# Application of the Stacking Regression in Context of the Soil-Water Modelling

Milan Cisty, Juraj Bezak, Jana Skalova
Department of Land and Water Resources Management
Slovak University of Technology
Bratislava, Slovakia
milan.cisty@stuba.sk, juraj.bezak@stuba.sk, jana.skalova@stuba.sk

***Abstract*— Modelling water transport in soil has become an important tool in simulating hydrological systems and agricultural productivity. Some of the data necessary for this modelling are usually easily available in competent institutions, but hydraulic soil properties (namely water retention curve) are only rarely easily available. The aim of this paper is to contribute to solving this deficit by evaluating so-called pedotransfer functions by data-driven modeling methods. Multi-linear regression, artificial neural networks, support vector machines and combination of these three methods in stacking model was evaluated. Work proves that stacking model yields more precise results than individual data-driven models and could be suggested for soil water modelling.**

***Keywords-soil-water modelling; pedotransfer function; data-driven model; stacking.***

## I. INTRODUCTION

Modelling water transport in soil has become an important tool in simulating hydrological systems and agricultural productivity. Models that deal with the transport of water and solutes range in scale from physically-based, fully distributed catchment models to the land parameterization scheme of general circulation models. Their practical application includes, e.g., systematic estimation of soil-water status to determine both the appropriate amounts and timing of irrigation. As is usual in any modelling, it depends on knowledge of the input data which are needed for the numerical simulations. Some of the data necessary for modelling water transport in soil (meteorological, climatic, hydrological or crop characteristics) are usually easily available in competent institutions, but hydraulic soil properties are only rarely easily available. These characteristics are therefore a key problem in the numerical simulation of a soil-water regime, and a modeller must deal with the problem of how to obtain them. The aim of this paper is to contribute to solving this task.

The water retention curve is one of the main soil hydraulic properties, which is used in simulating the water regime of soils. It represents the relationship between the water content and the soil's water potential (the potential energy of water per unit volume, which quantifies the tendency of water to move from one place to another). This curve is characteristic of different types of soil. It is used to predict a soil's water storage, the water supply to plants, and for other tasks in soil water modelling. A relatively large number of works have appeared in the past which were devoted to determining the water retention curve from more easily available soil properties such as particle size distribution, dry bulk density, organic C content, etc., e.g., [1][2][3]. In this context, Bouma [4] introduced the term "pedotransfer function" (PTF), which he described as "translating data that we have (soil survey data) into data that we need (soil hydraulic data)." In this paper, we will focus on point estimation methods of the PTFs, which follow the direct approach by estimating the water content at predetermined pressure heads.

Besides the application of the standard regression methods for solving this task, data-driven techniques appeared in the scientific literature in the second half of the previous decade as a tool for solving regression tasks in developing PTFs. However, there is no overall best data-driven technique which could be used in building hydrology models, because their suitability depends on the details of the problem, the data structure, the input data used, etc. For this reason various data-driven techniques are compared in this case study.

In the following part of the paper, the methods used in this study 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 "Conclusion" of the paper evaluates these experiments on the basis of the statistical indicators.

## II. METHODS USED TO FIT THE PEDOTRANSFER FUNCTIONS (PTFS)

The first approach for modelling the PTFs used in this paper is the application of *artificial neural networks* (ANNs). Briefly summarized, a neural network consists of input, hidden and output layers, all containing neurons. 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" or "training" 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 back propagation

 ISBN: 978-1-61208-237-0 

training algorithm. 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.

The basic idea behind the second methodology applied – *support vector regression regression* - 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 which is to be solved. The results of the regression are then mapped back to the input space. The kernel trick is a mathematical tool which can be applied to any algorithm which solely depends on the dot product between two vectors. Wherever a dot product is used, it is replaced by a kernel function. However, because kernels are used, the function never needs to be explicitly computed. This is highly desirable, because this higher-dimensional feature space could be unfeasible to compute.

The next important concept in SVM methodology is to fully ignore small errors (by introducing the variable ε, which defines what the "small" error is) to make the regression task dependent on a smaller number of inputs than were given in the original task, which makes the methodology much more computationally treatable. These crucial vectors of the inputs are called the support vectors.

In an ε-SVM regression [5], 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:

$$f(x) = w.\Phi(x) + b \quad w \in X, b \in R \tag{1}$$

where *f(x)* is the model's output, and input *x* is mapped into a feature space by a nonlinear function Φ(*x*) with the weight vector *w* and bias *b*.

The goal of a regression algorithm is to fit a flat function to the data points. "Flatness" means that one seeks a small w. One way to ensure this flatness is to minimize the norm, i.e. $\|w\|^2$. Thus, the regression problem can be written as a quadratic optimization problem:

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

$$\text{subject to: } y_i - (w.\Phi(x) + b) \le \varepsilon + \xi_i$$
$$(w.\Phi(x) + b) - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0$$

where ξi, ξi* are slack variables that specify the upper and lower training errors, subject to an error tolerance ε (soft margin), and C is a positive constant that determines the degree of the penalized loss when a training error occurs. In Equation system (2), 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 trade-off 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 over-fitting.

The third approach applied is to build the ensemble of the data-driven models, is so-called stacking model based on the base learners described above. Approach evaluated in this study is to generate ensemble model by applying different learning algorithms contained in the ensemble (other ensemble schemes, e.g., bagging or additive regression usually consist of one type of model). This approach to ensemble modelling deals with the task of training a meta-level base model to combine the predictions of multiple base-level base models. In other words, stacking introduces the concept of 1) base models and 2) a meta model, which computes the final results and replaces the averaging procedure used, e.g., in bagging. In such a way, stacking tries to learn which base models are more reliable than others, using mentioned meta-model (it could be a different algorithm than the base models) to discover how best to combine the output of the base models to achieve the final results. The results of the base learners are de facto new data for another learning problem, and in the second step a meta learning algorithm is employed to solve this problem. Variant of this approach is described on Fig. 1, where SVM is abbreviation for support vector machines, ANN is artificial neural network and MLR is multiple-linear regression – which are models from which stacking ensemble model consist from in our study.

```
D = {(x1, y1), (x2, y2) ... (xm, ym)};   % input data set, xi is vector
Base learning algorithms: A = {SVM, ANN, MLR}
Meta learner: SVM
TRAINING:
For j=1:3 do          % Train a base learner hj by applying
    hj = Lj(D)        % corresponding algorithm Aj
 end
For i=1: m
    For j=1:3
    zit = hj(xi)      % Use hj to predict the training example xi
    end
D' = {((zi1, zi2, zi3),yi)}
                          end;
h'=L (D')             % Train the meta learner h' (SVM)
                      % applying it to data set D'
                      % cross-validation was used
                      % to optimize its parameters
TESTING:              H (x) = h' (h1(x), h2(x), h3 (x))
```

Figure 1. Stacking algorithm scheme

## III. STUDY AREA AND DATA COLLECTION

The data used in this study were obtained from a previous work [6]. An area of the Zahorska lowland was selected for testing the methods described. A total of 226 soil samples was taken from various localities in this area.

The soil samples were air-dried and sieved for a physical analysis. A particle size analysis according to four grain

categories was performed utilizing Cassagrande's methods. Category I means the percentages of the clay (diameter < 0.01 mm), category II - silt (0.01–0.05 mm), category III - fine sand (0.05–0.1 mm) and category IV - sand (0.1–2.0 mm). 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 PTFs for the pressure head values of -2.5, -56, -209, -558, -976 and -3060 cm were estimated using overpressure equipment (set for pF-determination with ceramic plates).

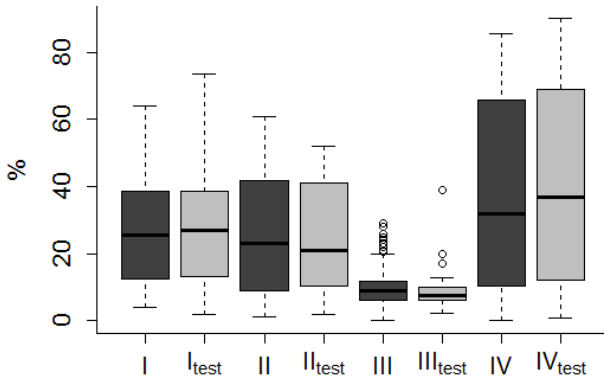


Figure 2. Comparison of grain categories (I, II, III, IV) in the training and testing data

A full database of the 226 samples and their properties were used for creating the input data for the modelling from which the training and testing subsets of the data were produced. The training data consist of 181 data samples and test data from 45 data samples. Statistically similar data should be in both data subsets; this condition is visualized by the boxplots on Fig. 2. In this figure, I, II, III and IV are grain categories in training set of data and the same identification with the subscript "test" is used for the test set. From this evaluation, it can be seen that category III will probably have the lowest impact on the pedotransfer function evaluation, but it will be included in the input data, anyway.

## IV. RESULTS

### A. Artificial neural networks

The first approach applied to determining the water retention curves in the presented work was the *artificial neural networks* methodology (ANN). In this work a multilayer perceptron with 4, 5, and 6 neurons in the hidden layer was tested; an ANN with 5 neurons in the hidden layer was finally chosen for the final neural network model used in the comparisons (it has the best results). A neuron with a hyperbolic tangent activation function was used in the hidden layer and a linear activation function in the output layer. The Levenberg-Maquardt method was used in the context of the back propagation method. The networks were trained to compute the water content at the pressure head value $h_w$ = -2.5, -56, -209, -558, -976, -3060 cm. The "hold-out" method was used for stopping the ANN to avoid overtraining, and this "hold-out" sample was 20% of the data from the training set.

Then the testing dataset was computed with the trained ANNs. The results with the regression coefficients are summarized in Table I. Three variants of the ANN with different hidden layer sizes and SVM are evaluated ($h_w$ - pressure head, H4 – H6 is the number of neurons in the hidden layer).

### B. Support vector machines

For a comparison with the ensemble approach, the given regression problem was also solved using *support vector machines (SVM)*. The estimation of the practical steps of the SVM regression are as follows: 1) selecting a suitable kernel and the appropriate kernel's parameter; 2) specifying the $\varepsilon$ parameter (2); and 3) specifying the capacity $C$ (2).

The radial basis function was chosen as the kernel function on a trial and error basis for which parameter $\gamma$ should be specified. The cross-validation methodology with 10 folds was used for finding the mentioned parameters of the SVM model.

In the training phase, SVM models for computing the water content for the pressure head values of $h_w$ = -2.5, -56, -209, -558, -976 and -3060 cm were created (on the basis of the particle size distribution as in the multi-linear regression case). Then the testing dataset was computed with the models obtained, and the final results were summarized with the help of the regression coefficients in Table I. The calculations of the SVM were performed using the LIBSVM library developed by Chang and Lin [8].

TABLE I. CORRELATION OF THE MODEL'S RESULTS WITH THE ACTUAL VALUES OF THE PTFS.

| $h_w$ [cm] | ANN – H4 | ANN – H5 | ANN – H6 | SVM | Stacking |
|---|---|---|---|---|---|
| -2.5 | 0.874 | 0.883 | 0.879 | 0.872 | 0.881 |
| -56 | 0.846 | 0.857 | 0.849 | 0.872 | 0.905 |
| -209 | 0.874 | 0.874 | 0.866 | 0.898 | 0.898 |
| -558 | 0.866 | 0.872 | 0.873 | 0.896 | 0.904 |
| -976 | 0.853 | 0.859 | 0.860 | 0.882 | 0.885 |
| -3060 | 0.833 | 0.846 | 0.852 | 0.880 | 0.890 |

### C. Stacking ensemble model

Stacked generalization (or stacking) is a way of combining the multiple models used in this work; it introduces the concept of a meta learner, the task of which is to combine the predictions of multiple base-level learners. In this work ANN, SVM and multi-linear regression were used as base learners. For stacking an ANN with four hidden units, the hyperbolic tangent activation function in the hidden layer and the linear function in the output layer were selected. The SVM as a base learner was not optimized (we did not include the parameter searching into the computational scheme), and the radial basis function kernel was chosen to maintain the nonlinearity. Parameter C was set as equal to the range of the output values [9], and parameter $\varepsilon$ in the $\varepsilon$-insensitive loss function was set to its default value 0.1 [7]. However, support vector regression was also used in the stacking model as a meta-learner. The SVM built a stacked model on top of the predictions of the base learners.

 ISBN: 978-1-61208-237-0 

In this case, its parameters γ, C and ε were optimized by tenfold cross-validation. The schema of this approach is in Fig. 1. The results with the regression coefficients are summarized in Table I.

From the results expressed by the correlation coefficient in Table I it can be seen that the stacking ensemble methodology evaluated in this case study give better results than when individual learners are used solo (ANN, SVM). Also, the application of the linear regression to the development of the pedotransfer function was evaluated in this work. Its main advantages are simplicity of implementation and interpretability; on the other hand, its shortcoming is that if the relationship between the input and output cannot be reasonably approximated by a linear function, the model will give poor predictions. This was also confirmed in this case study; the results in Table II, which were obtained by multi-linear regression, are generally worse than the results of the ANN and SVM alone (Table I) and significantly worse than the results obtained by stacking ensemble methodology evaluated in this work.

A more detailed evaluation of the various data-driven methods applied in this work is presented in Table II. For practical reasons (the limited extent of this paper) it is restricted only to an evaluation of the prediction of the water content for the pressure head value $h_w$ = -3060 cm. The results for the other pressure heads are similar from point of view of effectiveness of the algorithms used. In Table II the mean error (ME), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), normalized root mean square error (NRMSE), percent bias (PBIAS), correlation coefficient (r), maximal difference between the simulated and actual values (maxD) and the minimal difference between the simulated and actual values (minD) are evaluated. The names of the models in the heading of Table II are clear from their abbreviations. From this analysis, it is evident that it is worthwhile to pay attention to the development and choice of the proper regression model when evaluating the pedotransfer function, because it can be seen that a relatively big difference is between the effectiveness of the worst performing model (MLR) and the best model. The models are ordered in columns according to their quality from worst to best.

TABLE II. EVALUATION OF THE VARIOUS MODELS FOR PREDICTION OF THE WATER CONTENT AT $h_w$ = 3060 CM BY DIFFERENT STATISTICS

| | **MLR** | **ANN** | **SVM** | **Stacking** |
|---|---|---|---|---|
| ME | -0.39 | -0.53 | -1.02 | -0.45 |
| MAE | 4.21 | 3.75 | 3.31 | 3.32 |
| MSE | 30.40 | 25.98 | 21.44 | 19.73 |
| RMSE | 5.51 | 5.10 | 4.63 | 4.44 |
| NRMSE | 57.50 | 53.10 | 48.30 | 46.30 |
| PBIAS | -1.80 | -2.40 | -4.70 | -2.10 |
| r | 0.82 | 0.85 | 0.88 | 0.89 |
| maxD | 9.83 | 7.12 | 6.24 | 6.40 |
| minD | -15.27 | -15.02 | -12.91 | -11.88 |

## V. CONCLUSION AND FUTURE WORK

This paper proposed and evaluated data-driven models for the development of pedotransfer functions for the point estimation of the soil-water content for six pressure head values $h_w$ from the basic soil properties (particle-size distribution, bulk density). The ensemble data-driven model (stacking) was compared to single data-driven models (artificial neural networks and support vector machines) and to a multiple linear regression methodology. The accuracy of the predictions was evaluated by the correlation coefficient between the measured and predicted parameter values and by other statistics. From the results obtained it was proved that nonlinear data-driven methods work significantly better than multi-linear regression and that even better results were obtained by using data-driven methods in an ensemble context.

However, several issues remain to be addressed by further research. Although in this work stacking performs well, it is not easy to give the reasons for selecting its particular components. This process is subjective in the present state of our knowledge (on the basis of trial and error), which should be improved in the future.

## ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under Contract No. LPP-0319-09, and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/1044/11 and 1/0243/11.

## REFERENCES

[1] S.C. Gupta and W.E. Larson, “Estimating soil water retention characteristics from particle size distribution, organic matter percentage, and bulk density”, Water Resour. Res. 15, pp. 1633-1635, 1979.

[2] W.J. Rawls, D.L. Brakensiek and K.E. Saxton, “Estimating soil water retention properties”, Trans. ASAE 25, pp. 1316-1320, 1982.

[3] B. Minasny, A.B. McBratney and K.L. Bristow, ”Comparison of different approaches to the development of pedotransfer functions for water retention curves”, Geoderma, 93, pp. 225–253, 1999.

[4] J. Bouma, “Using Soil Survey Data for Quantitative Land Evaluation”, Adv. Soil Sci., 9, pp. 177–213, 1989.

[5] V. Vapnik, ”The Nature of Statistical Learning Theory”, Springer, NY, 1995.

[6] J. Skalova, ”Pedotransfer functions of the Zahorska Lowland soils and their application to soil-water regime modelling”, Faculty of Civil Engineering STU Bratislava, (in Slovak), 2001.

[7] I.H. Witten, E. Frank and M.A. Hall, “Data mining”, Morgan Kaufmann Publishers, 2011.

[8] C.Ch. Chang and C.J. Lin, ”A library for support vector machines”, 2001. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf [retrieved: August, 2012]

[9] V. Cherkassky and Y. Ma, ”Practical selection of SVM parameters and noise estimation for SVM regression”, Neural Networks 17(1), pp. 113-126, 2004.

 ISBN: 978-1-61208-237-0 