Alternative methods for determining water content in soil

MARTIN SUCHÁR, MILAN ČISTÝ

Slovak University of Technology Bratislava, Faculty of Civil Engineering, Radlinskeho 11, Bratislava 813 68, Slovak Republic

milan.cisty@stuba.sk

Abstract Soil moisture content is an important hydrologic component. The paper presents two data driven methods, which makes possibility to estimate water retention curves important for soil moisture forecasting for the soils of Záhorská lowland. These methods are based on supposed dependence of the water content on the percentage content of the Kopecký grain categories, and on the dry bulk density. Artificial neural networks (ANN) and support vector machines (SVM) were used to estimate the pedotransfer functions that can be applied for prediction of the drying branch of the water retention curve. The SVMs formulate a quadratic optimization problem that ensures a global optimum, which makes them superior to traditional learning algorithms such as multi-layer perceptron (MLP) type of neural network. Results from the SVM modelling are compared with predictions obtained from MLP models and shows that SVM models are performing better for soil moisture forecasting than MLP models.

Keywords soil water regime; pedotransfer function; neural networks; support vector machines

INTRODUCTION

Modelling water and solute transport in soil has become an important tool in simulating agricultural productivity as well as environmental quality. Over the last decades many studies have been devoted to the development of methods for estimating soil hydraulic parameters. In general, two categories of methods can be distinguished: (1) measurement techniques and (2) predictive methods (mathematical modelling). However, despite the progress that has been achieved, the measurement techniques remain time consuming and costly, especially when data are needed for large areas (Wösten et al., 2001). On the other hand, usage of the models depends on knowledge of the input data, which are needed for the numeric simulations. Some of this data (meteorological, climatic, hydrologic or crop characteristics) are usually available in competent institutions, but namely soil properties are available only for some parts of Slovakia (or elsewhere). These characteristics appear as key problem in the numerical simulation of soil water regime, mainly water retention curve (WRC) determination. The soil water retention curve describes the ability of a soil to store water at different suctions. Measurement of the water WRC points in the laboratory is very expensive, time consuming and labour intensive. During last ten years relatively many works appears which were devoted to the determination of WRCs from available soil properties as particle size distribution, dry bulk density, organic C etc., e.g. (Šútor, Štekauerová 1999), (Štekauerová, V., Skalová, J. 1999) in Slovak scientific literature. Pedotransfer functions (PTF) became term for such relationships between soil hydraulic parameters and the easier measurable properties usually available from soil survey (Bouma, 1989). Standard method for solving this task uses various types of regression analyses. Recently artificial neural networks (ANNs) became the tool of choice in PTF development, e.g. (Schaap, Leij and van Genuchten, 1998). Artificial neural networks refer to computing systems whose central theme is borrowed from the analogy of biological neural networks. They represent simplified mathematical models of biological neural networks. They include the ability to learn and generalize from examples to produce meaningful solutions to problems even when input data contain errors. Training of ANNs consists of finding of minimum of the mean-squared error as dependent on the neuron weights. Recent developments in machine learning methods include the growing research and application of the alternative data driven method called support vector machines (SVMs). SVMs have gained popularity in many traditionally ANNs dominated fields. Using the SVMs eliminates the local minimum issue - the minimum found is always the global one. The objective of this work was to see whether using the SVM to develop PTFs may have some advantages compared with the ANN.

METHODOLOGY

Methods used to fit PTFs

The most common method used in estimation PTFs is to employ *multiple linear regression*. Multiple linear regression (MLR) analysis is generally used to find the relevant coefficients in the model equations. For example:

$$Y = aX_1 + bX_2 + cX_3 + \ldots + X_n, \qquad (1)$$

where Y denotes depended variable, X_n is independent variable.

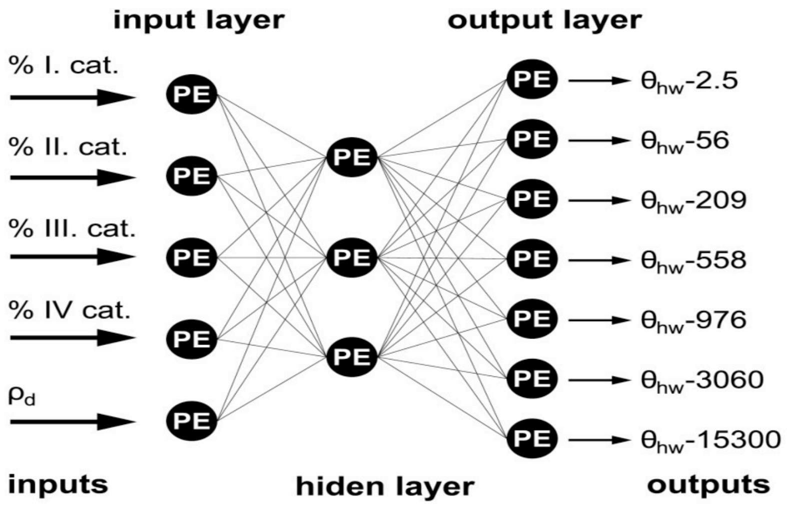


Fig. 1 ANN model for pedotransfer function regression problem solving.

Second approach to model PTFs used in this paper is the application of the *artificial neural networks* (ANNs) This approach was described in some previous works and information about the subject could be found in (Schaap et al., 1998), (Minasny et al., 1999), (Minasny and McBratney, 2002) etc. The mathematical model of an ANN comprises of a set of simple functions linked together by weights. ANN is a simplified simulation of the human brain which is able to learn and generalize from experimental data even if they are noisy and imperfect. This ability allows this computational system to learn constitutive relationships directly from the result of experiments. Unlike conventional models, it needs no prior knowledge. In brief, a neural network consists of an input, a hidden, and an output layer all containing "nodes" or "processing elements - PE" (Figure 1). The number of nodes in input layer (e.g. soil bulk density, soil particle size data, etc.) and output layer (different soil properties) corresponds to the number of input and output variables of the model. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well. Important step in developing an ANN model is training of its weight matrix which represent synaptic connections. The weights are initialized randomly between suitable ranges, and then updated using certain training mechanism. A type of ANN known as multilayer perceptron (MLP), which uses a back-propagation training algorithm, is used for generating PTFs in our study. Training process was performed by the neural network simulator NeuroSolution, which includes a number of training algorithms including the back propagation training algorithm. This is a gradient descent algorithm that has been used successfully and extensively in training feed forward neural networks. Basic information about application of the ANN to regression problems are available in literature and known enough, so we will not go to more detailed explanation here.

ANNs for forecasting and regression problems in hydrology are almost always trained using a multi-layer perceptron (MLP) with the backpropagation algorithm. This may be due in part to the fact that MLPs were the first successful models to be implemented (Rumelhart et al., 1986), and because the algorithm is simple to program and apply. However, there are now many different types of model available, some of which may be more suited to soil water content forecasting and prediction. This may be due in part to the fact that MLPs were the first successful models to be implemented (Rumelhart et al., 1986), and because the algorithm is simple to program and apply. However, there are now many different types of model available, some of which may be more suited to soil water content forecasting and prediction.

For this third approach for pedotransfer function estimation used in this study is *support vector machines* (SVM) type of data driven model. SVMs are learning machines that can be used to solve both classification and regression problems (Vapnik, 1995, 1998). The basic idea is to project the input data into a higher dimensional space by means of kernel functions, called the *feature space*, where linear regression can be performed.

Suppose we are given training data $\{(x_1, y_1),\ldots, (x_n, y_n)\} \subset X \times R$, where X denotes the space of the input patterns (e.g. $X = R^d$). In ε-SVM regression (Vapnik, 1995), goal is to find a function *f(x)* that has at most *ε* deviation from the actually obtained targets y_i for the training data, and at the same time is as flat as possible. In the case of linear functions *f*, it is taking the form:

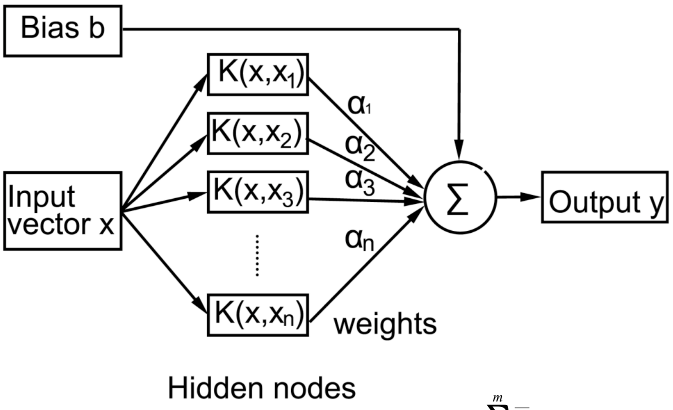
$$f(x) = \langle w, x \rangle + b = 0 \quad w \in X,\ b \in R \tag{2}$$

where ⟨w, x⟩ denotes the dot product in X. Flatness in the case of (1) means that one seeks a small w. One way to ensure this is to minimize the norm, i.e. ||w||2 = ⟨w, w⟩. We also may want to allow for some errors. Slack variables ξi, ξi* were introduced to cope with this. Consequently we arrive at the formulation of the convex optimization problem:

$$\begin{aligned} &minimize \qquad \frac{1}{2}\|w\|^2 + C\sum(\xi_i + \xi_i^*) \\ &subject\ to\ y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i \\ &\qquad \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ &\xi_i,\ \xi_i^* \ge 0 \end{aligned} \tag{3}$$

where ξ_i, ξ_i^* are slack variables that specify the upper and the lower training errors subject to an error tolerance *ε*, and *C* is a positive constant that determines the degree of penalized loss when a training error occurs (the trade-off between the flatness of *f* and the amount up to which deviations larger than *ε* are tolerated). Support vector machine is formulated based on the concept of structural risk minimization (SRM) principle. In (3), the first term of the objective function indicates model complexity and the second term is the empirical risk. The SRM principle theoretically minimizes the expected risk based on the simultaneous minimization of both the empirical risk and the model complexity. Accordingly, a best learned function that minimizes the expected risk by controlling the two terms in (3) can be obtained.

The basic challenge then is to make the SVM algorithm nonlinear. This could be achieved by preprocessing the training patterns x_i by a map $\Phi\colon X \rightarrow F$ into some feature space F and then applying linear SVM regression algorithm. Dot product from (1) is made computationally cheaper by application of the kernel function. Moreover, SVM can be solved by transforming the optimization problem into a quadratic programming algorithm (utilizing Lagrange multipliers), which is a convex function, and the solution to the quadratic programming is unique and optimal. Therefore, support vector machine analytically obtain the optimal network architecture. Architecture of the SVM is very similar to ANN (Fig. 2), only training algorithms differ. The model produced by SVM depends only on a subset of the training data (support vectors), because the cost function for building the model ignores any training data close to the model prediction (within a threshold ε).



$$y = f(x) = \sum_{k=1}^{m} \overline{\alpha}_k \cdot K(x, x_k) + b$$

Fig. 2 SVM model for regression problem solving, weights are Lagrange multipliers, hidden nodes are support vectors and K(x, xk) is kernel function.

Study area and data collection

The data used in this study were obtained from the previous work (Skalová et al., 2003). In that study an area of Zahorská lowland was selected for investigation. A total of 140 soil samples were taken from various localities of this area. The sandy soils occur mainly here.

Soil samples were air-dried and sieved for physical analysis. Particle size analysis according Kopecký grain categories (from 1st till 4th in %) was made utilizing hydrometer methods. Kopecký grain categories is most often used soils texture classification system in Slovakia. Dry bulk density, particle density, porosity and saturated hydraulic conductivity were measured on soil samples too. The points of draying branch WRCs for pressure head values -2.5, -56, -209, -558, -976, -3060 and -15300 cm were estimated in the overpressure equipment.

Full database from 140 samples and their properties was used for creating input data for modelling. At first various usual analyses were accomplished with aim to find possible errors in basic database. According to biggest errors found, 8 samples were excluded, which could affect the correctness of final calculations. Rest of measured data (132 samples) was used for creation of files needed for calculation and consequently three subsets of data were produced:

1. Training data 60% of data
2. Validation data on 20% of data
3. Test data 20% of data

Training and validation data are both used in models calibration, e.g. the data set were actually divided into two subsets for calibration (80%) and testing (20%). Only in the case of MLP computations are first two subsets used separatly, but both in calibration (or training) part of MLP application.

RESULTS

Three methods of WRCs assessment for the soils of Záhorská lowland were used: multiple linear regression, multi-layer perceptron type of artificial neural networks and support vector machines. It was applied to the data set, which contained data from 132 soil samples.

Multi-linear regression for PTFs assessment was used in form:

$$\theta_{hw} = a*1^{st}\ cat. + b*2^{nd}\ cat. + c*4^{th}\ cat. + d*\rho_d + e, \tag{4}$$

where θ_{hw} is water content [cm^3.cm^{-3}] for the particular pressure head value h_w [cm], 1^{st} *cat.*, 2^{nd} *cat.*, 4^{th} *cat.* are percentage of clay, sild and sand, ρ_d is dry bulk density [g.cm^{-3}] and *a, b, c, d, e* are parameters determined by regression analysis.

Results of multi-linear regression are listed in Table 1. Correlation coefficients (*R*) for each of the PTFs are also there. The *R* testifies to a high degree of relationship between correlated elements ($0.81 < R < 0.89$). Designed PTFs were checking on testing dataset that consist of 26 soil samples.

Table 1 Results of multi-linear regression

h_w [cm]	a	b	c	d	e	R
-2.5	0.025	-0.166	-0.141	-38.575	109.780	0.895
-56	0.073	-0.250	-0.296	-27.309	93.465	0.850
-209	0.216	-0.161	-0.250	-19.073	68.286	0.864
-558	0.226	-0.176	-0.247	-21.009	68.889	0.865
-976	0.180	-0.217	-0.281	-19.977	68.712	0.842
-3060	0.253	-0.219	-0.230	-17.932	58.326	0.831
-15300	0.192	-0.270	-0.281	-14.718	54.796	0.810

Second approach to determination of water retention curves in presented work was artificial neural networks (ANN). Proposed network consists of a set of input units (inputs are fractions of soil particle in the same classes as in the previous method and dry bulk density), a set of output units (θ_{hw} - volume water content) and a set of hidden units, which links the inputs to outputs (Figure 1). The hidden units extract useful information in learning phase and use them to predict the outputs. We should to determine network architecture for this purpose. It means to determine the number of neurons in hidden layer (what means number of connection weights - free parameters) and the way information flows through the network. Neuron, with a bias and tanh function was used. This will squash the range of each neuron in the layer between - 1 and 1. Such nonlinear elements provide a network with the ability to make soft decisions. In this work cross validation was used in training process an independent test set is used to assess the performance of the model at various stages of learning. As the testing set must not be used as part of the training process, a different, independent testing set is needed for the purposes of cross-validation. This means that the available data need to be divided into three subsets; a training set, a testing set and a validation set, as was mentioned.

As learning law we used Levenberg-Maquardt method. We trained networks for pressure head value h_w = -2.5, -56, -209, -558, -976, -3060, -15300 cm with hidden layer with 2, 3, 4 neurons. Then testing dataset was computed with this ANNs. Results with regression coefficients are summarized in Table 2. As could be seen there, ANN provides better results in comparing with multi-regression analysis.

Table 2 Regression coefficients for testing dataset obtained from ANN computations

	Number of neurons in hidden layer		
hw [cm]	4	3	2
-2,5	0.871	0.900	0.912
-56	0.925	0.907	0.906
-209	0.904	0.888	0.866
-558	0.901	0.871	0.864
-976	0.833	0.761	0.851
-3060	0.727	0.822	0.864
-15300	0.812	0.864	0.867

Given regression problem was solved by using *support vector machines*, too. In order to minimize error and speedup convergence, the original samples are normalized by:

$$y_i = kx_i + q, \tag{5}$$

where x_i is original value, y_i is normalised value and k, q are following constant for each data column:

$$k = \frac{2}{x_{\max} - x_{\min}}; \quad q = \frac{x_{\min} + x_{\max}}{x_{\min} - x_{\max}}, \tag{6}$$

where x_{max} and x_{min} is maximal and minimal value in column. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Scaling by applying (5) and (6) is linear scaling to interval (-1, 1).

SVM regression estimation steps are the following: 1) selection of a suitable kernel and appropriate kernel's parameter; 2) specifying the ε parameter; and 3) specifying the capacity C. On trial and error basis radial basis function was chose as kernel function. This function has following form:

$$K(x_i, x_j) = exp(-\lambda|| x_i - x_j||^2), \lambda>0 \tag{7}$$

This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. Parameter γ of this kernel function, tube size ε for ε-insensitive loss function and parameter C which controls the tradeoff between errors of the SVM on training data and margin maximization were found by grid search programmed in Matlab environment. As mentioned above, a common strategy was used which is to separate the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the ”unknown” set more precisely reflects the performance on an independent data set, so correlation coefficient from test set was evaluation criterion in grid search. The grid search is straightforward and seems naive because there are several more advanced methods which can save computational cost, for example application of various heuristic search methods. However, we prefer the simple grid search approach. The reason is that the computational time required to find good parameters by grid search is not much more than that by advanced methods since there are only three parameters.

The analysis and calculation of SVM were performed by using the LIBSVM software, developed by Chang and Lin (2001). We trained SVM model for pressure head value h_w = -2.5, -56, -209, -558, -976, -3060, -15300 cm. Then testing dataset was computed with obtained models and results are summarized with regression coefficients in Table 3. There could be seen, that this results are clearly better in comparing with multi-linear regression and somewhat better in comparing with ANN.

Table 3 Regression coefficients for testing dataset obtained from SVM computations

hw [cm]	-2.5	-56	-209	-558	-976	-3060
R	0.923	0.933	0.937	0.962	0.927	0.921

CONCLUSIONS

The research presented in this paper attempts to develop a more precise model using multi-layer perceptron and support vector machines instead of traditional models like multiple linear regression for predicting some soil hydrological properties. Pedotransfer functions for point estimation of soil hydraulic parameters from basic soil properties such as particle-size distribution, bulk density were developed and validated using multiple-linear regression, artificial neural network and support vector machine methods and the predictive capabilities of the three methods was compared. Total of 132 soil samples was divided into two groups as 105 for the development and 27 for the validation of PTFs. Accuracy of the predictions was evaluated by the correlation coefficient (*R*) between the measured and predicted parameter values. The *R* varied from 0.81 to 0.895 for regression, for 0.851 to 0.925 for ANN and varied from 0.913 to 0.962 for SVM, respectively. The correlation coefficients of these methods shows that the SVM in our experiment resulted from all methods with the best result, e.g. comparison with multi-linear regression shows about 9% better results and 4.5% better result in comparing with ANN was achieved (Tables 1, 2, 3). This suggests the greater accuracy of the calculations as well as their greater stability and need of less time devoted to calculations, since the ANN training sometimes stuck in a local minimum so the training process has to be reset and run many times. The development of ANNs followed a heuristic path, with applications and extensive experimentation preceding theory. In contrast, the development of SVMs involved sound theory first, then implementation and experiments. A significant advantage of SVMs is that whilst ANNs can suffer from multiple local minima, the solution to an SVM is global and unique.

Acknowledgement This work was supported by the Slovak Research and Development Agency under the contract No. LPP-0319-09, APVV-0443-07, APVV-0271-07 and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/0585/08.

REFERENCES

Bouma, J. (1989) Using soil survey data for quantitative land evaluation, Adv. Soil Sci., **9**, 177–213.

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.

Minasny, B., McBratney, A. B. (2002) The neuro-m methods for fitting neural network parametric pedotransfer functions", Soil Sci. Soc. Am. J. 66: 352–361.

Minasny, B. A., McBratney, B., Bristow, K. L. (1999) Comparison of different approaches to the development of pedotransfer functions for waterretention curves", Geoderma, 93: 225–253.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986) Learning internal representation by error propagation, Rumelhart, D.E. & McClelland, J.L. (eds.) *Parallel distributed processing: explorations in the microstructure of cognition*, Vol.**1**. Cambridge MA, MIT Press, pp. 318-362.

Schaap, M. G., Leij, F. J., Van Genuchten, M. (1998) Neural network analysis for hierarchical prediction of soil hydraulic properties, Soil Sci. Soc. Am. J, 62: 847–855.

Schaap, M. G., Leij, F. J., Van Genuchten, M. (1998) Neural network analysis for hierarchical prediction of soil water retention and saturated hydraulic conductivity, Soil Sci. Soc. Am. J, 62:847-855.

Skalová, J. (2001) Pedotransfer functions of the Záhorská nížina soils and their application by soil water regime modelling, Faculty of Civil Engineering STU Bratislava, 112 p.

Skalová, J., Šútor, J., Štekauerová, V. (2003) Pedotransferové funkcie pre pôdy Záhorskej nížiny, *Acta horticulturae et regiotectura,* ISSN 1335-2563. - Roč. 6. č. mimoriadne, s. 66-69.

Štekauerová, V., Skalová, J. (1999) Calculation of the drying brunch of water retention curves from the easily measured soil properties, VII. Poster day. UH SAS Batislava, 133-134.

Šútor, J., Štekauerová, V. (1999) Determination of the water retention curve points from the basic physical characteristics of soil, Influence of anthropogenic activity for water regime of plain area. ÚH SAV, Michalovce, 151-157.

Vapnik, V. (1998) Statistical Learning Theory", Wiley, NY.

Vapnik, V. (1995) The Nature of Statistical Learning Theory, *Springer*, NY.

Wösten, J. H. M., Pachepsky, A. Ya., Rawls, W. J. (2001) Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics", *J. Hydrol*, 251, 123–150.