Determining Soil – Water Content by Data Driven Modeling when Relatively Small Data Sets are Available

Milan Cisty[1],

[1] Slovak University of Technology Bratislava, Faculty of Civil Engineering,
Radlinskeho 11, Bratislava 813 68, Slovak Republic
Milan.Cisty@stuba.sk

Abstract. 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. 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 performance of the models are evaluated and compared in case study for the Zahorska Lowland in the Slovak Republic. 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.

Keywords: soil water regime, pedotransfer function, neural networks, support vector machines, harmony search.

1 Introduction

Modeling water content and transport in soil has become an important tool in simulating agricultural productivity as well as in solving various hydrological tasks. For instance, optimum irrigation management requires a systematic estimation of the soil-water status to determine both the appropriate amounts and timing of irrigation. That is why soil characteristics appear as a important input in the numerical simulation of a soil-water regime. 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. [1], [2], [4], 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 [1]. 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 developing PTFs, e.g., [1], [5], [7], [10], etc.). Authors of above works confirm that they received better results from ANN-based pedotransfer functions than from standard linear regression-based PTFs.

Artificial neural networks include the ability to learn and generalize from examples with the aim of providing meaningful solutions to the problems to be solved. This process is called "training". When the training of an ANN is accomplished with a set of input and output data, the aim is to adjust the parameters of the ANN and make the ANN also provide corresponding outputs for other sets of input data (for which the outputs are not known).

Also second data driven method was used in this study- support vector machines (SVMs), which were developed by Vapnik [12] and are gaining in popularity due to their attractive features and promising empirical performance. The formulation embodies the structural risk minimization principle in addition to the traditional empirical risk minimization principle employed by conventional neural networks. 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 compare abovementioned methods while developing PTFs for the Zahorska Lowland in Slovakia, 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).

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

2 Methodology

The first approach for modeling the PTFs used in this paper is the application of *artificial neural networks* (ANNs). This approach has been described e.g. in [4] or [7]. Briefly summarized, a neural network consists of input, hidden and output layers, which contains processing elements. The number of processing elements in the input layer and output layer correspond to the number of input (e.g., the soil's bulk density, the soil's particle size data, etc.) and output variables of the model. So-called "learning" involves adjustment of the synaptic connections that exist between the neurons or weights in hidden layer, which are used for the transformation of the inputs to the outputs. 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 this study. The training process was performed by back propagation training algorithm of an MLP. 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 second approach called *support vector machines* (SVM) for estimating the pedotransfer functions used in this study is explained hereinafter, with brief explanations of its principles. A more detailed description of the methodology can also be found in available sources, e.g., in [9].

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 original input-output space. The linear regression is maintained by quadratic programming, which ensures a global optimum and an optimal generalization. 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. The uniqueness of solution produced by SVMs 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 process of selection of them will be described later.

Objective function of mentioned linear regression in feature space SVM 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 form of objective function is that a better ability to generalize could be expected (by choosing appropriate parameter *C*), compared, e.g., with ANNs, because unnecessarily complex models usually suffer from over fitting.

The radial basis function was chosen on a trial and error basis as the kernel function for this work. This function has the following form:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \quad \gamma > 0 \ . \tag{1}$$

The parameter γ of this kernel function, the tube size ε for the ε-insensitive loss function, and parameter *C* should be found, which the basic task is when SVMs are applied to particular task. The *harmony search methodology (HS)* was used for this purpose instead of the usual trial-and-error principle. A harmony search is a metaheuristic search algorithm introduced by Geem [3] 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 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).

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.

3 Study Area and Data Collection

The data used in this study were obtained from a previous work [8]. An area of the Zahorska Lowland in Slovakia was selected for testing the methods described. A total

of 140 soil samples were taken from various localities in this area.

The soil samples were air-dried and sieved for a physical analysis. A particle size analysis was performed utilizing Cassagrande's methods. 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.

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), testing data (30 data samples).

A practical way to find a better generalization data driven model is to set aside a small percentage of the training set and use it for the cross validation. When the error in the validation set increases, the training should be stopped. The five divisions of the data used to the training and validation data set were used.

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. 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 (average value was taken as result). 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.

4 Results

The first 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.

Table 1. Regression coefficients (R1-5) of five ANN ensemble models and resulted R

h_w [cm]	R1	R2	R3	R4	R5	R
-2.5	0.914	0.921	0.937	0.934	0.888	0.930
-56	0.905	0.922	0.934	0.897	0.872	0.916
-209	0.886	0.92	0.934	0.897	0.871	0.912
-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	0.858	0.92	0.93	0.846	0.848	0.892
-15300	0.192	-0.270	-0.281	-14.718	54.796	0.864

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 five regression coefficients are summarized in Table 1 also with final (average) regression coefficient R.

Given regression problem was also solved by using *ensemble* of *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.1); 2) specifying the ε parameter and specifying the capacity *C*.

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.

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 because there is seven variables computed. 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 Table 2.

Table 2. Regression coefficients (R1-5) of five SVM ensemble models and resulted R

h_w [cm]	R1	R2	R3	R4	R5	R
-2.5	0.910	0.908	0.902	0.910	0.891	0.907
-56	0.865	0.859	0.863	0.848	0.852	0.861
-209	0.860	0.857	0.867	0.862	0.855	0.863
-558	0.856	0.854	0.857	0.857	0.848	0.857
-976	0.833	0.842	0.846	0.840	0.834	0.846
-3060	0.845	0.847	0.852	0.846	0.837	0.851
-15300	0.799	0.822	0.819	0.815	0.817	0.816

A multi-linear regression for assessing the PTFs was accomplished for comparison and it was used in the form:

$$\theta_{hw} = a*1^{st}\ \text{cat.} + b*2^{nd}\ \text{cat.} + c*3^{th}\ \text{cat.} + d*4^{th}\ \text{cat.} + e*\rho_d + f\ . \qquad (2)$$

where θ_{hw} is the water content [$cm^3.cm^{-3}$] for the particular pressure head value h_w [cm]; *1st cat., 2nd cat., 3th cat.* and *4th cat.* are the percentages of the clay (*d*< 0.01 mm), silt (0.01–0.05 mm) and sand (0.1–2.0 mm); ρ_d is the dry bulk density [$g.cm^{-3}$]; and *a, b, c, d, e, f* 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 and 30 samples for testing.

The PTFs designed were evaluated on a testing dataset. The results of the multi-linear regression are listed in Table 3.

Table 3. Results of the multi-linear regression (coefficients of Eq. 2 and regression coeficient)

hw [cm]	a	b	c	d	e	f	R
-2.5	-0.224	-0.427	-0.2728	-0.403	-37.251	133.466	0.888
-56	-0.803	-1.157	-0.905	-1.223	-25.193	180.218	0.798
-209	-0.794	-1.163	-1.060	-1.2781	-17.761	167.584	0.819
-558	-0.531	-0.932	-0.833	-1.024	-19.859	143.738	0.856
-976	-1.7	-2.091	-1.984	-2.174	-18.405	255.177	0.685
-3060	-1.166	-1.624	-1.499	-1.655	-1.166	199.374	0.770
-15300	-2.007	-2.459	-2.288	-2.487	-14.882	275.784	0.631

As can be seen, the results from both data driven techniques are clearly better compared with the multi-linear regression and from SVM are 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 [11]. It should be mentioned that the authors of mentioned paper worked with larger data sets (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.

5 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 0.631 to 0.888 for the multi-linear regression, from 0.864 to 0.930 for the MLP, and from 0.816 to 0.907 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 significantly 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 in one run, which is the advantage versus SVMs.

Because other authors have reported the better regression ability of SVMs compared with ANNs [11], 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 paper worked with larger data sets (they used 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. Also other types of data driven models should be tested, e.g., generalized regression neural network or radial basis network.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under Contract No. LPP-0319-09 and APVV-0496-10 and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grants No. 1/1044/11.

References

1. Baker, L., Ellison, D.: Optimisation of pedotransfer functions using an artificial neural network ensemble method. Geoderma, 144 (1-2), 212-224 (2008)
2. Bouma, J.: Using Soil Survey Data for Quantitative Land Evaluation. Adv. Soil Sci., 9, 177--213 (1989)
3. Geem, Z. W., Roper, W. E.: Various Continuous Harmony Search Algorithms for Web-Based Hydrologic Parameter Optimisation, International Journal of Mathematical Modelling and Numerical Optimisation, vol. 1, 213-226 (2010)
4. Minasny, B., McBratney, A. B.: The Neuro-M Methods for Fitting Neural Network Parametric Pedotransfer Functions, Soil Sci. Soc. Am. J., vol. 66, 352-361 (2002)
5. Mohammadi, J.: Testing an artificial neural network for predicting soil water retention characteristics from soil physical and chemical properties. 17th World Congress of Soil Science. Thailand (2002)
6. Pachepsky, Y., and Rawls, W.J.: Developmennt of Pedotransfer Functions in Soil Hydrology. Elsevier (2004)
7. Schaap, M.G., Leij F.J, Van Genuchten M.Th.: Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci. Soc. Am. J., vol. 62, 847-855 (1998)
8. Skalova, J.: Pedotransfer Functions of the Zahorska Lowland Soils and Their Application to Soil-Water Regime Modeling. Thesis, Faculty of Civil Engineering STU Bratislava, 112 pp. (2001) (in Slovak)
9. Smola, A.J. and Schölkopf, B.: A Tutorial on Support Vector Regression. In: Statistics and Computing, vol. 14, 199-222 (2004)
10. Tamari, S., Wosten, J.H.M., and Ruiz-Suarez, J.C.: Testing an Artificial Neural Network for Predicting Soil Hydraulic Conductivity. Soil Sci. Soc. Am. J., vol. 60, 1732--1741 (1996)
11. Twarakavi, N. K. C., Simunek, J. and Schaap, M. G..: Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters Using Support Vector Machines. Soil Sci. Soc. Am. J., vol. 73, 1443–1452 (2009)
12. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, NY (1995)