artificial outliers in the training set. We'll use the same split into sub-sets as in the previous section. Hyper-parameters were tuned by calculation the testing error as it was discussed above. Although it was shown that cross-validation error calculation is preferable for tuning, testing error surfaces shows all the features of SVR model applied to data with noise or outliers. Obtained testing error surfaces are presented on figure. Parameter ε was set to 0.02 as before. Additional noise has zero mean and variance $(0.02)^2$.

Artificial outliers were obtained from real data by increasing the order of measured value. This can be interpreted as someone's misprint during data collection. Four outliers were added to the training set. Testing set was left free of outliers.

As one can see from the testing error surfaces in figure 9 (compare with figure 3 (bottom left)), though testing error increases, the behavior of the model is stable. "Optimal" parameters were chosen and validation error was obtained. The results are presented in the table below.

Table 2. Characteristics of SVR models. Noise and outliers were added to the training set

	Original data	Noise N(0, 0.02)	Outliers
"Optimal" σ	3	3	4
"Optimal" C	4	2	0.6
Training error	0.019	0.02	0.42
Testing error	0.064	0.061	0.066
NSV	0.64	0.72	0.63
Validation error	0.078	0.079	0.079

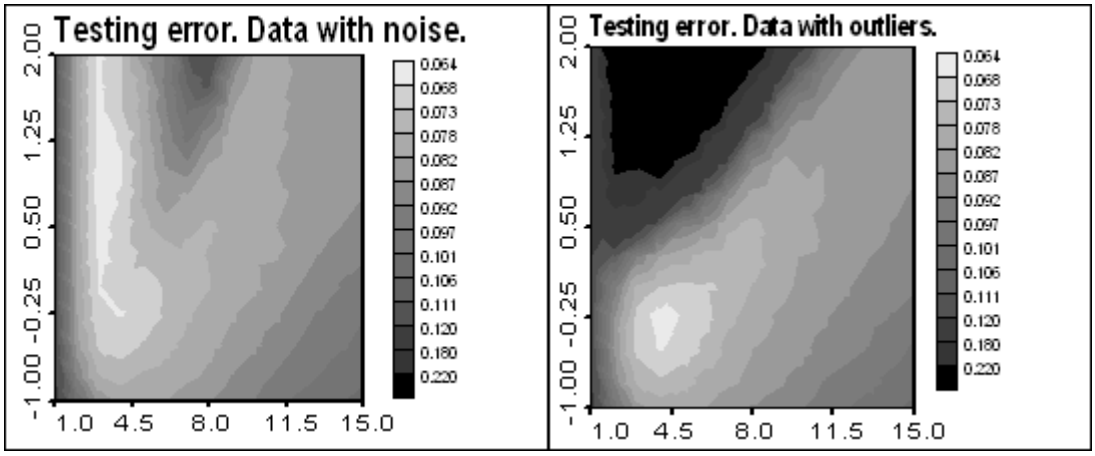

Figure 9. Testing error surfaces: training data with noise (left) and with artificial outliers (right)

As it is seen from the error surfaces, the model "ignores" the outliers when C parameter is small. This causes very high training error, but it is just due to outliers, the rest data is modeled well within the ε-insensitive tube. The ε-insensitiveness also let the model be robust to noise. The results with original data and noised data are quite similar.

The main conclusion one can make is: testing and validation error increased insignificantly, it was possible to select parameters to obtain appropriate model in spite of presence of noise/outliers in training data.

5 Additional information in SVR modeling

This section is devoted to incorporation of additional available information into SVR model. Two types of additional information can be taken into account. First type is an information accessible in all training points and *all the points of the prediction grid*, and the second type is an information that is not accessible in the prediction grid points but known for training set points. Information of the first type (altitude, soil type, etc.) can be incorporated into the model as an extra input variable that will lead to the type of ANN + External Drift (ANNEX) model.

Information of the second type (as a rule related to the measurement's procedure) can be incorporated into SVR by different ways. We'll consider the extended data on the number of probes N_i taken in the given point (from 10 to 300) to obtain the pollution value. Postplot of these data is presented in the figure 10, left.

Large number of probes in a given point mean more exact measurement, thus this information can be used in SVR by setting individual ε_i and/or C_i values for every training point. Both ε_i and C_i are connected with the accuracy of the measurements as it was stated above. Three models will be considered below: SVR^{ε} (fixed C and individual ε_i), SVR^{C} (fixed ε and individual C_i) and $SVR^{C\varepsilon}$ with both individual parameters ε_i and C_i. We'll set $\varepsilon_i = \varepsilon\sqrt{N_{Ave}/N}$, and $C_i = C^{N_{Ave}/N}$, where N_{Ave} = 20 is the average number of probes taken in a point, and ε and C are still free parameters has to be tuned as it was described before. Our aim here is not to stress on tuning the parameters but to show that this additional information incorporated into model can improve the result in a sense of validation error. The characteristics of the models are presented in a table 3.

One can see (compare with SVR^{loo}, the model described before) that an improvement in validation error is achieved with all advanced models. $SVR^{C\varepsilon}$ model even outperformed the "theoretically" best possible SVR (in a sense of validation error) with fixed parameters – see the validation error surface above (figure 5). Training error decreased for all models and number of NSV increased. Mapping results of the $SVR^{C\varepsilon}$ model are presented in the figure 10, right. The spots of high Cs activity are reproduced since these measurements are quite confident according to the number of probes taken at corresponding locations.

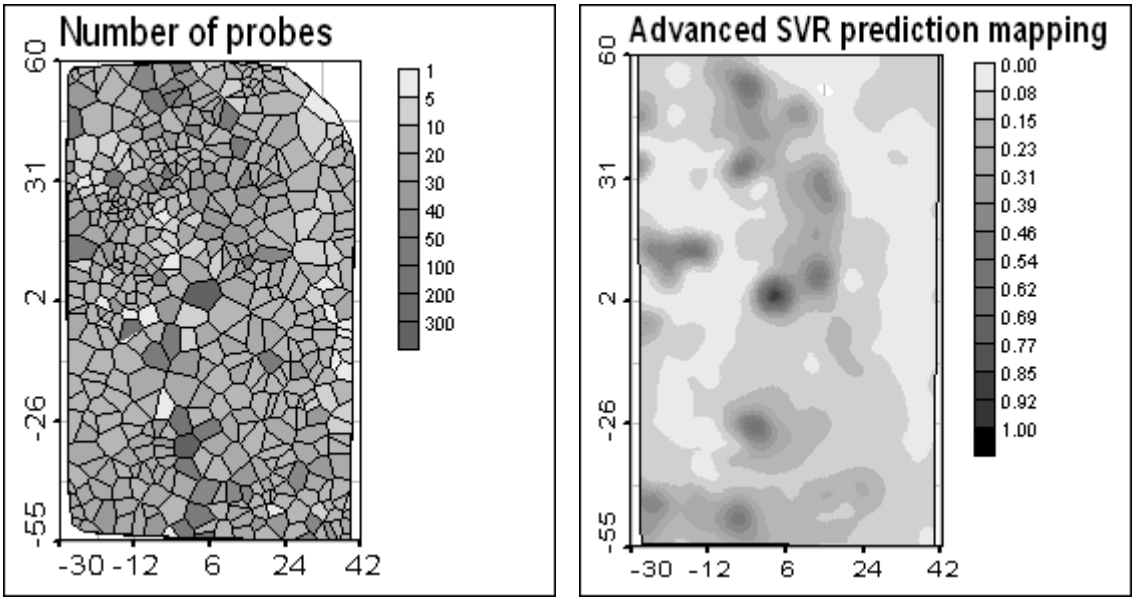

Figure 10. Number of probes taken for the measurements (left). Prediction mapping with advanced SVR model – information on number of probes included (right)

Table 3. Characteristics of SVR models with additional information included. SVR^{loo} added for the convenience of the comparison

	Training RMSE	Normalized Number of SV	Validation RMSE	Correlation coefficient
SVR^{ε} (ε=0.02, σ=3, C=0.5)	0.033	0.62	0.060	0.885
SVR^{C} (ε=0.02, σ=3, C=0.3)	0.030	0.61	0.061	0.880
$SVR^{C\varepsilon}$ (ε=0.02, σ=3, C=0.5)	0.029	0.63	**0.058**	**0.891**
SVR^{loo}	0.048	0.53	0.064	0.86

6 Multi-scale SVR modeling

The models considered before are single-scaled. The solution given by Eq. 19 is a linear combination of kernel functions. We've used isotropic RBF kernel functions with one parameter – kernel bandwidth. As it was discussed above, this parameter is common for all the training points and can't be varied in space. There is a SVR-type model following from Vicinal Risk Minimization learning principle that allows to use individual bandwidths for every training point [14]. But this model is not well studied yet to be applied to environmental data.

Here we consider another extension of the "standard" SVR model – a multi-kernel SVR [13]. This SVR-type model uses the so-called dictionary representation of the regression function. Several types of kernels (or several kernels of the same type with different parameters) are used in this model. The main idea of the adaptation of this model to environmental data is to use Gaussian RBF kernels with different bandwidths to model different correlation scales in complex non-stationary multi-scale data.

6.1 Multi-kernel SVR

Instead of ordinary SVM where we seek for the regression function in the form

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

we'll seek the regression function in a form

$$f(x,\alpha) = \sum_{i=1}^{N} \alpha_i^{\,1} K_1(x, x_i) + \alpha_i^{\,2} K_2(x, x_i) + ... + \alpha_i^{\,k} K_k(x, x_i) \; + b \tag{20}$$

where we denote α_i^p the weight corresponding to *i*-th training point and *p*-th kernel. We'll construct a SVR-type algorithm which would tune the α parameters. This is achieved by using the support-vector type regularizer with an ε-insensitive empirical risk term weighted by a trade-off constant C. The resulting optimization problem is a QP-problem what guarantees the uniqueness of the solution. This means that the α coefficients in (20) are uniquely determined automatically from the solution. In the case of several RBF kernel functions with different bandwidths it means that the model is a multi-scale decomposition of the solution.

6.2 Multi-kernel LP-SVR

The LP-SVR model is an analogue of the ordinary SVR that uses the linear regularization term hence gives a Linear Programming problem as the final optimization problem. We use standard linear ε-insensitive loss function. The regularizer for LP-SVM's is as follows:

$$Q_{LP}(\alpha) = \sum_{p=1}^{k} \sum_{i=1}^{N} (\alpha_i^p + \hat{\alpha}_i^p) \tag{21}$$

where summation by *i* corresponds to training data and summation by *p* corresponds to kernels.

The resulting optimization problem is a Linear Programming problem:

$$\begin{cases} \min \;\; Q_{LP}(\alpha) + C\sum_{i=1}^{N} (\xi_i + \xi^*_i) \;\; \text{subject to} \\ y_i - \varepsilon - \xi_i \le \sum_{i=1}^{N} \sum_{p=1}^{k} \; (\hat{\alpha}^p{}_i - \alpha^p{}_i) K_p(x_i, x_j) + b \le y_i + \varepsilon + \xi_i^*, \; \text{i = 1,...N} \\ \hat{\alpha}_i^p \ge 0, \quad \alpha_i^p \ge 0, \qquad \xi_i^* \ge 0, \quad \xi_i \ge 0 \end{cases} \tag{22}$$

The number of variables in the optimization problem is 2*N*K+2*N, where N is a number of the training points and K is a number of kernels. The more kernels we use the larger is the problem size and the computational time. LP problem can be solved faster then QP, and that is why LP-SVRs are used here for multi-scale modeling. LP-SVR also possesses the main properties of ordinary QP-SVR including sparseness of the solution for non-zero ε.

Two-scale RBF functions will be used below in the case study:

$$f(x,\alpha) = \sum_{i=1}^{N} \alpha_i^1 e^{-\frac{(x-x_i)}{2\sigma_1^2}} + \alpha_i^2 e^{-\frac{(x-x_i)}{2\sigma_2^2}} + b \tag{23}$$

6.3 Case study: two-scale model

Consider the same Cs137 data as we've modeled before with the ordinary single-scale SVR. Figure 11 presents the testing error surface, kernel bandwidths σ_1 and σ_2 are plotted along the axis. It is obvious that the resulting error surface is reflection symmetric with respect to σ_1=σ_2 line. One can observe the minima of the testing error at the values σ_1 = 1 and σ_2 = 5 (or σ_1 = 5 and σ_2 = 1). Let us remind that the observed optimal kernel bandwidth in a single-scale SVR model was σ = 3 – some average value in comparison to two-scale SVR.

Validation error of this two-scale model is 0.058, what is better than single-scale SVR and is the same as for complicated model with incorporated additional information (see Table 3).

The results of the prediction mapping of the two-scale SVR are presented in the figure 12. Notice that the spot of the high activity on the west is mainly modeled by the "short-scale" part of the model, while the spots in the center of the area are modeled with "long-scale" part. The drawback of the single-scaled model is that it uses some averaged spatial scale.

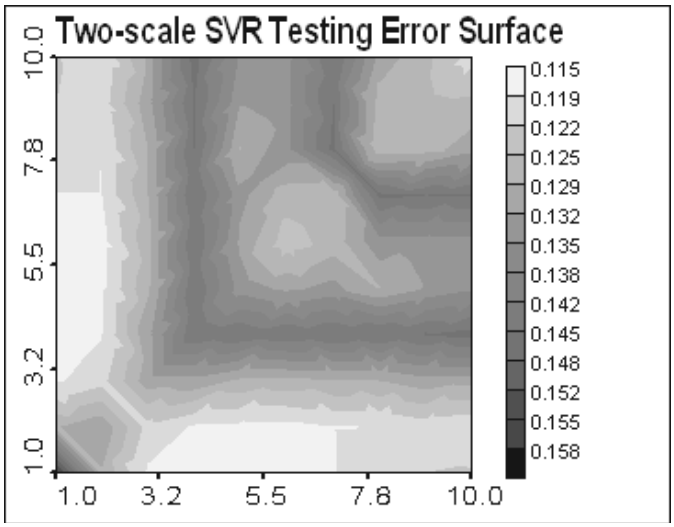

Figure 11. Two-scale SVR testing error surface. X, Y axis: kernel bandwidths

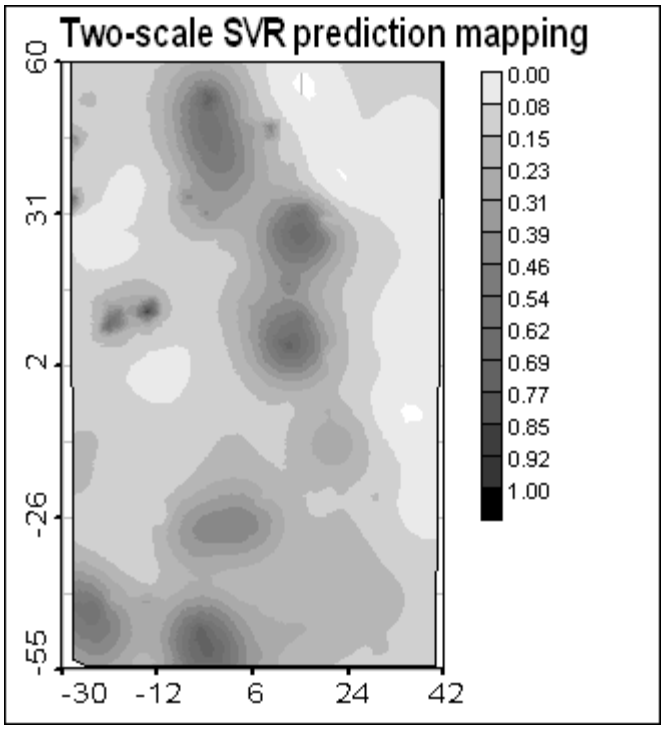

Figure 12. Two-Scale SVR prediction mapping

7 Conclusions

Support Vector Regression - a model based on the Statistical Learning Theory - was considered. The methodology of application of the SVR models to the problem of spatial data prediction mapping was considered on a real case study. Parameters' influence on the solution and procedure of selecting appropriate values of all the parameters was considered in details.

It was shown that flexibility of SVR makes it an attractive mapping tool and allows obtaining promising results even when data set is noisy or contains outliers.

Advanced SVR modeling in presence of additional information on measurements quality was considered. It was shown how to incorporate this additional information into SVR model. Advanced model outperformed "standard" SVR in the considered real case study.

Multi-scale SVR model was considered. This model adapts to non-stationary data containing spatial structures of different scales. Preliminary results on the multi-scale SVR modeling showed that two-scale model outperformed simple single-scale SVR in the considered case study.

Further research can be devoted to finding closer theoretical links between geostatistical models and SVR (both single and multi-scaled).

8 Acknowledgements

The work was partly supported by INTAS grant 99-00099 and CRDF grant RG2-2236.

9 References

1. Cherkassky V. and F. Mulier. Learning from data. New York: John Wiley Interscience, 1998.
2. Cristianini N. and J. Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge: Cambridge University Press, 2000.
3. Deutsch C. V. and A. G. Journel. GSLIB. Geostatistical Software Library and User's Guide, New York: Oxford University Press, 1997.
4. Hastie T., Tibshirani R., and Friedman J. The elements of Statistical Learning. New York: Springer, 2001.
5. Kanevski M., Pozdnukhov A., Canu S., Maignan M., Wong P. M., Shibli S. Support Vector Machines for Classification and Mapping of Reservoir Data. In "Soft Computing for Reservoir Characterization and Modeling". Wong P., Aminzadeh F., and Nikravesh M. (Eds.), Heidelberg: Springer-Verlag, pp. 531-558, 2002.
6. Kanevski M., Pozdnukhov A., Canu S., Maignan M. Advanced Spatial Data Analysis and Modeling with Support Vector Machines, International Journal of Fuzzy Systems, Vol. 4, No. 1, pp. 606-616, March 2002.
7. Kanevski M., N. Gilardi, M. Maignan, E. Mayoraz, "Environmental Spatial Data Classification with Support Vector Machines," IDIAP Research Report. IDIAP-RR-99-07, pp. 24, www.idiap.ch, 1999.
8. Kanevski M., R. Arutyunyan, L. Bolshov, V. Demyanov, M. Maignan. "Artificial neural networks and spatial estimations of Chernobyl fallout". Geoinformatics, Vol. 7, No.1, pp. 5-11, 1996.
9. Kanevski M., S. Canu. "Spatial Data Mapping with Support Vector Regression," IDIAP Research Report, RR-00-09, www.idiap.ch, 2000.
10. Kanevski M, V. Demyanov, S. Chernov, et al. "Geostat Office for Environmental and Pollution Spatial Data Analysis". Mathematische Geologie, N3, April 1999, pp. 73-83.
11. Huber P. Robust Estimation of Location Parameter, Annals of Mathematical Statistics, 35(1), 1964.
12. Smola A. Regression Estimation with Support Vector Learning Machines, 1996.
13. Weston J. Extensions to the Support Vector Method. Ph.D. Thesis, 1999.
14. Vapnik V. Statistical Learning Theory. New York: John Wiley & Sons, 1998.