Evaluation of Water Retention Curves by Regression and Machine Learning Methods

Milan Cisty, Barbora Povazanova

Faculty of Civil Engineering, Slovak University of Technology in Bratislava, Bratislava, Slovakia
milan.cisty@stuba.sk

Abstract. The paper presents two methods that simplify the estimation of the water retention curves. The case study is evaluated for the soils of Záhorská lowland in the paper. These methods are based on the supposed dependence of the soil water content on the percentage content of the 1st, 2nd, 3rd and 4th Kopecký grain categories, and the dry bulk density. The representative set of the drying branch of water retention curves was measured using soil samples from the Záhorská lowland region in a laboratory. Particle size distribution and dry bulk density were also determined. In this paper support vector machines and multiple linear regression is compared to estimate the pedotransfer functions that can be used for the prediction of the drying branch of the water retention curve. Both methods were verified on other data set of measured water retention curves than the one which was used for building the models with a close agreement to measured results.

1. Introduction
While solving various ecological and agricultural studies and projects in the landscape, an important variable that determines the properties and processes of the environment is soil moisture [1]. Modelling of this variable is used as an essential tool in, e.g., land drought management or groundwater issues [2]. Characteristics of soil water regime are possible to determine by two methods. Generally accepted method is a method of the direct monitoring of soil moisture in a given area [3]. However, complete monitoring of soil moisture on agricultural soils in Slovakia wasn’t done. The situation in many other countries is often similar and usually, only some information published in research studies and scientific works related to irrigation, ecology, hydrology and pedology incompletely maps the situation. Monitoring of the soil water regime characteristics produces a knowledge base for the assessment and optimization of soil water regime (SWR) in agricultural land, e.g., for increasing crop production, but it is time consuming activity [4]. The alternative method for the determination of SWR could be achieved by mathematical modelling. Using models depends on the availability of input data. Some input data - meteorological, climatic, hydrologic or crop characteristics are usually well available in corresponding institutions. On the opposite, the soil water properties appear as the key problem in the numerical simulation of SWR. From data needed, mainly the water retention curve is usually not available.

Measurement of the water retention curve (WRC) points, its drying or wetting branch in the laboratory is quite expensive, time-consuming, and labour-intensive. During the last decades, some works were published, which are dedicated to the determination of WRCs without using laboratory procedures, based on various easier available soil properties, e.g., particle size distribution, dry bulk density, organic C content, for example [4,5]. This methodic approach, which is based on the supposed
dependence of the soil water content on various easier obtained characteristics uses regression techniques, such as linear regression. Obtained functions are named pedotransfer functions (PTFs). The other method that was used for the determination of WRCs is an artificial neural network [6].

For the so-called point estimation of the drying branch of WRC, we compared in this work multiple linear regression and machine learning method called support vector machines. The case study was accomplished for soils of the Záhorská lowland region.

2. Case study description
The pedotransfer functions can be used for the calculation of volume water contents \( \theta(h_w) \) by its so-called point estimation at the corresponding pressure head values. In this work we used pressures \( h_w = -2.5, -56, -209, -558, -976, -3060 \) cm. Computation of volume water contents \( \theta \) is based on the percentage content of the 1\(^{st}\), 2\(^{nd}\) and 4\(^{th}\) grain categories by Kopecký and the dry bulk density \( \rho_d \).

From various localities on Záhorská lowland, 140 soil samples were taken with known soil characteristics and were used for the determination of WRC points. Soil types of the Záhorská lowland region are presented in figure 1 and table 1. As we can see the sandy soils prevails in the investigated area.

![Figure 1. Soil types on Zahorská lowland.](image)

| % of 1. grain category for soil type by Novák | Soil samples | Area |
|---------------------------------------------|-------------|------|
| % of I. grain category for soil type by Novák | Number | % | % |
| 1. Sandy soil | 0-10% | 30 | 21 | 47 |
| 2. Loam sandy soil | 10-20% | 30 | 21 | 15 |
| 3. Sandy loam soil | 20-30% | 23 | 17 | 6 |
| 4. Loam soil | 30-45% | 50 | 36 | 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 |

**Table 1.** Classification of soil samples of Záhorská lowland by particle size distribution.
Correlation analysis of this data (with all 140 soil samples) was performed. Soil moisture was in this task as dependent variable and grain categories and dry bulk density were taken as independent variables. Results are in table 2. Based on these results could be seen that the amount of the 3rd cat. has the smallest influence on all pressure heads that documents the corresponding value of the correlation coefficient. Therefore, it was considered, that computations will use as explanatory variables only the 1st, 2nd and 4th grain categories of Kopecký classification and the dry bulk density ρ_d.

Review of correlation coefficients (r) between grain categories and bulk density ρ_d and suction pressure h_w is in table 2. The r testifies to a good degree of relationship between correlated elements, so these variables could be used for developing regression dependencies.

| Independent variables | Correlation coefficient r for h_w [cm] |
|-----------------------|----------------------------------------|
|                       | -2.5 | -56 | -209 | -558 | -976 | -3060 | -15300 |
| 1st cat.              | 0.5726 | 0.7739 | 0.8414 | 0.8386 | 0.8471 | 0.8476 | 0.8630 |
| 2nd cat.              | 0.2795 | 0.5696 | 0.6353 | 0.6025 | 0.6006 | 0.5526 | 0.5801 |
| 3rd cat.              | 0.0418 | 0.0890 | 0.0907 | -0.0173 | -0.0487 | -0.0948 | -0.0859 |
| 4th cat.              | -0.4448 | -0.7253 | -0.7783 | -0.7509 | -0.7465 | -0.7069 | -0.7328 |
| ρ_d                   | -0.8259 | -0.5580 | -0.4684 | -0.5213 | -0.5100 | -0.5215 | -0.4925 |

The full database from 140 samples was divided into two subsets of data:
1. Training data 75% (105 samples)
2. Test data 25% (35 samples)

In the following calculations, training data are used to create the model and test data were used to verify the accuracy of these models.

3. Methodology

Firstly, multiple linear regression (MLR) was applied to assess PTFs in form:

\[ \theta_{hw} = a \times 1^{st} \text{cat.} + b \times 2^{nd} \text{cat.} + c \times 4^{th} \text{cat.} + d \times \rho_d + e \]  \[ \text{[cm}^3 \cdot \text{cm}^{-3}] \]  \[ (1) \]

Where:

- \( \theta_{hw} \) is volume water content \([\text{cm}^3 \cdot \text{cm}^{-3}]\) in the concrete pressure head value \( h_w \) [cm],
- 1\text{st} cat. - percentage of clay (with diameter < 0.01 mm),
- 2\text{nd} cat. - percentage of silt (diameter 0.01–0.05 mm),
- 4\text{th} cat. - percentage of sand (diameter 0.1–2.0 mm),
- \( \rho_d \) - dry bulk density \([\text{g} \cdot \text{cm}^{-3}]\) and
- a, b, c, d, e are regression parameters determined by mean square error method.

The second approach to the determination of water retention curves in this work is support vector machines (SVM). Machine learning method support vector machine was proposed by Vapnik [7], as a statistical learning method with a promising ability to generalize [8]. 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 principle in addition to the traditional empirical risk minimization principle employed by conventional neural networks. Structural risk minimization minimizes the upper boundary on the risk expected, as opposed to empirical risk minimization, which only minimizes an error on the training data. It is this difference...
that gives SVMs a great ability to generalize, which is the goal of statistical learning. In addition, a support vector machine [9] are specific by using kernel trick - nonlinear mapping, which is used to transform the original training data of a nonlinear problem (which is also our task) into a higher dimension. They learn a linear function in the space induced by the kernel, which matches to a non-linear function in the original space.

4. Results and discussion

Two methods of WRCs assessment for the soils of Záhorská lowland were used: multiple linear regression and support vectors machines. It was applied to the data set, which contained 140 soil samples.

Results of multiple linear regression are listed in table 3. Designed PTFs were checking on testing dataset that contents of 35 soil samples.

| $h_w$ [cm] | a (I.cat) | b (II.cat) | c (IV.cat) | d ($\rho_d$) | e | r     |
|------------|-----------|------------|------------|--------------|---|-------|
| -2.5       | 0.02504   | -0.16559   | -0.14090   | -38.5748     | 109.780 | 0.89476 |
| -56        | 0.07266   | -0.24990   | -0.29580   | -27.3086     | 93.4649 | 0.85031 |
| -209       | 0.21626   | -0.16102   | -0.25044   | -19.0725     | 68.2862 | 0.86419 |
| -558       | 0.22592   | -0.17563   | -0.24695   | -21.0091     | 68.8893 | 0.86525 |
| -976       | 0.17952   | -0.21693   | -0.28093   | -19.9774     | 68.7116 | 0.84164 |
| -3060      | 0.25277   | -0.21897   | -0.22958   | -17.9315     | 58.3262 | 0.83101 |

A grid search combined with a repeated cross-validation methodology was used for finding the parameters of SVM. In this approach, a set of SVM parameters from a predetermined grid is sent to the parameter-evaluating algorithm. In this work, two parameters were optimized, namely C and sigma, which is a parameter of the RBF kernel which was used in this work. C is a regularization parameter that controls the trade-off between achieving a low training error and a low testing error that is the ability to generalize SVM to unseen data.

5 times repeated 10-fold cross-validation was used. The data set was divided into 10 subsets, and the training-testing-evaluation was repeated 10 times. Each time, one of the 10 subsets is used as the test set, and the other 9 subsets are put together to form a training set. Then the average error across all 10 trials is computed, and the case with the lowest errors determines the combination of the model parameters in an actual repetition. These parameters were used to train the final model. The division of the data into ten subsets was repeated differently five times [10]. All computations were accomplished in R statistical language [11].

The results of the SVMs obtained by this procedure, together with the model parameters for each suction pressure value, are given in table 4. From this table, we can see that all correlation coefficients are better (closer to 1) when using the SVM method than when using the MRL (table 3).

In figure 2 are graphically presented measured and calculated WRCs by both methods (MLR and SVM). In the next two figures also scatter plots are used, on which measured and calculated volume water content with MLR and SVM (figure 3) is compared. From these analyses results, that both methods can be used for the prediction of WRCs, but SVMs offers more precise results.
Table 4. Obtained parameters of SVM based pedotransfer functions (C, sigma) for calculation of points of retention curve for Záhorská nížina soils (\(h_w\) is pressure head, RMSE is the root mean square error, pbias is percentual bias, \(r\) is correlation coefficient).

| \(h_w\) [cm] | C   | sigma | RMSE | Pbias % | \(r\) |
|--------------|-----|-------|------|---------|------|
| -2.5         | 100 | 0.003 | 3.773| -2.3    | 0.912|
| -56          | 30  | 0.004 | 4.678| -0.7    | 0.925|
| -209         | 80  | 0.005 | 4.919| -0.6    | 0.911|
| -558         | 60  | 0.002 | 4.919| -0.6    | 0.915|
| -976         | 60  | 0.006 | 5.203| -1.5    | 0.899|
| -3060        | 50  | 0.002 | 5.297| -5.1    | 0.878|

Figure 2. Comparison of measured water retention curves (black) and calculated water retention curves by MLR (red) and by SVM (blue) on the example of 9 soil samples from the testing dataset.
basic input to mathematical models of soil water regime. Machines were tested against measured WRCs of soils of Záhorská lowland. Comparison of measured both regression methods, the dependence of the soil water content on the amount of 1st, 2nd, 3rd and 4th grain categories (according to Kopecky classification) and dry bulk density. In this paper, two methods are presented, which helps to simplify and speed up the assessment of WRCs, which are the basic input to mathematical models of soil water regime.

4th grain categories (according to Kopecky classification) and dry bulk density. In this paper, two methods are presented, which helps to simplify and speed up the assessment of WRCs, which are the basic input to mathematical models of soil water regime.

5. Conclusion
In this paper, two methods are presented, which helps to simplify and speed up the estimation of the water retention curves (WRCs). These methods - multiple linear regression and support vector machines, were applied on the soils of Záhorská lowland.

For creating of WRCs, the database of 140 soil samples from this area was used. This database consisted of WRCs points estimated in the laboratory, particle size distribution and dry bulk density. In both regression methods, the dependence of the soil water content on the amount of 1st, 2nd, 3rd and 4th grain categories (according to Kopecky classification) and dry bulk density for assessment of PTFs. Obtained PTFs for 6 points of WRCs determined by multiple linear regression and support vector machines were tested against measured WRCs of soils of Záhorská lowland. Comparison of measured and calculated values documents close agreement (tables 3, 4, figures 3, 4). From the analysis of the results could be evaluated, that assessment of drying branch WRC points it is possible by using both methods but SVM offers more precise results.

Using MLR and SVM significantly simplifies and speeds up the assessment of WRCs, which are the basic input to mathematical models of soil water regime.
Acknowledgements
This work was supported by the Slovak Research and Development Agency under Contract No. APVV-19-0383 "Natural and technical measures oriented to water retention in sub-mountain watersheds of Slovakia" and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/0662/19.

References
[1] M. Červeňanská, D. Baroková, and A. Šoltész, “Modeling the groundwater level changes in an area of water resources operations”, Pollack Periodica, vol. 11, no. 3, pp. 83-92, 2016.
[2] A. Šoltész, D. Baroková, Z. D. Shenga, and M Červeňanská, “Hydraulic assessment of the impacts of gate realization on groundwater regime”, Pollack Periodica, vol. 15, no. 3, p. 162-171, 2020.
[3] V. Novak, H. Hlavacikova, “Applied Soil Hydrology” Springer, Cham, ISBN: 978-3-030-01805-4, 2018.
[4] J. Skalova and V.Stekauerova, “Pedotransfer Functions and Their Application in the Modelling of a Soil Water Regime.” Bratislava: STU in Bratislava, pp. 101, 2011. ISBN 978-80-227-3431-8E (in Slovak)
[5] K. Lamorski, J.Simunek, C. Sławinski and J. Lamorska, “An estimation of the main wetting branch of the soil water retention curve based on its main drying branch using the machine learning method.” Water Resources Research, vol. 53(2), pp. 1539-1552, 2017.
[6] B. Minasny, J.W. Hopmans, T. Harter, A. Eching, A. Tuli, M.A. Denton “Neural Networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data,: Soil Sci. Soc. Am. J., 68, 417-429, 2004
[7] V. Vapnik, “The nature of statistical learning theory,” Springer, 1999.
[8] L. Wang, “Support vector machines: theory and applications”. Springer Science & Business Media, vol. 177, pp. 341, ISBN: 978-3-540-24388-5, 2005.
[9] M. Awad, R. Khanna, “Support Vector Regression”. In: Efficient Learning Machines. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_4, 2015
[10] T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction” Springer series in statistics, Springer, New York, ISBN: 978-0-387-84857-0, 2009.
[11] R Core Team: “R: a language and environment for statistical computing” R Foundation for Statistical Computing, Vienna, available at: https://www.R-project.org/, 2017.