Geoderma 154 (2010) 340–347

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Soil type classification and estimation of soil properties using support vector machines

Miloš Kovačević [a,1], Branislav Bajat [a,*], Boško Gajić [b,2]

[a] Faculty of Civil Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia
[b] Faculty of Agriculture, University of Belgrade, Nemanjina 6, 11080 Zemun, Serbia

ARTICLE INFO

Article history:
Received 1 April 2009
Received in revised form 1 November 2009
Accepted 5 November 2009
Available online 4 December 2009

Keywords:
Support vector machines
Classification
Regression
Soil types
Chemical properties
Physical properties

ABSTRACT

Quantitative techniques for prediction and classification in soil survey are developing rapidly. The paper introduces application of Support Vector Machines in the estimate of values of soil properties and soil type classification based on known values of particular chemical and physical properties in sampled profiles. Comparison of proposed approach with other linear regression models shows that Support Vector Machines are the model of choice for estimation of values of physical properties and pH value when using only chemical data inputs. They are also the model of choice in the cases where chemical data inputs are not strongly correlated to the estimated property. However, in classification task, their performance is similar to that of the other compared methods, with an increasing advantage when a data set consists of a small number of training samples per each soil type.

1. Introduction

Measuring and assessment of soil components and properties is generally a time-consuming and costly procedure. Lack of sampling data is often compensated by results of predictions or modelling. Various modelling procedures, known as predictive soil mapping, are specially developed to estimate spatial distribution of soil variables. Most of them are based on numerical or statistical models of relationship among other environmental variables and soil properties, applied to geographic databases in order to create a predictive map or to derive the values of soil properties at unmeasured sites from field-collected data (Burrough et al., 1997; Scull et al., 2003). Spatial modelling methods based on spatial variability and correlation between different soil and environmental properties are referred to as geostatistical methods (Goovaerts, 1999). The current progress in data gathering technologies and computational resources have provided geostatistician with the large amounts of information of different types, which can be stored, processed and visualised in Geographic Information Systems (McBratney et al., 2003).

The other approach to the estimate of soil variables is more focused on direct estimation of values of unknown soil parameters, based on measured or available values of some other parameters. The most comprehensive research in this field is embedded in developing Pedotransfer Functions (PTF) (Bouma, 1989), which estimate hydraulic soil parameters based on already known soil properties. PTFs are not restricted to hydraulic properties, they are also developed for estimating soil physical, mechanical, chemical and biological properties by following up effort principle *not predict something that is easier to measure than the predictor* (McBratney et al., 2002).

The development of fast and cheap microprocessors has resulted in a growing usage of sophisticated statistical and machine learning methods, such as Artificial Neural Networks (ANN) or Support Vector Machines (SVM), in a wide variety of environmental sciences. ANNs are used for PTFs by Pachepsky et al., 1996. SVM has shown considerable advantage over ANN in PTF development, especially in soil matric potentials (Lamorski et al., 2008). SVM has confirmed the viability of modelling of complex relationships between seismic and soil parameters to solve classification-related problems with assessment of soil liquefaction potential (Goh and Goh, 2007).

There are many examples of supervised learning methods used in agriculture, especially in precision agriculture. The classification task in distinguishing between weeds and crops by their spectral proprieties, for the purpose of precise herbicide use, could reduce the input costs and also mitigate the environmental impact. Moshou et al. (2001) compared the efficiency of several neural network techniques in classifying different kinds of crops vs. weeds. A study prepared by Karimi et al. (2006) demonstrated the capability of SVM method, analyzing hyperspectral data for identification of weed and nitrogen stresses in early growth stage of a cornfield. Machine learning methods could also be used to predict the quantity forward

* Corresponding author. Tel.: +381 11 3218 579; fax: +381 11 3370 223.
E-mail addresses: milos@grf.bg.ac.rs (M. Kovačević), bajat@grf.bg.ac.rs (B. Bajat), bonna@agrif.bg.ac.rs (B. Gajić).
[1] Fax: +381 11 3370 223.
[2] Fax: +381 11 2193 659.

0016-7061/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2009.11.005

in time, based on training sets which use the past data (Gill et al., 2006).

Many applications in environmental studies use SVM in conjunction with some other contemporary techniques, for instance cellular automata in spatial simulation of land use changes (Yang et al., 2008) or fuzzy *k*-means in image pattern recognition in precise farming (Tellaeche et al., 2007). Most machine learning methods used in bioinformatics and environmental sciences have been proven supportive of contemporary acquisition technologies such like remote sensing (Nemmour and Chibani, 2006; Zhai et al., 2006). A distinctive characteristic of all such applications is a huge number of observations. This research is focused on a "small" data set and feasibility of using the SVM approach in related environment.

Determination of most physical and chemical soil properties needs laborious and time-consuming laboratory tests. For those reasons, it is economically justified to develop methods which are capable of estimating some of them based on knowledge of other, already identified properties. When assessing the soil properties, one generally assumes that input data need to be split into more homogeneous soil groups. Seybold et al. (2005) attempted to estimate the cation exchange capacity from organic C content, clay and silt content, and soil pH, by linear regression models. The very first step in their research was to stratify all data into exact soil type groups according to certain characteristics. The model parameters obtained after stratification are related to the soil types division. In order to improve both accuracy and reliability of PTFs prediction, Pachepsky and Rawls (1999) suggested soil grouping according to taxonomic unit, moisture and temperature regime, and soil textural classes. Therefore, the objective of this research is to find the appropriate regression model for such estimation, regardless of the soil type or class. The possibility to classify soil samples according to soil types, based on 7 chemical and 3 physical properties of samples obtained in laboratory analyses, shall also be investigated. The comparison of the Support Vector Regression and classification approach with the other commonly used techniques is presented in the experimental section.

The structure of the paper is as follows: Section 2 deals with formulation of the problem of soil type classification and estimation of values of chemical and physical soil properties, using a machine learning framework. It contains a brief description of theoretical foundations of SVM approach in classification and regression. Section 3 provides details of experiments performed on different soil samples taken from a case study area in eastern Serbia. It contains a description of testing methodology and analysis of the experiment results. Section 4 concludes the paper.

2. Material and methods

2.1. Problem statement

Let S be the set of all the possible soil samples covering a geographical area, presented in the following form: $S=\{x|x\in R^n\}$. Each soil sample is represented as an n-dimensional real vector and every particular coordinate x_i represents a value of some measured property such as available K_2O or P_2O_5, which were determined by extraction with 0.1 M ammoniumlactate and 0.4 M acetic acid (Egner et al., 1960). Let $C=\{c_1,c_2,...,c_l\}$ be the set of l classes that correspond to some predefined soil types such as Chernozems or Leptosols. The function $f_c:S\rightarrow C$ is called a classification if for each $x_i\in S$ it holds that $f_c(\mathbf{x}_i)=c_j$ if $\mathbf{x}_i$ belongs to the class c_j. In practice, one only has a limited set of m labeled *examples* (x_i,y_i), $\mathbf{x}_i\in R^n$, $y_i\in C$, $i=1,...,m$. Dimension n is equal to the number of properties used to describe each example (soil sample). Labeled examples form the training set for the classification problem at hand. The machine learning approach tries to find the function $\tilde{f}_c$, which is a good approximation of the real, unknown function f_c, using only the examples from the training set and a specific learning method such as ANN or Decision Trees (DT) (Mitchell, 1997).

Another relevant task is to estimate the value of an unknown property for particular soil sample using any other available properties known in advance. Suppose one is given a training set with m soil samples (x_i,y_i), $\mathbf{x}_i\in R^n$, $y_i\in R$, $i=1,...,m$, where y_i is the known real value of the target property one tries to estimate for samples in S not contained in the training set. The function $f_r:S\rightarrow R$ is called a *regression* if it estimates the value of the target property given the values of other properties for any random sample $x\in S$. As in the classification task, the machine learning approach is focused on finding the function $\tilde{f}_r$, which is a good approximation of the unknown function f_r, using the training examples and a specific learning method such as ANN or DT.

The issue addressed in this paper is finding the most accurate possible classification $\tilde{f}_c$ and regression $\tilde{f}_r$ from training examples, using the learning method called Support Vector Machines (Vapnik, 1995). In order to test the quality of the proposed method, we compared SVM (linear and Gaussian kernel) to the other commonly used linear classification and regression algorithms. Linear methods were chosen for reference because of the small number of soil properties used for representation of a sample and satisfactory results provided by these methods, for instance in CEC estimation using linear regression (Seilsepour and Rashidi, 2008; Seybold et al., 2005). The linear methods we tested in the classification task are: Logistic Regression (LR) (Hosmer and Lemeshow, 2000) and Multinomial Naïve Bayes (MNB) (Lewis, 1998). In the regression task, linear and Gaussian SVR are compared to Ordinary Least Squares (OLS) and Ridge Linear Regression (RLR) (Hoerl and Kennard, 1970).

In this research, each soil sample is represented by measured values of 7 chemical and 3 physical properties on top soil (depth of 0–30 cm), as described in Table 1. Soil types used in the paper (Table 2) are classified according to the World Reference Base (WRB) for Soil Resources (FAO, 2006). Methods for appraisal of quality of the proposed classification and regression tasks are addressed in Section 3.

2.2. Brief theory of support vector machines

2.2.1. SVM classification

Support Vector Machines method is a recent approach in pattern classification and it deals with the *binary classification model* (Vapnik, 1995). The binary model assumes that a soil sample belongs to one class only, and that there are just two classes ($C=\{c_1,c_2\}$). Each classification task with n classes can be modelled as a sequence of $\binom{n}{2}$ binary tasks using the *one-vs-one* approach in which one trains $n*(n-1)/2$ binary classifiers, one for each pair of classes. The final decision is made by voting, i.e. the most frequently predicted class is selected as the output. Let $(\mathbf{x}_i,y_i)$, $\mathbf{x}_i\in\boldsymbol{R}^n$, $y_i\in\{-1,1\}$, $i=1,...,m$ be the training set (-1 stands for class c_1 and 1 for c_2). Fig. 1 is used to illustrate the basic idea of SVM classification. White and grey squares represent samples from a training set comprised of two distinct classes.

Let us assume for a moment that classes are linearly separable, and neglect the circled examples in Fig. 1. During the learning phase, one seeks the separating hyper-plane which best separates the examples of two classes. Let $h_1:\mathbf{w}\cdot\mathbf{x}+b=1$ (where "·" denotes the dot-product) and h_{-1}: $\mathbf{w}\cdot\mathbf{x}+b=-1$, $\mathbf{w},\mathbf{x}\in R^n, b\in R$, be possible hyper-planes, with all the white examples lying above h_1 ($y_i=1$) and all the grey examples lying below h_{-1} ($y_i=-1$). Hence for all training examples $(\mathbf{x}_i, y_i)$ it follows that:

$$y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1,\ i=1,2,...m \tag{1}$$

One chooses h: $\mathbf{w}\cdot\mathbf{x}+b=0$ to be the best separating hyper-plane lying in the middle between the already fixed hyper-planes h_1 and h_{-1}. The notion of the best separation can be formulated to find the maximum margin M that separates the data from both classes. Since

Table 1
Chemical and physical properties describing the soil samples.

Property (units)	Description	Chemical/physical	Min	Max	Mean
SOM %	Soil organic matter.	C	1.02	11.43	2.93
pH	Soil pH (1:2.5 suspension of soil in distilled water).	C	5.0	8.26	6.33
Total N, %	Total soil nitrogen	C	0.03	0.535	0.141
K_2O, mg/100 g soil	Available soil potassium	C	1.1	41.0	17.28
P_2O_5, mg/100 g soil	Available soil phosphorus	C	0.9	42.0	4.96
S, $cmol^+$/kg soil	Exchangeable bases	C	4.9	60.2	20.91
CEC, $cmol^+$/kg soil	Cation exchange capacity	C	10.5	61.6	27.79
Sand, % (2–0.05 mm)	Soil particles between 0.05 mm and 2 mm in equivalent diameter.	P	12.4	72.4	39.02
Clay, % (<0.002 mm)	Soil particle smaller than 0.002 mm equivalent diameter.	P	10.3	51.92	31.02
Physical sand, %	Soil particle size >0.01 mm equivalent diameter.	P	34.0	84.0	58.24

the margin is equal to $\frac{2}{||w||}$, maximizing the margin is equal to minimizing the $||\mathbf{w}||$. The best separating hyper-plane can now be found by solving the following non-linear convex programming problem (for solving of optimization problem refer to Fletcher, 1987): find $\mathbf{w}$, b so that

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{w.r.t} : 1 - y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \le 0,\ i = 1,2,\ldots m \tag{2}$$

In practical classification problems, examples are usually not linearly separable (circled examples from Fig. 1). Therefore, some additional positive slack variables ξ_i are introduced, representing the distance of points on the wrong side of the separating hyper-plane (circled squares). The non-linear convex program (2) now becomes:

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_i \xi_i \quad \text{w.r.t} : 1-\xi_i - y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \le 0,\ -\xi_i \le 0,\ i = 1,2,\ldots m \tag{3}$$

The parameter C models the penalty for misclassified points in a training set. One wants to find a hyper-plane to minimize misclassification errors while maximizing the margin between classes. The optimization problem (3) is usually solved in its dual form and the solution is:

$$\mathbf{w}^* = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i,\ C \ge \alpha_i \ge 0,\ i = 1,\ldots m \tag{4}$$

Herewith the solution $\mathbf{w}^*$ for optimal hyper-plane is a linear combination of training examples. However, it can be shown that $\mathbf{w}^*$ represents a linear combination of those vectors $\mathbf{x}_i$ (support vectors) for which the corresponding α_i is a non-zero value. Support vectors for which $C > \alpha_i > 0$ holds belong either to h_1 or h_{-1} (depending on y_i). Let $\mathbf{x}_a$ and $\mathbf{x}_b$ be two support vectors ($C > \alpha_a, \alpha_b > 0$) for which holds $y_a = 1$ and $y_b = -1$. Now $b^* = -\frac{1}{2}\mathbf{w}^*\cdot(\mathbf{x}_a + \mathbf{x}_b)$ and finally the classification function becomes:

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i\cdot\mathbf{x}) + b^*\right) \tag{5}$$

In order to deal with non-linearity of the classification problem, the SVM approach goes one step further. One can define mapping of examples to a so-called feature space of very high dimensions: ϕ: $R^n \to R^d$, $n << d$ i.e. $\mathbf{x} \to \phi(\mathbf{x})$. The basic idea of this mapping into high-dimensional space is to transform the non-linear case into the linear one, as illustrated in Fig. 2, and then to use the above explained linear algorithm. In such a space, the dot-product from Eq. (5) transforms into $\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x})$. It is recognized that there exists a class of functions called kernels (Burges, 1998) for which holds $k(\mathbf{x}_i,\mathbf{x}) = \phi(\mathbf{x}_i)\cdot\phi(\mathbf{x})$. These functions represent the dot-products in some high-dimensional spaces, but can be easily computed in the input space. Using kernels equation, Eq. (5) becomes:

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i k(\mathbf{x}_i,\mathbf{x}) + b^*\right) \tag{6}$$

In this paper, Gaussian $k(\mathbf{x}_i,\mathbf{x}) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}\|^2)$ and linear kernel [no mapping, Eq. (5)] are used. After removing of all training data which are not the support vector points but retraining the classifier, the same solution is reached once again. Hence, support vectors represent examples from the training set that best describe the classes. The ability to distinguish between the support vectors and the noisy data points enables SVM to increase its generalization capacity in the learning process. For detailed review of SVM for pattern classification please refer to Burges (1998).

Table 2
Soil types and their respective parent materials and class labels.

Soil type		Class label
WRB soil group	Parent materials	
Rendzic leptosols	Loess-like loams and/or marls	1
Calcic chernozems	Calcareous silty colluvial deposits from sandstones or sandy limestones	2
Arenic chernozems	Slightly calcareous sands	3
Humic leptosols	Non-calcareous colluvial deposits over compact clay sediments	4
Eutric cambisols	Calcareous loams and/or calcareous sandstones	5
Haplic vertisols	Marly clays	6
Siltic chernozems	Highly calcareous, weakly cemented, silty loams	7
Haplic luvisols	Non-calcareous or slightly calcareous loamy lacustrine and colluvial deposits	8

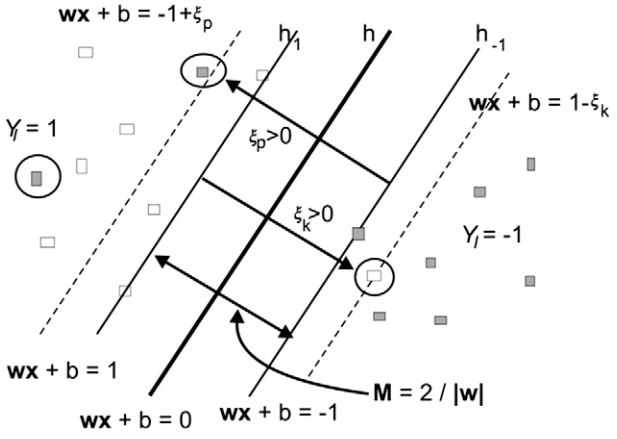

Fig. 1. SVM used for classification: construction of separation hyper-plane in two-dimensional case (hyper-plane is a line).

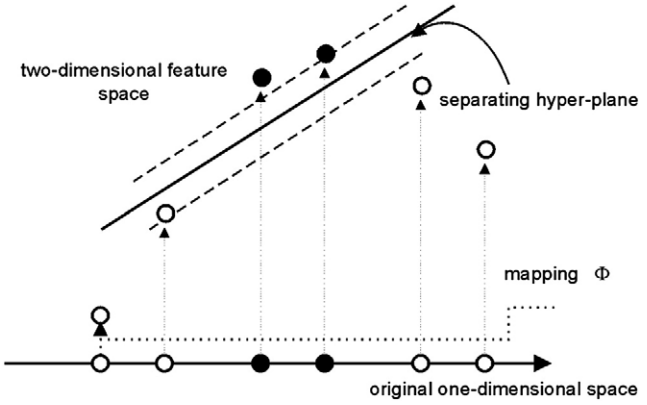

Fig. 2. Mapping examples (here one-dimensional) into high-dimensional space (two-dimensional).

2.2.2. Support Vector Regression (SVR)

Let $(\mathbf{x}_i,y_i)$, $\mathbf{x}_i\in R^n$, $y_i\in R$, $i=1,\ldots,m$ be the training set where y_i denotes the target property of an already known i-th example. One tries to estimate the target values in points other than the training set. Fig. 3 is used to illustrate the basic idea of SVR. Linear case is again the starting point. During the learning phase, one tries to find the linear function $f(\mathbf{x})=\mathbf{w}\cdot\mathbf{x}+b$, $\mathbf{w}$, $\mathbf{x}\in R^n$, $b\in R$ for which the difference between the actual training value y_i and estimated value $f(\mathbf{x}_i)$ would be at most ε or $|y_i-f(\mathbf{x}_i)|<\varepsilon$. The other criterion is as flat as possible function f, which is related to the small norm of the vector **w.**

If one should tolerate all the points which are out of the shaded region for some small positive values ξ_i then the following convex optimization program could be formulated:

$$\begin{aligned}&\min_{w,b}\ \frac{1}{2}\|w\|^2+C\sum_i(\xi_i+\xi_i^*)\\&\text{w.r.t}\quad y_i-(w\cdot x_i+b)\le\varepsilon+\xi_i,\\&:\qquad (w\cdot x_i+b)-y_i\le\varepsilon+\xi_i^*,\\&\qquad -\xi_i,-\xi_i^*\le 0,\ i=1,2,\ldots m\end{aligned}\tag{7}$$

Using similar approach as in the classification setting, one can solve Eq. (7) using the dual program and after introducing the kernel function in order to achieve the non-linearity, the solution is given in the following form:

$$f(\mathbf{x})=\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)k(\mathbf{x}_i,\mathbf{x})+b^*\ \text{where}\ C\ge\alpha_i,\alpha_i^*\ge 0,\quad i=1,\ldots m\tag{8}$$

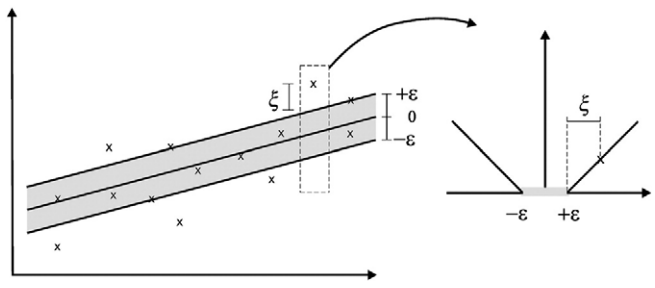

Fig. 3. ε-insensitive loss function for SVR (right). Points in shaded region are assumed to be exactly estimated (left).

A good tutorial on SVR can be found in Smola and Schölkopf 2004.

2.3. Dataset

The Institute of Soil and Melioration at the Faculty of Agriculture, University of Belgrade, carried out a detailed survey in 1981, in order to produce a soil map intended for detailed design of drainage and irrigation networks, as well as for establishing new vineyards in the Negotin municipality in eastern Serbia, close to the Romanian and Bulgarian borders (N 44°12′10″, E 22°28′17″). Mapping of soil units have been based on field and laboratory observations of chemical, physical, biological, and mineralogical properties of horizons, geological properties of parent rock material, and geomorphological features of landscape (Soil Survey Staff, 1951). The total area covered by the survey approximates 1272 ha, with altitudes between 106 m and 290 m above the sea level i.e. morphologically it belongs to a moderate hilly area. The land use predominantly consists of croplands (45%) and grasslands (meadows) (34%) (Table 3) subdivided into numerous individual lots (average area 1 ha).

The parent material is composed solely of sediments or sedimentary rocks. These parent materials are considerably different in terms of particle size, levels of coherence, calcareousness and porosity (Table 2). These soils are usually characterised by eutric properties, with sand loamy, sand clay loamy, loamy, clay loamy or clay texture. The most representative soil types within the study area belong to the WRB groups of Leptosols, Chernozems, Cambisols, Vertisols and Luvisols. The covered terrain is characterised by the presence and mixture of various soil types and turning of one soil type into another at short intervals.

Polygons of soil type classes were delineated based on the topographic map at 1:25,000 scale. The total number of measured soil samples was 151 or one sample per each 8.5 ha (Fig. 4). The areas marked grey in Fig. 4 represent the terrain with specific geomorphological features, such as ravines, water retentions or urban areas, where the soil was not mapped.

Physical properties, i.e. particle size distribution, cation exchange capacity, exchangeable bases, pH in water (1:2.5), and total nitrogen (N), were determined following the standard methodologies (Van Reeuwijk, 1995). Soil organic matter was estimated by the method of Tiurin (Bogdanović et al., 1966). Available P and K values were determined by AL method (Egner et al., 1960). Distribution of measured samples among the soil types is presented in Table 3.

Table 4 (correlation matrix) contains Pearson correlation coefficients between the soil properties calculated from the available data set. Many chemical soil properties are significantly correlated, especially SOM and N, and S and CEC. The same stands for Send and Clay, Send and Physical Send, and Clay and Physical Send in the group of physical properties. The correlation matrix will be used to explain certain findings related to the regression task in Section 3.3.

Table 3
Land use and the distribution of measured samples among the soil types.

Soil type (WRB)	Land use				No. of sample profiles (%)
	Forest	Grassland	Cropland	Orchards and vineyards	
Rendzic leptosols	20.0%	16.0%	52.0%	12.0%	25 (16.5%)
Calcic chernozems	–	33.3%	33.3%	33.3%	6 (4.0%)
Arenic chernozems	–	35.3%	41.2%	23.5%	17 (11.3%)
Humic leptosols	–	–	–	100%	2 (1.3%)
Eutric cambisols	10%	50%	30%	10%	10 (6.6%)
Haplic vertisols	4.6%	60.5%	30.2%	4.7%	43 (28.5%)
Siltic chernozems	–	14.3%	78.6%	7.1%	14 (9.3%)
Haplic luvisols	8.8%	20.6%	55.9%	14.7%	34 (22.5%)
Total number	7.3%	34.4%	45.0%	13.3%	151 (100%)

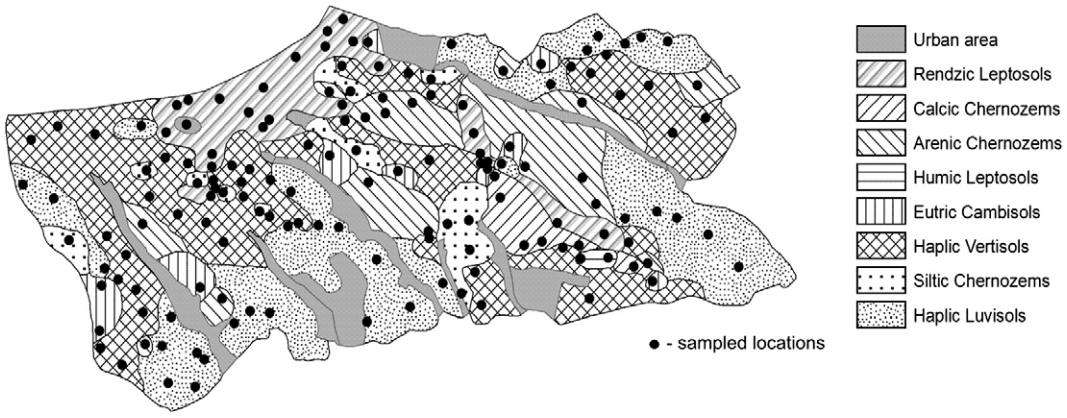

Fig. 4. Map of soil types with locations of sampled profiles.

Table 4
Pearson correlation coefficients among chemical and physical properties (correlation matrix is symmetric).

Target property	SOM	pH	N	K_2O	P_2O_5	S	CEC	Sand	Clay	Phy. sand
SOM	1.000	0.567	0.987	0.141	0.375	0.663	0.675	−0.469	0.531	−0.533
pH		1.000	0.570	0.264	0.406	0.860	0.813	−0.273	0.352	−0.345
N			1.000	0.158	0.378	0.656	0.667	−0.449	0.506	−0.510
K_2O				1.000	0.485	0.132	0.110	−0.132	0.214	−0.172
P_2O_5					1.000	0.434	0.427	−0.114	0.118	−0.119
S						1.000	0.995	−0.403	0.414	−0.434
CEC							1.000	−0.440	0.440	−0.466
Sand								1.000	−0.903	0.942
Clay									1.000	−0.984
Phy. sand										1.000

Shaded cells represent exceptionally strong correlations between related chemical properties. These properties will be removed in some regression experiments dealing with the estimation of SOM, N, S and CEC.

3. Results and comments

3.1. Performance measures and testing protocol

The following metrics were used to measure the performance of the classification task: *micro averaged F_1* measure (F_1) (van Rijsbergen, 1979; Sebastiani, 2002) and *kappa* statistics (κ) (Cohen, 1960). The performance of the regression task was measured using *normalized root mean squared deviation* (NRMSD) and *coefficient of determination* (R^2).

Classification methods were compared using the *leave one out cross-validation* protocol, since our small data set contains certain number of classes with few examples. In order to measure the effects of the number of training examples on a classification performance, we conducted a series of *holdout* experiments with different percentage of training examples in related train-test splits. The same proportion of training examples was preserved in each class. Each holdout experiment was repeated 100 times and the final κ and F_1 were obtained after averaging over the iterations. The comparison of regression models was conducted using 10-fold cross-validation repeated in 100 different iterations. Final estimates for NMRSD and R^2 were obtained after averaging over the iterations.

All experiments were conducted in WEKA machine learning software (Witten and Frank, 2005). When testing the SVM classification and regression algorithms, an open-source package LIBSVM (Chang and Lin, 2001) was used together with its standard interface to WEKA.

3.2. Classification experiments

The initial experiment included classification of soil samples into 8 original soil types using all chemical (7) and physical (3) properties as defined in Table 1. In order to ensure numerical stability of all the compared classification methods, the values of all parameters were normalized to be between 0 and 1. We trained 18 different classifiers divided into four groups: two LR with different Ridge parameter, one MNB, three linear SVM (LSVM) with $C \in \{1, 10, 100\}$ and twelve Gaussian SVM (GSVM) with $(C, \gamma) \in \{1, 10, 100\} \times \{0.1, 0.5, 1, 2\}$. The best-performing classifier from each group was selected, and the results are presented in the first row of Table 5.

LSVM and GSVM yield similar κ and F_1 values. Both SVM methods perform slightly better than LR concerning κ value, and are equal when comparing F_1. MNB is the worst performing classifier. The related confusion matrix is presented in Table 6. Obviously, all the methods except for MNB make similar misclassifications (MNB tends to classify towards bigger classes since the *a priori* class probabilities produce a significant impact in its model). In the next experiment we decided to remove the classes with few examples (Humic Leptosols 2, Calcic Chernozems 6, Eutric Cambisols 10) in order to reduce the negative impact of incomplete class description on overall classification results. Results presented in the second row of Table 5 showed the improvement for all methods, when compared to the initial

Table 5
The classification performance of Logistic Regression, Multinomial Naïve Bayes and SVM soil classifiers.

	LR (κ, F_1)	MNB (κ, F_1)	LSVM (κ, F_1)	GSVM (κ, F_1)
All classes	0.45; 0.55	0.36; 0.42	0.47; 0.55, $C=100$	0.46; 0.54, $C=10$, $\gamma=2$
Removed classes with ≤10 examples	0.53; 0.64	0.42; 0.51	0.53; 0.63, $C=100$	0.52; 0.63, $C=100$, $\gamma=1$

All experiments were performed using leave one out cross-validation, measured accuracy in Kappa statistics κ and F-measure F_1.

Table 6
Confusion matrices for all classes: Rendzic Leptosols (1), Calcic Chernozems (2), Arenic Chernozems (3), Humic Leptosols (4), Eutric Cambisols (5), Haplic Vertisols (6), Siltic Chernozems (7) and Haplic Luvisols (8).

Logistic regression									
True		1	2	3	4	5	6	7	8
Classified as	1	13	2	2	1	1	5	1	0
	2	3	2	1	0	0	0	0	0
	3	1	2	5	3	0	0	3	3
	4	0	0	0	1	0	1	0	0
	5	0	0	4	1	2	2	1	0
	6	3	0	1	1	1	30	5	2
	7	1	1	0	0	2	5	4	1
	8	0	0	2	1	1	4	0	26

Multinomial Naive Bayes									
True		1	2	3	4	5	6	7	8
Classified as	1	13	0	0	0	0	12	0	0
	2	3	0	0	0	0	2	0	1
	3	1	0	0	0	0	4	0	12
	4	0	0	0	0	0	1	0	1
	5	0	0	0	0	0	2	0	8
	6	0	0	0	0	0	38	0	5
	7	0	0	0	0	0	11	0	3
	8	0	0	0	0	0	7	0	27

Linear SVM									
True		1	2	3	4	5	6	7	8
Classified as	1	17	1	0	0	0	6	1	0
	2	4	0	2	0	0	0	0	0
	3	1	3	4	0	1	0	2	6
	4	0	0	1	0	0	1	0	0
	5	0	0	4	0	2	2	1	1
	6	2	0	1	0	0	34	2	4
	7	0	1	1	0	1	6	3	2
	8	0	0	3	0	0	4	0	27

Gaussian SVM									
True		1	2	3	4	5	6	7	8
Classified as	1	17	3	0	0	0	4	1	0
	2	2	2	2	0	0	0	0	0
	3	1	2	2	0	4	1	3	4
	4	1	0	0	0	1	0	0	0
	5	0	0	3	0	4	1	1	1
	6	2	0	0	0	0	31	7	3
	7	1	1	1	0	1	7	1	2
	8	0	0	1	0	0	5	0	28

setting. LSVM, GSVM and LR exhibited the same performance, which strongly suggested the linear nature of the classification problem at hand. Since LR method is less complex than L(G)SVM, one could conclude that it is the natural choice for the soil classification problem.

However, leave one out cross-validation tends to be an overoptimistic testing protocol. Hence, we wanted to test how the number of training examples influences the performance of the classification methods. For that purpose we performed a series of holdout experiments (see Section 3.1) on a reduced dataset (three smallest classes were removed) using LR, MNB and LSVM approaches. It is evident from Figs. 5 and 6 that LSVM performs the best over the whole range of training set sizes, while MNB performs the worst. Obviously, LSVM is a classification model of choice when dealing with smaller training set sizes. When 50% of samples (or more) were used for training the classifier (smallest class ≥ 7, largest class ≥ 21 examples), LR constantly moved towards LSVM matching its performance in the case when all data were used for the training (leave one out case).

3.3. Regression experiments

The following step of research was focused on application of SVR and its comparison with the proposed reference methods. First we tried to estimate the value of certain physical property from the already measured chemical characteristics of a soil sample. For that purpose the data set was transformed to have each sample represented by 7 chemical properties (Table 1). The physical characteristic to estimate was used as the target property. All coordinates were scaled to be between 0 and 1 except for the target values which remained in their original range. The obtained results are presented in Table 7.

Results in Table 7 suggest that linear models (OLS, RLR and LSVR) behave alike and are not appropriate for the regression task. It is not surprising since the chemical parameters are weakly correlated with the physical ones in our data set (see Table 4). However, it is possible to estimate the values of clay and physical sand reasonably well from the set of already known chemical parameters using GSVR. While not satisfactory, the performance measures for sand are significantly better in case of GSVR when compared to other linear models. This finding indicates to non-linearity of the regression task in such cases when target values are physical properties, and suggests that GSVR could be the method of choice.

In the next experiment we represented the soil samples using only chemical properties and tried to estimate the value of one particular property using the values of the others. This task makes sense in such situations when data about one property is lost or incomplete and some sort of reconstruction is required. Results from this experiment are presented in Table 8. All the compared methods performed

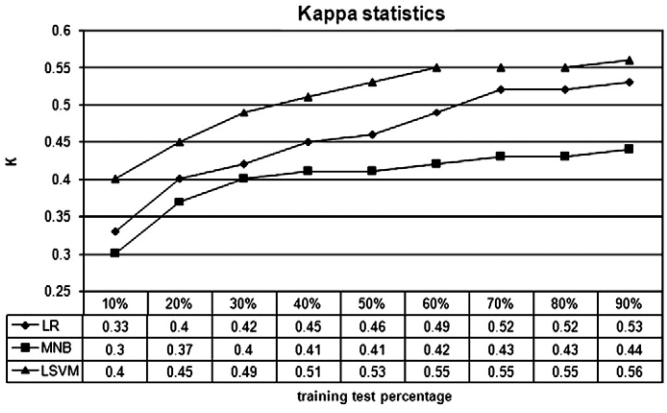

Fig. 5. Kappa statistics: LSVM is a classification model of choice when dealing with the small number of training data.

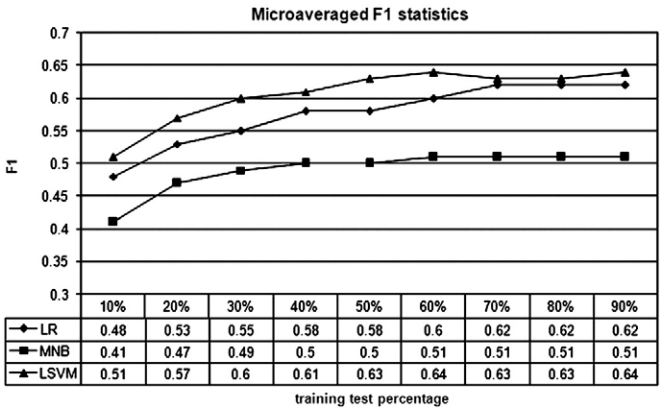

Fig. 6. Micro averaged F1 measure: LSVM is a classification model of choice when dealing with the small number of training data.

Table 7
Performances of certain physical property estimation based on the measured chemical characteristics of a soil sample.

	OLS	RLR	LSVM	GSVM
	R^2; NRMSD	R^2; NRMSD	R^2; NRMSD ($C=10$)	R^2; NRMSD ($C=10$, $\gamma=1$)
Clay	0.37, 0.19	0.35, 0.18	0.36, 0.18	0.76, 0.11
Sand	0.26, 0.21	0.28, 0.20	0.28, 0.20	0.59, 0.14
Physical sand	0.37, 0.19	0.35, 0.18	0.36, 0.18	0.76, 0.11

All experiments are performed using 100 times 10-fold cross-validation, squared correlation coefficient R^2 (degree of linearity between the real and the predicted value, higher the better) and the normalized root mean squared deviation NRMSD (lower the better).

Table 8
Performances of chemical property estimation based on values of others known chemical properties.

	OLS	RLR	LSVR	GSVR
	R^2, NRMSD	R^2, NRMSD	R^2, NRMSD	R^2, NRMSD
P_2O_5	0.14, 0.13	0.32, 0.11	0.31, 0.12 ($C=1$)	0.37, 0.11 ($C=1$, $\gamma=1$)
K_2O	0.19, 0.16	0.22, 0.15	0.24, 0.15 ($C=10$)	0.30, 0.15 ($C=1000$, $\gamma=0.1$)
pH	0.71, 0.13	0.61, 0.14	0.90, 0.07 ($C=100$)	0.94, 0.06 ($C=100$, $\gamma=0.5$)
N	0.96, 0.02	0.96, 0.02	0.85, 0.08 ($C=1$)	0.83, 0.08 ($C=10$, $\gamma=0.1$)
SOM	0.96, 0.02	0.96, 0.02	0.96, 0.02 ($C=10$)	0.98, 0.02 ($C=100$, $\gamma=0.1$)
S	0.98, 0.05	1, 0.03	1, 0.03 ($C=100$)	1, 0.02 ($C=1000$, $\gamma=0.1$)
CEC	0.98, 0.05	1, 0.03	1, 0.03 ($C=100$)	1, 0.02 ($C=1000$, $\gamma=0.1$)

All experiments are performed using 100 times 10-fold cross-validation; Shaded rows represent properties that we are able to predict. Dark-shaded row represents the case in which GSVR could be the model of choice.

excellently for SOM, S and CEC. Both SVR methods performed worse than linear models in case of N. This finding suggests that the method of choice for SOM, S and CEC is OLS (or RLR). OLS model is much simpler than SVR and is also interpretable (regression weights suggest the importance of particular properties).

Despite the presented results, in case of pH, both SVR models significantly outperform OLS and RLR (dark-shaded row in Table 8). However, GSVR shows a slightly better performance than LSVR. It is obvious that the performance of regression models for P_2O_5 and K_2O are not applicable, but are much better when using SVR with Gaussian kernel. Since the two properties are strongly related to the impact of applied agrotechnical measures, it is reasonable to expect that introduction of land use as an additional parameter might improve the regression results for all the compared methods.

The obtained results reflect the existence of very strong correlation between SOM and N, and S and CEC in our data set (see Table 4). However, target properties such as pH, K_20 and P_2O_5 are less correlated to other available chemical properties in our data set, when compared to SOM, N, S and CEC (see Table 4). We believe that existing correlations explain the results sufficiently well. In the final experiment an attempt is made to remove the highly correlated properties when predicting SOM, N, CEC and S (shaded cells in Table 4). Results presented in Table 9 suggest that GSVM clearly outperforms other methods and should be the model of choice when reconstructing missing chemical values from the set of other, less correlated chemical soil properties.

Table 9
Performances of chemical property estimation based on values of others, less correlated chemical properties (removed shaded properties from Table 4).

	OLS	RLR	LSVR	GSVR
	R^2, NRMSD	R^2, NRMSD	^{22}, NRMSD	R^2, NRMSD
N	0.50, 0.10	0.45, 0.11	0.41, 0.11 ($C=1$)	0.45, 0.11 ($C=1$, $\gamma=0.1$)
SOM	0.50, 0.10	0.45, 0.11	0.51, 0.10 ($C=100$)	0.59, 0.09 ($C=100$, $\gamma=0.1$)
S	0.71, 0.25	0.77, 0.22	0.77, 0.26 ($C=100$)	0.88, 0.14 ($C=100$, $\gamma=2$)
CEC	0.59, 0.28	0.69, 0.25	0.71, 0.28 ($C=100$)	0.86, 0.15 ($C=100$, $\gamma=2$)

All experiments are performed using 100 times 10-fold cross-validation; Shaded rows represent the case in which GSVR could be the model of choice.

4. Conclusions

Soil type classification based on samples with known values of particular chemical and physical soil properties, exhibits high degree of linearity. Classification models such as Logistic Regression or Linear Support Vector Machines could be used to automate this task with satisfactory accuracy. Logistic Regression could be a method of choice for data sets with enough training examples per each class. However, when the number of training examples per class is much smaller, Linear SVM approach has clear advantage over the other linear methods.

In the regression task, the obtained results suggest that linear methods are not able to estimate the values of physical properties using the already measured chemical properties. However, the non-linear SVR is able to estimate the value of clay and physical sand sufficiently well. This finding bears a practical effect, since the determination of physical properties of soil samples in laboratory is more laborious and time-consuming than the determination of chemical properties. It might be interesting for further research to take into consideration some additional physical properties such as bulk density, water retention characteristics, saturated hydraulic conductivity, and water stable aggregates.

Finally, it has been confirmed that one is able to reconstruct some of the chemical properties using the rest of them, except for K_2O and P_2O_5. The reconstruction could be done regardless of our knowledge about the soil type. When compared to other linear models, SVR did not show any practical advantage except in the regression model for pH, where both SVR variants significantly outperformed the other compared methods, with the Gaussian model being better than the linear one. The regression results which favour Ordinary Least Squares were not surprising, concerning the nature of existing correlations between the chemical soil properties in our data set. However, when the highly correlated data is missing (i.e. estimating S from other parameters, CEC is missing) Gaussian SVR clearly outperforms other linear methods.

The recently published paper (Ballabio, 2009) dealing with spatial aspects of relations between soil properties and environmental variables confirms the possibility of SVMs use in pedometrics. Such results naturally suggest that our future research should be directed to comparison of SVM approach with geostatistical techniques in soil mapping. We believe that these techniques could be applied in Geographical Information System environment for appropriate soil analyses and studies, and in other decision mapping and land use management tasks.

References

Ballabio, C., 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151 (3–4), 338–350.
Bouma, J., 1989. Using soil survey data for quantitative land evaluation. Advances in Soil Scienece 9, 177–213.
Bogdanović, M., Velikonja, N., Racz, Z., 1966. Manual for Soil Chemical Analysis. Yugoslav Society of Soil Science, Belgrade, p. 275 (in Serbian).
Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Kluwer Academic Publishers, Boston. 43 pp.
Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77, 115–135.
Chang, C.C., Lin, C.J., 2001. LIBSVM: a library for support vector machines. . Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1), 37–46.
Egner, H., Riehm, H., Domingo, W.R., 1960. Untersuchungen über die chemische Bodenanalyse als Grundlage für die Beurteilung des Nährstoffzustandes der Boden. II. Chemische Extraktionsmethoden zur Phosphor und Kaliumbestimmung. Kungliga Lantbrukshögskolans Annaler 26, 199–215.
FAO, 2006. World reference base for soil resources, by FAO-UNESCO-ISRIC. World Soil Resources Report No. 103, Rome.

Fletcher, R., 1987. Practical Methods of Optimization, 2nd ed. John Wiley, New York. 436 pp.
Gill, M.K., Mariush, T.A., Kemblowski, W., McKee, M., 2006. Soil moisture prediction using support vector machines. Journal of the American water resources association 42 (4), 1033–1046.
Goh, A.T.C., Goh, S.H., 2007. Support vector machines: their use in geotechnical engineering as illustrated using seismic liquefaction data. Computers and Geotechnics 34, 410–421.
Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and Perspectives. Geoderma 89, 1–45.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics 12 (3), 55–67.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd ed. Wiley, New York.
Karimi, Y., Prasher, S.O., Patel, R.M., Kim, S.H., 2006. Application of support vector machine technology for weed and nitrogen stress detection in corn. Computers and electronics in Agriculture 51, 99–109.
Lamorski, K., Pachepsky, Y., Stawiński, C., Walczak, R.T., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Science Society of America Journal 72, 1243–1247.
Lewis, D., 1998. Naive (bayes) at forty: The independence assumption in information retrieval. in ECML'98: Tenth European Conference On Machine Learning.
McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109, 41–73.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Mitchell, T.M., 1997. Machine Learning. McGraw Hill, New York. 414 pp.
Moshou, D., Vrindts, E., De Ketelaere, B., De Baerdemaeker, J., Ramon, H., 2001. A neural network based plant classifier. Computers and Electronics in Agriculture 31, 5–16.
Nemmour, H., Chibani, Y., 2006. Multiple support vector machines for land cover change detection: an application for mapping urban extensions. ISPRS Journal of Photogrammetry & Remote Sensing 61, 125–133.
Pachepsky, Y.A., Rawls, W.J., 1999. Accuracy and reliability of pedotransfer functions as affected by grouping soils. Soil Science Society of America Journal 63, 1748–1756.
Pachepsky, Y.A., Timlin, D., Várallyay, G., 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Science Society of America Journal 60, 727–773.
Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography 27 (2), 171–197.
Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Computing Surveys 34 (1), 1–47.
Seilsepour, M., Rashidi, M., 2008. Prediction of soil cation exchange capacity based on some soil physical and chemical properties. World Applied Sciences Journal 3, 200–205.
Seybold, C.A., Grossman, R.B., Reinsch, T.G., 2005. Predicting cation exchange capacity for soil survey using linear models. Soil Science Society of America Journal 69, 856–862.
Smola, A., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14 (3), 199–222.
Soil Survey Staff, 1951. Soil Survey Manual, vol. 18. U.S. Dept. Agriculture Handbook, Washington D.C.. 503 pp.
Tellaeche, A., Burgos-Artizzu, X.P., Pajares, G., Ribeiro, A., 2007. On combining support vector machines and fuzzy K-means in vision-based precision agriculture. Proceedings of World Academy of Science. Engineering and Technology 22, 33–38.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.
Van Reeuwijk, L.P., 1995. Procedures for soil analysis. Technical Paper, vol. 9. International Soil Reference and Informacion Centre, ISRIC-FAO, Wageningen.
van Rijsbergen, C.J., 1979. Information Retrieval, 2nd ed. Butterworths, London, UK.
Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, San Francisco.
Yang, Q., Lia, X., Shid, X., 2008. Cellular automata for simulating land use changes based on support vector machines. Computers & Geosciences 34, 592–602.
Zhai, Y., Thomasson, J.A., Boggess III, J.E., Sui, R., 2006. Soil texture classification with artificial neural networks operating on remote sensing data. Computers and Electronics in Agriculture 54, 53–68.