J. Hydrol. Hydromech., *59*, 2011, 4, 275–284
DOI: 10.2478/v10098-011-0023-7

COMPARISON OF THREE REGRESSION MODELS FOR DETERMINING WATER RETENTION CURVES

JANA SKALOVÁ, MILAN ČISTÝ, JURAJ BEZÁK

Slovak University of Technology Bratislava, Faculty of Civil Engineering, Radlinského 11, 813 68 Bratislava, Slovak Republic; Mailto: jana.skalova@stuba.sk

A key physical property used in the description of a soil-water regime is a soil water retention curve, which shows the relationship between the water content and the water potential of the soil. Pedotransfer functions are based on the supposed dependence of the soil water content on the available soil characteristics, e.g., on the relative content of the particle size in the soil and the dry bulk density of the soil. This dependence could be extracted from the available data by various regression methods. In this paper, artificial neural networks (ANNs) and support vector machines (SVMs) were used to estimate a drying branch of a water retention curve. The paper compares the mentioned methods by estimating the water retention curves on regional scale for the Záhorská lowland in the Slovak Republic, where relatively small data set was available. The performance of the models was evaluated and compared. These computations did not fully confirm the superiority of SVMs over ANNs as is often proclaimed in the literature, because the results obtained show that in this study the ANN model performs somewhat better and is easier to handle in determining pedotransfer functions than the SVM models. Nevertheless, the results from both data-driven models are quite close, and the results show that they provide a significantly more precise outcome than a traditional multi-linear regression does.

KEY WORDS: Soil-Water Retention Curve, Pedotransfer Function, Neural Networks, Support Vector Machines, Multiple Linear Regression.

Jana Skalová, Milan Čistý, Juraj Bezák: POROVNANIE TROCH REGRESNÝCH MODELOV NA URČENIE VLHKOSTNÝCH KRIVIEK PÔD. J. Hydrol. Hydromech., *59*, 2011, 4; 21 it., 2 obr., 5 tab.

Autori sa v príspevku venujú určovaniu pedotransferových funkcií (PTF), ktoré umožňujú stanoviť body vlhkostných retenčných kriviek pôdy z ľahšie merateľných pôdnych vlastností a sú dôležitým prvkom modelovania vodného režimu pôdy. Ešte v minulej dekáde sa objavili snahy využívať na ich určenie umelé neurónové siete (UNS). Multi-layer perceptron (MLP) čiže viacvrstvový perceptrón je najčastejšie používaný model doprednej umelej neurónovej siete s kontrolovaným typom učenia. Vstupné signály prechádzajú sieťou typu MLP iba dopredným smerom, teda postupne od vrstvy k vrstve. MLP používa tri a viac vrstiev neurónov rozdelených na vstupnú, skrytú a výstupnú vrstvu s nelineárnou aktivačnou funkciou a vie rozpoznať alebo modelovať informácie, ktoré nie sú lineárne oddeliteľné alebo závislé. Novší vývoj v oblasti učiacich algoritmov poskytuje ďalšie možnosti, z ktorých sa v tomto príspevku venujeme tzv. mechanizmom podporných vektorov (Support Vector Machines – SVM). SVM využíva pri svojom kalibrovaní na riešený problém princíp tzv. štrukturálnej minimalizácie namiesto iba minimalizácie chyby – (*Vapnik*, 1995). Pri trénovaní siete MLP je jediným cieľom minimalizovať celkovú chybu. Pri SVM sa simultánne minimalizuje chyba aj zložitosť modelu. Použitie tohto princípu vedie zvyčajne k vyššej schopnosti generalizácie, t.j. umožneniu presnejších predpovedí pre dáta, ktoré neboli použité pri trénovaní SVM. Vhodnosť štandardnej umelej neurónovej siete, SVM a viacnásobnej lineárnej regresie sa v článku vyhodnocuje na základe údajov získaných z pôdnych vzoriek odobratých v lokalite Záhorskej nížiny. Pôvodné údaje a ich aplikáciu pri vyhodnocovaní vodného režimu pôd uvádza *Skalová* (2001, 2007), odkiaľ boli prevzaté vstupné dáta a to percentuálny obsah zrnitostných kategórií (I až IV podľa Kopeckého), redukovaná objemová hmotnosť (ρ_d) a vlhkosti pre vlkostné potenciály h_w= –2.5, –56, –209, –558, –976, –3060, –15300 cm, ktoré boli stanovené laboratórne pre potreby určenia a testovania regresných závislostí. Vzhľadom na to, že pri odvodzovaní regionálnych PTF je častým prípadom nedostatok dát pre odvodenie dátovo riadených modelov, autori navrhli riešiť úlohu pomocou ansámblu MLP resp. SVM. Ansámbel dátovo riadených modelov bol vytvorený variabilným rozdelením údajov na trénovacie a validačné (validačnými údajmi sa testuje presnosť modelu vo fáze jeho tvorby, ešte sa používajú konečné testovacie dáta,

ktoré neboli pri tvorbe modelu použité). Výsledky ukázali lepšie regresné schopnosti oboch dátovo riadených modelov (SVM aj MLP) voči multilineárnej regresii a o niečo lepšie výsledky boli získané z viacvrstvového perceptrónu než zo SVM. Keďže v niektorých iných prácach mal zvyčajne vyššiu výpočtovú presnosť model založený na SVM než na UNS, autori odporúčajú pre budúci výskum preveriť vhodnosť kombinácie SVM a MLP modelov v dátovo riadenom skupinovom modeli.

KĽÚČOVÉ SLOVÁ: retenčná krivka, pedotransférová funkcia, neurónová sieť, mechanizmus podporných vektorov, viacnásobná lineárna regresia.

Introduction

Modeling water and solute transport in soil has become an important tool in simulating agricultural productivity as well as in dealing with various environmental quality issues. For instance, optimum irrigation management requires a systematic estimation of the soil-water status to determine both the appropriate amounts and timing of irrigation. A possible ecological application of soil-water modeling is described, e.g., in *Skalová* et al., 2009. In general, two categories of methods for evaluating a soil-water regime can be distinguished: (1) the measurement techniques and (2) predictive methods (mathematical modeling). 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, the use of mathematical models depends on knowledge of the input data which are needed for the numerical simulations. Some of this data (meteorological, climatic, hydrological or crop characteristics) are usually available in competent institutions, but hydraulic soil properties are only available for some sites in Slovakia (the same situation is usual in other countries). That is why these characteristics appear as a key problem in the numerical simulation of a soil-water regime. During the last ten years a relatively large number of works have appeared which were devoted to determining the water retention curve which is needed for this purpose from more easily available soil properties such as particle size distribution, dry bulk density, organic C content, etc., e.g. *Gupta* and *Larson*, 1979; *Rawls* et al., 1982; *Minasny* et al., 1999, and in Slovak scientific literature (*Šútor*, *Štekauerová*, 1999; *Štekauerová*, *Skalová*, 1999, etc.). Pedotransfer functions (PTF) have become the term for such relationships between soil hydraulic parameters and the more easily measurable properties usually available from a soil survey (*Bouma*, *Van Lanen* 1987; *Bouma*, 1989). Consequently, the method for the quantification of these relationships uses various types of regression analyses. The aim of this paper is a comparison of three regression models for determining pedotransfer functions.

Besides the standard regression methods, artificial neural networks (ANNs) have become the tool of choice in the last decade in developing PTFs, e.g., *Schaap* et al., 1998; *Pachepsky* et al., 1996; *Tamari* et al., 1996 etc.. The above authors confirm that they received better results from ANN-based pedotransfer functions than from standard regression-based PTFs.

One of the advantages of ANN-based PTFs compared to traditional regression PTFs is that they do not require an a priori regression model, which relates the input and output data, and which in general is difficult to define, because these models are not known (*Minasny* and *McBratney*, 2002). The training of an ANN is basically an iterative process; on the other hand, some problems from its character may sometimes arise. In *Twarakawi* et al., 2009, the possible weaknesses of the ANN approach are confirmed and summarized as follows: 1. ANNs have a number of coefficients (weights) that do not permit easy physical interpretation (*Schaap* et al., 2001); 2. the ANN's structure has to be selected a priori and therefore may not be optimal since there are many types of neurons and many types of possible connections (*Wösten* et al., 2001); 3. a higher number of neurons and connections than required can result in overfitting and over parameterization (*Hastie* et al., 2001), which can negatively affect the ability of the models obtained to generalize. That is why also other possibilities for solving given regression task were evaluated in this study.

Recent developments in machine learning methods have forced the application of alternative data-driven methods in hydrology applications, e.g., radial basis function networks (*Tamari* et al., 1996; *Kumar* et al., 2010) or support vector machines (*Lamorski* et al, 2008; *Twarakawi* et al., 2009). The foundations of support vector machines (SVMs) were developed by *Vapnik* (1995) and are gaining in popularity due to their attractive features and promising empirical performance. A support vector

machine was proposed by *Vapnik* (1995), as a statistical learning method with a promising ability to generalize. It maps the training vectors into a high dimensional feature space and constructs a hyperplane that maximizes the margin (i.e. maximizes the distance between the hyperplane and the closest training vector in the feature space). The SVMs formulate a quadratic optimization problem for finding such a hyperplane, which ensures a global optimum for a given parameter set. The formulation embodies the structural risk minimization (SRM) principle in addition to the traditional empirical risk minimization (ERM) principle employed by conventional neural networks. SRM minimizes the upper boundary on the risk expected, as opposed to ERM, which only minimizes an error on the training data. It is this difference which gives SVMs a greater ability to generalize, which is the goal of statistical learning.

The objective of this work is to verify the abovementioned advantages of SVMs while developing PTFs for the Záhorská lowland, which was selected as a representative region for the investigation (e.g., while solving the regression task of determining the water retention curve from easily available soil properties). The data used in this study were obtained from previous work (*Skalová*, 2001).

In the following part of the paper ("Methodology") the three methods used in this study – multiple linear regression, ANN and SVM are briefly explained. Then the data acquisition and preparation is presented. In the "Result" part, the settings of the experimental computations are described in detail, and the "Conclusion" of the paper evaluates these experiments on the basis of the statistical indicators.

Materials and methods

Methods used to fit the PTFs

The most common method used in estimating 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:

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

where Y denotes a dependent variable, X_n – an independent variable.

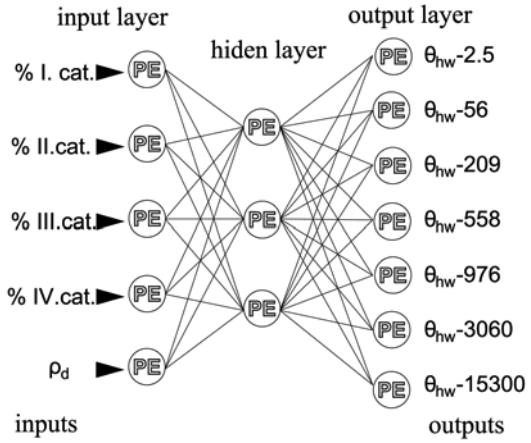

Fig. 1. ANN model for evaluating a pedotransfer function. Each connection is characterized by its weight, by which the data flow is transformed from the inputs to the outputs.

The second approach for modeling the PTFs used in this paper is the application of *artificial neural networks* (ANNs). This approach has been described in various previous works, and information about the subject can be found in *Schaap* et al., 1998; *Minasny* et al., 1999; *Minasny* and *McBratney*, 2002, etc. Briefly summarized, a neural network consists of input, hidden and output layers, all containing "nodes" or "processing elements (PE)" (Fig. 1). The number of nodes in the input layer (e.g., the soil's bulk density, the soil's particle size data, etc.) and output layer (various soil properties) correspond to the number of input and output variables of the model. So-called "learning" involves adjustment of the coefficients (i.e., the synaptic connections that exist between the neurons or weights), which are used for the transformation of the inputs to the outputs. For that reason, an important step in developing an ANN model is the training (computing) of its weight matrix. A type of ANN known as a multi-layer perceptron (MLP), which uses a back-propagation training algorithm, was used for generating the PTFs in our study. The training process was performed by the NeuroSolution neural network simulator, which includes a number of training algorithms, including a back propagation training algorithm of an MLP. This is a gradient descent algorithm that has been successfully and extensively used in training feed-forward neural networks. The basic information about the application of an ANN to regression problems is available in the literature and is well known, so we will not provide a more detailed explanation here.

A third approach called *support vector machines* (SVM) for estimating the pedotransfer functions

used in this study is explained hereinafter, with explanations of its principles only to such an extent considered necessary for understanding its possible advantages versus ANN and for explanation of the various settings of this methodology, which are necessary for its application. A more detailed description of the methodology can be found, e.g., in *Vapnik*, 1995.

The architecture of a SVM is similar to that of an ANN, but the training algorithm is significantly different. The basic idea is to project the input data by means of kernel functions into a higher dimensional space called the *feature space*, where a linear regression can be performed for an originally nonlinear problem, the results of which are then mapped back to the input space. The linear regression is maintained by quadratic programming, which ensures a global optimum and an optimal generalization. The uniqueness of this solution is often emphasized, but the actual truth is that this solution is only unique for a given set of performance parameters, which should be chosen and will be described later.

Suppose we are given training data $\{(x_1, y_1),\ldots, (x_l, y_l)\}$, where $x_i \in X = R^n$ denotes the input pattern for the i-th sample, and y_i is the desired model's output. A non-linear transformation function $\Phi(\cdot)$ is defined to map the input space to a higher dimension feature space. The important idea is to fully ignore small errors (by introducing the "tube" variable ε, which defines what the "small" error is) to make the regression sparse, that is, dependent on a smaller number of inputs (called the support vectors), which makes the methodology much more computationally treatable. In an ε-SVM regression (*Vapnik*, 1995), the goal is to find a function $f(x)$ that at most has an ε deviation from the actually obtained targets y_i (or $f(x)$) for the training data:

$$f(x) = w.\Phi(x) + b \quad w \in X, b \in R, \tag{2}$$

where $f(x)$ is the model's output, and input x is mapped into a feature space by a nonlinear function $\Phi(x)$ with weight vector w and bias b. According to the structural risk minimization principle, solving the optimal fitting function $f(x)= y_i$ can be expressed as the following optimization problem:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3}$$

$$\text{subject to} \quad y_i - (w.\Phi(x) + b) \le \varepsilon + \xi_i$$

$$(w.\Phi(x) + b) - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

where ξ_i, ξ_i^* are slack variables that specify the upper and lower training errors, subject to an error tolerance ε, and C is a positive constant that determines the degree of the penalized loss when a training error occurs. In (3), the first term of the objective function indicates the model's complexity, and the second term is the empirical risk. That is why this objective function simultaneously minimizes both the empirical risk and the model's complexity; the tradeoff between these two goals is controlled by parameter C. An important characteristic of SVMs as a consequence of this fact is that a better ability to generalize could be expected, compared, e.g., with ANNs (the better results for the data which were not used for building the model), because unnecessarily complex models usually suffer from overfitting.

Moreover, an SVM can be solved by transforming the optimization problem into its dual form via a quadratic programming algorithm (utilizing Lagrange multipliers), and the solution to the quadratic programming is unique and optimal. Therefore, a support vector machine analytically obtains the optimal network architecture, which partially avoids the problem with the local minima which arises in training the ANNs.

The radial basis function was chosen on a trial and error basis as the kernel function for this work (the function used to transform a nonlinear problem from an input space to a high dimensional space for the sake of the possibility of solving a linear problem instead of a nonlinear problem, which is a characteristic feature of an SVM). This function has the following form:

$$K(x_i, x_j) = exp(-\gamma|| x_i - x_j||^2), \gamma > 0. \tag{4}$$

The parameter γ of this kernel function, the tube size ε for the ε-insensitive loss function, and parameter C should be found, which is the basic task when SVMs are applied to any practical problem. The *harmony search methodology*, which was programmed in a Matlab environment, was used for this purpose instead of the usual trial-and-error principle, which is more efficient. A harmony search is a metaheuristic search algorithm introduced by Geem (*Geem* et al., 2002) and is inspired by the improvisational process of musicians. In an HS algorithm, each musician corresponds to one decision variable. A musical instrument's pitch range corresponds to a decision variable's value range; the musical harmony at a certain time corresponds to a solution vector at a certain iteration;

and the audience's aesthetics corresponds to an objective function. Just as a musical harmony is improved time after time, a solution vector is improved iteration by iteration by the application of the improvisation's operators (the random selection of a tone, a musician's memory considerations or a pitch adjustment). This methodology is described in more detail in the various works of Geem (*Geem*, 2001) or other authors (e.g. *Čistý*, 2010).

For each combination of the parameters (γ, C, ε) generated in the iterations of this search process, an SVM model using this actual combination of parameters based on the training data is created. As a criterion for selecting the appropriate combinations of the parameters, the correlation coefficient could be used in regression task of determining PTF as the value of the objective function of the harmony search methodology.

Study area and data collection

It should be noted that when modeling empirical data by the means described hereinbefore, a process of induction is used to build up a model of the system, from which it is hoped to deduce responses of the system that are unknown. Ultimately, the quantity and quality of the observations made while obtaining the training data will govern the performance of these empirical models. This paper is aimed at a situation when there is not enough data available, which could often be the case; and this has some impact which is described in results of this work.

The data used in this study were obtained from a previous work (*Skalová*, 2001). An area of the Záhorská lowland was selected for testing the methods described. A total of 140 soil samples were taken from various localities in this area (Tab. 1).

The soil samples were air-dried and sieved for a physical analysis. A particle size analysis according to the Kopecký grain categories from 1st to 4th was performed utilizing Cassagrande's methods. 1st category means the percentages of the clay (diameter < 0.01 mm), 2nd cat. – silt $d \in (0.01–0.05$ mm), 3rd cat. – fine sand $d \in (0.05–0.1$ mm) and 4th cat. – sand $d \in (0.1–2.0$ mm). The Kopecký grain categories are very often used in Slovakia for the soil texture classification. The dry bulk density, particle density, porosity and saturated hydraulic conductivity were also measured on the soil samples. The points of the drying branches of the WRCs for the pressure head values of –2.5, –56, –209, –558, –976, –3060 and –15300 cm were estimated using overpressure equipment (set for pF-determination with ceramic plates).

A full database of the 140 samples and their properties was used for creating the input data for the modeling from which the three subsets of the data were produced:

- Training data: 88 data samples;
- Validation data: 22 data samples;
- Test data: 30 data samples.

The training and validation data were both used in calibrating the models, e.g., the data set was actually divided into two subsets for the calibration (110 samples) and testing (30 samples). A practical way to find a better generalization model is to set aside a small percentage (around 20%) of the training set and use it for the cross validation. When the error in the validation set increases, the training should be stopped because the best generalization has been reached.

T a b l e 1. Classification of soil samples taken from Záhorská lowland (expressed by total number and %) and land area of soil types (expressed by %)

Soil type	[%] of I. grain category (< 0.01 mm)	Soil samples		Land area
		Number	[%]	[%]
1. Sandy soil	0–10%	56	40	47
2. Loam sandy soil	10–20%	34	24	15
3. Sandy loam soil	20–30%	21	15	6
4. Loam soil	30–45%	22	16	27
5. Clay loam soil	45–60%	4	3	3
6. Silty clay soil	60–75%	3	2	1
7. Clay	> 75%	0	0	0
Total		140	100	100

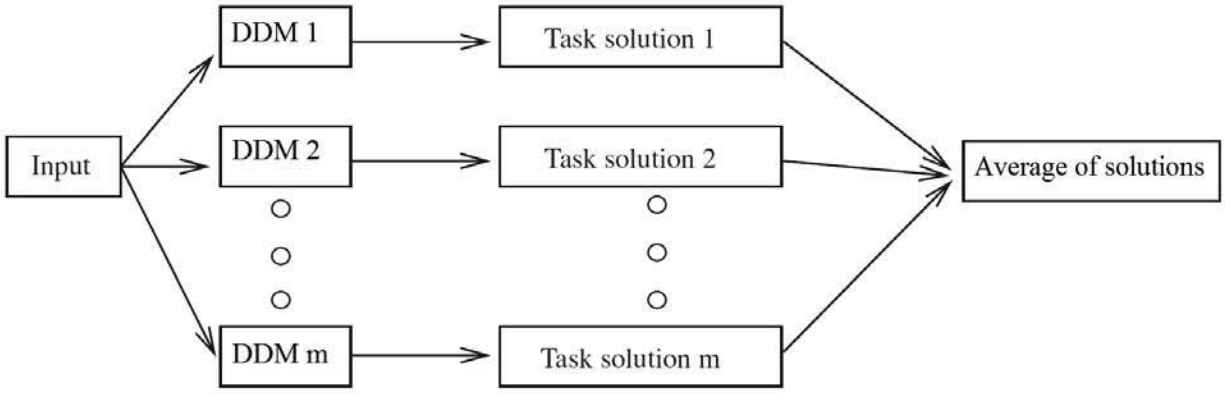

Fig. 2. Ensemble data-driven model (DDM) for evaluating a pedotransfer function.

Statistically similar data should be in all three data subsets, but this is not easy to accomplish, especially in this case, when the data set is relatively small. That is why an ensemble of data-driven modeling was used in the present work, which means a collection of a finite number of data-driven models that are trained for the same task (Fig. 2). This is meant as a simple variant of bootstrapping (the bootstrap scheme involves generating subsets of the data on the basis of random sampling with replacements as the data are sampled). Five data-driven models are trained independently, and their predictions are combined for the sake of obtaining a better generalization. For this reason the mentioned training fraction of the data (110 samples) was divided into the training and validation data sets alternatively in five different versions. This led to five models both for the ANN and SVM, the results of which were combined for the given type of model – final result is average value from all five models. Through this approach the authors intended to avoid obtaining a model based on a wrong data division in the data sets.

Results

A multi-linear regression for assessing the PTFs was used in the form:

$$\theta_{hw} = a * 1^{st}\text{ cat.} + b * 2^{nd}\text{ cat.} + c * 4^{rd}\text{ cat.} + d * \rho_d + e, \quad (5)$$

where θ_{hw} is the water content [$cm^3.cm^{-3}$] for the particular pressure head value h_w [cm], 1^{st} cat., 2^{nd} cat., and 4^{th} cat. – the percentages of the clay (d < 0.01 mm), silt d∈(0.01–0.05 mm), and sand (0.1–2.0 mm), ρ_d – the dry bulk density [g cm^{-3}], and *a, b, c, d, e* are the parameters determined by the regression analysis. In the case of the multi-linear regression it was not possible to use ensemble models, because in this case no iterative process is applied which involves cross validation, so all 110 training samples were used as a whole for the development of the model. Fine sand (3^{rd} cat.) was not included in regression equation because of avoiding correlation between independent variables.

The PTFs designed were evaluated on a testing dataset, which consisted of 30 soil samples. The results of the multi-linear regression are listed in Tab. 2. The correlation coefficients (*R*) for each of the PTFs testify also less degree of the relationship between the correlated elements in some cases (with smallest value 0.63).

The second approach used in determining the water retention curves in the presented work was an application of *ensemble neural networks*. The same network architecture for every ANN in the network was determined. In this work the multilayer perceptron (MLP) with 2, 3, and 4 neurons in the hidden layer was tested; an MLP with 3 neurons in the hidden layer was finally chosen for the ensemble neural network model. A neuron with a bias and tanh activation function was used. The Levenberg-Maquardt method was used in the context of the back propagation method.

The networks were trained for computing the water content at the pressure head value h_w = –2.5, –56, –209, –558, –976, –3060, –15300 cm. Then the testing dataset was computed with the ensemble ANN. The results with the regression coefficients are summarized in Tab. 3. As can be seen, the ensemble ANN provides significantly better results compared with the multi-regression analysis.

Lastly, we solved the given regression problem using *ensemble support vector machines*. The estimation of the steps of the SVM regression (described in the methodology part of this paper) are the following: 1. the selection of a suitable kernel and the appropriate kernel's parameter (γ in eq. (4)); 2. specifying the ε parameter (eq. (3)); and 3. specifying the capacity C (eq. (3)).

As was noted in the methodology part, the radial basis function was chosen as the kernel function on a trial and error basis, and the harmony search methodology was used for finding the parameters of the SVM.

As a criterion for selecting the appropriate combinations of the parameters, the correlation coefficient for the training and cross-validation data is calculated within the objective function of the harmony search, where the correlation coefficient of the cross-validation data was weighted by the coefficient 1.2 for the sake of a better generalization. This combination of correlation coefficients is generally not necessary (only the correlation coefficient of the cross-validation data is usually used), but due to the relatively small data set, this was shown to be more effective.

T a b l e 2. Results of the multi-linear regression – regional parametres of pedotransfer functions (*a*, *b*, *c*, *d*, *e*, *f*) for calculation of points of drying branch of water retention curve for Záhorská nížina soils (h_w – pressure head, R – correlation coefficient)

h_w [cm]	*a*	*b*	*c*	*d*	*e*	*R*
–2.5	0.044	–0.16	–0.135	–37.288	106.632	0.907
–56	0.091	–0.269	–0.33	–25.314	91.044	0.88
–209	0.253	–0.123	–0.233	–17.902	63.167	0.877
–558	0.291	–0.116	–0.203	–19.969	61.745	0.876
–976	0.259	–0.146	–0.219	–18.668	59.757	0.876
–3060	0.314	–0.153	–0.178	–17.386	51.733	0.877
–15300	0.252	–0.215	–0.231	–15.185	50.415	0.866

T a b l e 3. Results of the ANN for determining PTFs (h_w – pressure head, $R1$ – $R5$ partial correlation coefficients for members of ANN ensemble and R – final correlation coefficient by ANN for given pressure head)

h_w [cm]	*R*1	*R*2	*R*3	*R*4	*R*5	*R*
–2.5	0.914	0.921	0.937	0.934	0.888	0.930
–56	0.925	0.928	0.938	0.899	0.889	0.925
–209	0.898	0.926	0.934	0.905	0.878	0.921
–558	0.883	0.927	0.943	0.879	0.877	0.912
–976	0.872	0.923	0.937	0.871	0.865	0.905
–3060	0.888	0.920	0.930	0.866	0.918	0.914
–15300	0.882	0.901	0.935	0.871	0.915	0.910

T a b l e 4. Results of SVM (hw – pressure head, R1 – R5 partial correlation coefficients for members of SVM ensemble and R – final correlation coefficient by SVM for given pressure head)

h_w [cm]	*R*1	*R*2	*R*3	*R*4	*R*5	*R*
–2.5	0.916	0.914	0.908	0.916	0.897	0.913
–56	0.887	0.880	0.885	0.869	0.873	0.883
–209	0.891	0.888	0.898	0.893	0.886	0.894
–558	0.870	0.868	0.871	0.871	0.862	0.871
–976	0.879	0.888	0.892	0.886	0.880	0.892
–3060	0.876	0.878	0.883	0.877	0.868	0.882
–15300	0.856	0.881	0.878	0.873	0.876	0.874

The analysis and calculation of the SVM were performed using the LIBSVM library in C++ developed by *Chang* and *Lin* (2001), which was called from the Matlab code of the harmony search written by the authors of this paper. In the training phase, SVM models for a pressure head value of h_w = –2.5, –56, –209, –558, –976, –3060, –15300 cm were created. This was repeated five times because of the five divisions of the data used to train the model on the training and validation data set. A

total of 35 computations were run. Then the testing dataset was computed five times with the models obtained, and the final result is the average of the outputs from these five models; the results are summarized with the regression coefficients in Tab. 4. As can be seen, these results are clearly better compared with the multi-linear regression and somewhat worse compared with the ANN.

From these results, it seems that ANNs are more resistant to an insufficient amount of data (which is the case in this work), because, on the other hand, better results with the application of the SVM than with the ANN for the PTF evaluation were reported in the literature (*Lamorski* et al, 2008; *Twarakawi* et al., 2009). It should be mentioned that the authors of these papers worked with larger data sets (806 and 2134 soil samples). For this reason the authors of the present paper hypothesize that it is advisable to use combined SVM/MLP models, because of the variability of an adequate methodology, but this should be verified in future work.

On the basis of this work the authors also have some remarks on the effectiveness of the manipulation with these two data-driven models. Both MLP and SVM training involves searching for some parameters which should be set properly for the sake of obtaining good quality results from the model. Heuristic searching (which is more comfortable and objective) of such parameters could be applied to both of them, but in the case of SVMs this application is easier. This is due to the deterministic type of training of the SVM algorithm, which is faster and offers unique results for a particular combination of the parameters, which is not true in the case of iterative MLP training. On the other hand, from the point of view of manipulating the SVM and ANN models, the advantage of ANNs versus SVMs is that a multiple output (more variables) is possible to obtain from an ANN, whereas it is necessary to build different SVM models for the different pressure head values of h_w.

The quality of SVM models depends on the proper setting of their parameters. If parameter C (Eq. (3)), (which determines the trade-off between a model's complexity and the degree to which deviations larger than ε are tolerated) is too large, the rate of accuracy of the estimation is high in the training phase, but may be low in the testing phase. If C is too small, the accuracy of the estimation is unsatisfied, and the model is useless. The kernel parameter γ (eq. (4)) also has a great influence on any estimates. An excessively large value for parameter γ results in overfitting, while a disproportionately small value leads to underfitting. The complexity of the data-driven model (the number of free parameters) has a strong influence on the model's ability to generalize. The complexity of an ANN model is setting directly (as the number of neurons in the hidden layer), but the complexity of an SVM is defined by the number of support vectors, which is the aim of the computation on the basis of the abovementioned parameters. Consequently, there is a lower possibility of setting the complexity of the SVM model intuitively on the basis of some trials as in the case of an ANN. Although this necessity for setting the parameters of an ANN on the basis of a modeler's know-how is often criticized, it can be seen that it was advantageous in this work (on the basis of the results in Tabs. 3, 4). In Tab. 5 the parameter settings and number of support vectors in the first of the five SVM models working as an ensemble is displayed as an illustration of the ideas in this paragraph.

T a b l e 5. Parameters of the SVM for the R1model (C, γ, ε, number of SV – parameters of SVM explained in methodology part of the paper).

h_w [cm]	C	γ	ε	Number of SV	*R*1
–2.5	16.939	0.004	0.5	8	0.916
–56	9.6417	0.195	0.104	44	0.887
–209	29.045	0.0948	0.104	50	0.891
–558	9.155	0.08	0.129	45	0.870
–976	29.837	0.096	0.097	59	0.879
–3060	5.128	0.239	0.085	63	0.876
–15300	3.88	0.099	0.082	66	0.856

Conclusions

The results of this paper contain a description and evaluation of the models of an ensemble of multi-layer perceptrons and an ensemble of support vector machines for the development of pedotransfer functions for the point estimation of the soil-water content for the seven pressure head values h_w from the basic soil properties (particle-size distribution, bulk density). Both ensemble data-driven models were compared to a multiple linear regression methodology.

- The accuracy of the predictions was evaluated by the correlation coefficient (R) between the measured and predicted parameter values. The R varied from 0.866 to 0.907 for the multi-linear regression for various pressure heads, from 0.905 to 0.930 when using MLP, and from 0.871 to 0.913 for the SVM. The MLP models perform somewhat better than the SVM models. Nevertheless, the results from both data-driven models are quite close, and the results show that they provide a more precise outcome than traditional multi-linear regression.
- Although SVM training is faster, the whole process of ANN training for evaluating PTFs is accomplished in less time, because of the ability of ANNs to produce more outputs (Fig. 1), which is the advantage versus SVMs.

Because other authors have reported the better regression ability of SVMs compared with ANNs (*Lamorski* et al, 2008; *Twarakawi* et al., 2009), the authors of the present paper hypothesize that it is advisable to use combined SVM/MLP models, because of this variability in suitable methodology. This should be verified in future work. The authors of the mentioned papers worked with larger data sets (they used 806 and 2134 soil samples; 140 samples were used in our work), and the influence of the amount of data or other statistical data set properties on the choice of the methodology suitable to use should be evaluated.

Acknowledgement. This contribution is the result of the implementation of the project: Centre of Excellence of Integrated Flood Protection of Territory ITMS 26240120004, supported by the Research & Development Operational Programme funded by the ERDF. This work was also supported by the Slovak Research and Development Agency under Contract No. LPP-0319-09, APVV-0139-10, APVV-0496-10 and VEGA 1/0243/11 and 1/1044/11.

REFERENCES

BOUMA J., 1989: Using Soil Survey Data for Quantitative Land Evaluation. Adv. Soil Sci., 9, 177–213.

BOUMA J., VAN LANEN A.J., 1987: Transfer Function and Treshhold Values: From Soil Characteristics to Land Qualities. K.J. Beek, et al. (eds.) Quantified land evaluation. Proc Worksh. ISSS and SSSA, Washington, D.C. 106–110.

CHANG C.C. and LIN C.J., 2001: LIBSVM: A Library for Support Vector Machines (2001). (Version 2.91, April 2010). Software, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

ČISTÝ M.: Application of the Harmony Search Optimization in Irrigation. In: Recent advances in harmony search algorithm. Berlin: Springer Verlag London, 2010. ISBN 978-3-642-04316-1, 123–134.

GEEM Z.W., KIM J.H. and LOGANATHAN G.V., 2001: A New Heuristic Optimization Algorithm: Harmony Search. Simulation, 76, 60–68.

GUPTA S.C., LARSON W.E., 1979: Estimating soil water retention characteristics from particle size distribution, organic matter percentage, and bulk density. Water Resour. Res., 15, 1633–1635.

KUMAR B., SREENIVASULU G., RAMAKRISHNA A. RAO, 2010: Radial Basis Function Network Based Design of Incipient Motion Condition of Alluvial Channels with Seepage. J. Hydrol. Hydromech., *58*, 2010, 2, 102–113.

MINASNY B., MCBRATNEY A.B., 2002: The neuro-m methods for fitting neural network parametric pedotransfer function. Soil Sci. Soc. Am. J., 66, 352–361.

MINASNY B., MCBRATNEY A.B., BRISTOW K.L., 1999: Comparison of different approaches to the development of pedotransfer functions for water retention curves. Geoderma, 93, 225–253.

PACHEPSKY YA.A., TIMLIN D.J. and VARALLYAY G., 1996: Artificial neural networks to estimate soil water retention from easily measurable data. Soil Sci. Soc. Am. J., 60, 727–733.

RAWLS W.J., BRAKENSIEK D.L., SAXTON K.E., 1982: Estimating soil water retention properties. Trans. ASAE, 25, 1316–1320.

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.Th., 1998: Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci. Soc. Am. J., 62, 847–855.

SKALOVÁ J., 2001: Pedotransfer functions of the Záhorská nížina soils and their application to soil-water regime modeling. (In Slovak.) Faculty of Civil Engineering STU Bratislava, 112 pp.

SKALOVÁ J., JAROŠ B., NOVÁK V., 2009: The Influence of Different Canopies on Groundwater Table Level Changes at Kláštorské Lúky Natural Reserve. J. Hydrol. Hydromech., *57*, 2009, 4, 276–285.

ŠTEKAUEROVÁ V., SKALOVÁ J., 1999: Calculation of the drying brunch of water retention curves from easily measured soil properties. (In Slovak.) VII. Poster Day. UH SAS Bratislava, 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 wa-

ter regime of plain area. (In Slovak.) ÚH SAV, Michalovce, 151–157.
TAMARI S., WOSTEN J.H.M., and RUIZ-SUAREZ J.C., 1996: Testing an artificial neural network for predicting soil hydraulic conductivity. Soil Sci. Soc. Am. J., 60, 1732––1741.
VAPNIK V., 1995: The Nature of Statistical Learning Theory. Springer, NY.
VAPNIK V., 1998: Statistical Learning Theory. Wiley, 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.

Received 22 October 2010
Accepted 19 October 2011