PAPER • OPEN ACCESS

Evaluation of Linear and Machine Learning Models for Determining Pedotransfer Functions

To cite this article: Milan Cisty and Frantisek Cyprich 2020 *IOP Conf. Ser.: Earth Environ. Sci.* **609** 012083

View the article online for updates and enhancements.

Evaluation of Linear and Machine Learning Models for Determining Pedotransfer Functions

Milan Cisty[1], Frantisek Cyprich[1]

[1]Faculty of Civil Engineering, Slovak University of Technology in Bratislava,
Radlinskeho 11, 810 05 Bratislava, Slovakia

milan.cisty@stuba.sk

Abstract. Modelling of water content and the transport of water in soil has become a useful tool in simulating agricultural productivity or in solving various hydrological analyses. For instance, optimum irrigation management requires a systematic estimation of the soil-moisture to determine both the appropriate amount and timing of irrigation. Soil characteristics appear as an essential input in the numerical simulation of a soil-water regime. A critical physical property used in the description of a soil-water regime in such modelling is a soil water retention curve. This paper aims to evaluate so-called pedotransfer functions, which helps to assess the soil water retention curve easier than by standard complex and lengthy procedure involving both field and laboratory work. As a case study Zahorska Lowland, which is located in central Europe in the western part of Slovakia, was selected. The frequent occurrence of dry years in this area results in the necessity to construct irrigation systems in this area, so modelling water content in the soil is an important task here. This study aims to support such modelling with determining pedotransfer functions. Authors compare linear methods (multiple linear regression, LASSO regularized regression) and CatBoost machine learning model.

1. Introduction

When solving various ecological and agricultural tasks in the landscape, one of the main variables that determine 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 [2]. Some variables for modelling the climatic and soil characteristics of the landscape are relatively easy to obtain. For example, precipitations, temperatures, or evapotranspiration (which can be derived from the previous two) are almost always available. But, it is usually much more difficult to describe the characteristics of the heterogeneous soil environment of the landscape in such modelling. An essential characteristic or input of models used in the description of a soil-water regime is a soil water retention curve. It describes the relationship between the water content and the water potential of the soil. This dependence can be determined by measurements and laboratory, which is, however, relatively time-consuming. Even more challenging is to evaluate this characteristic by measuring for a larger area. For this reason, the concept of so-called pedotransfer functions (PTFs) was created, which simplifies/eliminate the mentioned laboratory procedure [3]. This paper deals with the point estimation of pedotransfer functions. Point estimation methods follow an approach of estimating water content at predetermined pressure heads.

PTFs indicate the dependence of the soil water content by using the easy available soil characteristics, e.g., the soil particle sizes in the soil or the dry bulk density of the soil. These data are very often already available in various geographic information systems. Soil water retention curve can be extracted from these data by different regression methods. In this paper, linear and machine learning methods were used to estimate a drying branch of a water retention curve by using concept of pedotransfer functions. Proposed methods are tested on the Zahorska lowland in the Slovak Republic. Presented computations show the superiority of machine learning methods over linear methods which is the main results obtained in this study.

In the next part of the paper ("Case Study and Data Description"), the acquisition and preparation of the data is described. The methods applied in this study are then briefly explained. In the "Results and Discussion" section, the computational experiments are described, as well the results are evaluated and discussed. Finally, the "Conclusion" part of the paper summarizes the main achievements of the work.

2. Case Study and Data Description

The area of interest – the Zahorska Lowland, is located in central Europe, in the western part of the Slovak Republic. It is bounded by the river Morava, and the Little Carpathians mountain. Most of the Zahorska Lowland territory is covered by clayed sands, drift sands and sandy clays [4]. The main pedogenetic factors of these lowlands are the accumulation activity of streams, soil-disrupting floods and soil erosion. The most widespread soils in this area are Chernozems, Arenosols, and fluvial soils on the river Morava's fluvial plains. Their profile is continuously loaded with layers of flood sediment sludge [5]. The spatial distribution of the different types of soil is shown in Figure 1 according to the USDA classification.

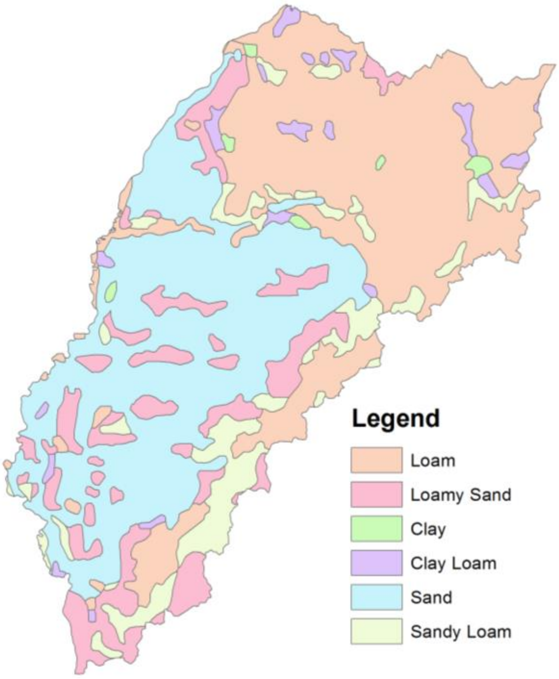

Figure 1. The spatial distribution of the Zahorska Lowland soil types according to the USDA classification

The soil samples which are used in this study for evaluation of PTFs were air-dried and sieved; soil particle distribution and other analyses were performed as well. After these analyses were accomplished, the data set contained the following parameters: four grain categories according to the Kopecky textural classification (which is used in Slovakia), reduced bulk density ρd, and the points of the drying branches of the water retention curve (WRC) for the pressure head values of -2.5, -56, -209, -558, -976, -3060 and -15,300 cm. The last variables were estimated using overpressure equipment.

3. Methods

The main goal of this work is to compare the existing regression methods for estimating of PTFs with new ones, proposed in this paper, and to choose the method, which is most suitable input for subsequent modelling of soil moisture. The standard method used for estimating PTFs is multiple linear regression. Multiple linear regression (MLR) search for the relevant coefficients in the equation of the linear model. In general, it can be written as follows:

$$Y = aX_1 + bX_2 + cX_3 + \cdots + X_n, \tag{1}$$

where Y denotes a dependent variable, X_n are independent variables. In linear regression, independent variables should not correlate too much. To remove multicollinearity in given problem, the Least Absolute Selection and Shrinkage Operator (LASSO) was used for the regularization of the linear regression. It was introduced in 1996 by R. Tibshirani [6] In a LASSO regularization, some coefficients are zeroed. Using this regularization, a simpler model is created, and a better generalization is provided.

As a representative of nonlinear machine learning models the CatBoost model [7] was chosen. Gradient boosting is a powerful machine-learning technique that achieves state-of-the-art results in a variety of practical tasks: web search, environmental variable predictions, spatial analysis of ecological factors distribution, such as distribution of contaminant concentration, weather forecasting, and many others [8, 9, 10].CatBoost is a machine learning algorithm that uses gradient boosting on regression trees [11] which builds a model in a stage-wise fashion, through increasingly refined approximations. Gradient boosting evaluates the precision of a model in a previous stage. It then develops the next model, which computes the differences between the current results computed and the known target values (i.e., not the original target values). Such "boosting" continues until the desired level of accuracy is reached. A fundamental building block in CatBoost is a regression tree – a type of tree in which decision node contains a test on some input variable's value, and the terminal nodes of the tree contain the predicted values. The cost for using CatBoost (compared, e.g., with a LASSO algorithm) is that CatBoost has more parameters to tune. In this work, the cross-validation method was used to find the parameters of this model. A description of the parameters and various recommendations for setting them can be found at CatBoost website [12].

4. Results and Discussion

For comparison, we firstly calculated points of PTFs using the standard method - multiple linear regression. Results obtained were statistically evaluated, so it will be possible to show whether the proposed methods described later in this chapter works better. A multi-linear regression for assessing the PTFs was used in the form:

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

where θ_{hw} is the water content [cm^3.cm^{-3}] for the particular pressure head value h_w [cm]; *1st cat., 2nd cat., 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*, are the parameters determined by the regression analysis. Fine sand (3rd cat.) was not included in regression equation because of avoiding correlation between independent variables.

The PTFs designed were from now on always evaluated on a testing dataset, which consisted of 50 soil samples. These samples were randomly selected from 180 samples that were measured at Zahorska Lowlands. Test samples were not used during models creation. The results of the multiple linear regression are listed in Table 1. The Coefficients of Determination (R2) for each of the pressure head values shows less degree of the relationship between the dependent and independent variables in some cases (with the smallest value of 0.671). This is why the authors tried to propose better regression methods for this task.

Table 1. Parameters of multiple linear regression (a, b,c, d, e) for calculation of points of drying branch of water retention curve (h_w - pressure head, R2 – Coefficient of Determination)

h_w [cm]	a	b	c	d	e	R2
-1.0	0.005	-0.163	-0.155	-35.10	105.2	0.736
-60	0.002	-0.258	-0.342	-23.02	90.01	0.670
-200	0.211	-0.125	-0.237	-14.72	59.71	0.701
-560	0.246	-0.115	-0.203	-17.65	59.42	0.740
-1000	0.256	-0.114	-0.188	-16.48	54.85	0.689
-3000	0.314	-0.093	-0.131	-14.68	44.40	0.700
-15300	0.275	-0.142	-0.171	-11.79	40.59	0.671

The next method used to solve this task was regression using LASSO regularization. The precision of the modelling using this method was enhanced by using the variable's interaction within the input data. Regarding the degree of interactions, the interaction of the three variables was used. The interaction was included by introducing new variables that are arithmetic products of the original variables. If we potentially consider all the possible products of three variables, many new combined variables were created. For this reason, the regularization by LASSO was very useful, which, as described in the methodological part, reduces the number of variables in the resulting model. This selection and reduction of a number of variables should ensure the accuracy and stability of the resulting model. This premise is tested in Table 2, which evaluates the model for calculating the water content for the pressure head value h_w of 1 cm. This table shows the usefulness of using variable interactions - as the table shows with the inclusion of interactions, the accuracy of the calculation increases (coefficient of determination (R2) increases, RMSE and percentage bias (PBIAS%) decrease). The models for the water content in the soil at all pressures (1, 60, 200, 560, 1000, 3000, 15300 cm h_w) are because of this finding assessed using only a triple interaction and are evaluated in Table 3. The results of the models in Table 3 can be compared with Table 1, where the basic linear regression was used. In the column for Coefficient of Determination (R) we see a higher, i.e., better value for all pressures h_w, which proves that LASSO is a more suitable model for calculating the points of the retention curve.

The functions from the R package *glmnet* were used for computation of the LASSO model [13]. The lambda regularisation parameter was optimized by cross validation, and its value is shown in the tables with results of modelling. An example of tuning this parameter for LASSO regression with a triple interaction of the explanatory variables is given in Figure 2.

Table 2. Three variants of LASSO regularized regression models for computing point of water retention curve for h_w 1 cm

model	RMSE	PBIAS%	R2	lambda
LASSO	3.415	0.500	0.728	0.0836
LASSO with double interactions	3.305	0.500	0.746	0.1156
LASSO with tripple interactions	3.092	0.300	0.779	0.0531

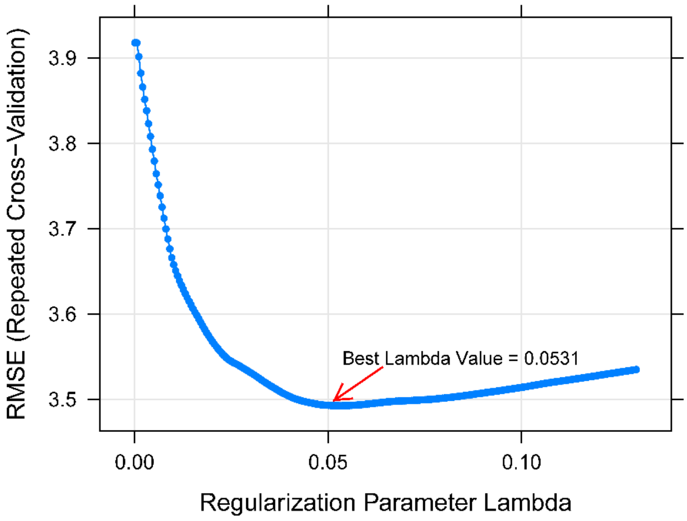

Figure 2. Example of tuning lambda parameter of the the LASSO with triple interactions using multiple cross-validation (for pressure h_w = 1 cm).

Table 3. Evaluation of models for computing points of drying branch of water retention curve by LASSO for various values of h_w

h_w [cm]	RMSE	PBIAS%	R2	lambda
-1.0	3.092	0.300	0.779	0.0531
-60	4.372	0.400	0.803	0.0290
-200	4.262	0.900	0.820	0.0230
-560	4.325	2.600	0.812	0.0690
-1000	4.798	4.300	0.758	0.1000
-3000	4.594	4.400	0.747	0.0660
-15300	4.703	5.300	0.709	0.0600

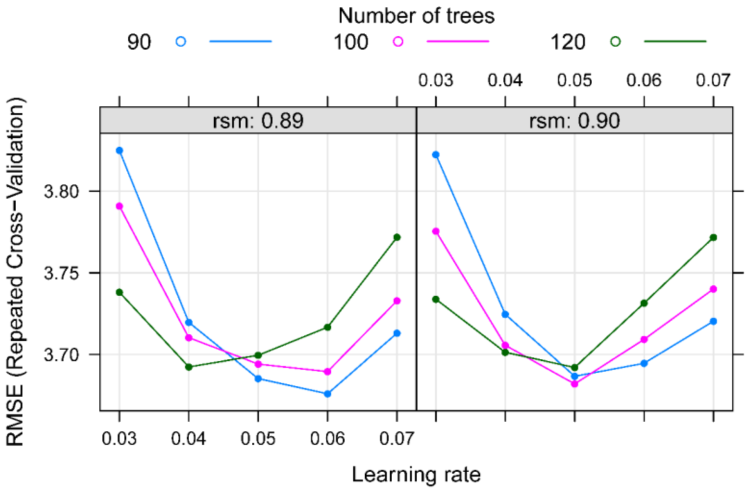

Figure 3. An Example of tuning the CatBoost parameters. The minimum value of RMSE and headings of the relevant panel of the plot indicate the optimum values of the parameters

CatBoost tuning is more complex than previous models, as it has more parameters. The following parameters have been tuned by cross-validation: 1) number of regression trees; 2) learning rate and 3) random subspace method (rsm) - the percentage of features to use at each regression tree split selection. This algorithm is known for its good default settings, so for other parameters, default settings have been used. An example of tuning these parameters is given in Figure 3. Point models for the water content in the soil at pressures 1, 60, 200, 560, 1000, 3000 and 15300 cm h_w are evaluated in Table 3. The CatBoost show better results even though interactions were not used, because interactions are already included in algorithm itself, which is based on regression trees.

Table 4. Evaluation of models for computing points of drying branch of water retention curve by CatBoost for various values of h_w

h_w [cm]	RMSE	PBIAS%	R2	random subspace method	Learning rate	Number of trees
-1.0	3.345	-0.600	0.785	0.89	0.060	90
-60	4.261	-0.200	0.829	0.90	0.050	100
-200	4.161	0.800	0.846	0.90	0.049	100
-560	4.235	1.100	0.832	0.90	0.053	105
-1000	4.654	3.400	0.792	0.90	0.052	105
-3000	4.574	4.200	0.774	0.89	0.050	110
-15300	4.793	5.900	0.731	0.89	0.052	120

5. Conclusions

The paper contains a description and evaluation of the models 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). Linear and machine learning model were compared with a standard multiple linear regression methodology. The accuracy of the predictions was evaluated by the Coefficient of Determination (*R2*) between the measured and predicted values. For the multiple linear regression and the various pressure heads the *R2* varied from 0.671 to 0.740, from 0.709 to 0.820 when using LASSO, and from 0.731 to 0.846 for the CatBoost.

The results show the suitability of the proposed methods, while the CatBoost method can be considered as the most suitable method from the point of view of its accuracy.

Nevertheless, innovative linear methods (with LASSO regularization), in applying of which authors also used the interaction of variables, offers comparable results. Advantage of these methods is that LASSO is more comfortable to handle than the compared machine learning model.

Acknowledgments

This work was supported by the Slovak Research and Development Agency under Contract No. APVV-15-0489 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] V. V. Barek, P. Halaj and D. Igaz, “The Influence of Climate Change on Water Demands for Irrigation of Special Plants and Vegetables in Slovakia” *Bioclimatology and Natural Hazards*, pp. 271-282, 2009. ISBN: 978-1-4020-8875-9

[2] L. Jurik, K. Halaszova, J. Pokryvkova and S. Rehak, “Irrigation of arable land in Slovakia: history and perspective.”*Water resources in Slovakia.* Part 1. Cham: Springer, pp. 81-96. 2018. ISBN 978-3-319-92852-4

[3] 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.

[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

[5] V. Hrdina et al., “Landscape-ecological plan, Regional territorial plan.” Bratislava Region. Aurex, 2010.

[6] T. Hastie, R. Tibshirani and R. J. Tibshirani, “Extended comparisons of best subset selection, forward stepwise selection, and the lasso.” *arXiv:1707.08692*. 2017

[7] A. V. Dorogush, V. Ershov and A. Gulin, “CatBoost: gradient boosting with categorical features support” *arXiv:1810.11363v1*. 2018

[8] R. Caruana and A. Niculescu-Mizil. “An empirical comparison of supervised learning algorithms.”*In Proceedings ofthe 23rd international conference on Machine learning, pages 161–168. ACM,* 2006.

[9] G. Huang, L. Wu, X. Ma, W. Zhang, J. Fan, X. Yu, W. Zengand H. Zhou, “Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions”*Journal of Hydrology*, vol. 574, pp. 1029-1041. 2019

[10] R. Z. Safarov, Z. K. Shomanova, Y. G. Nossenko, Z. G. Berdenov, Z. B. Bexeıtova, A. S. Shomanov and M. Mansurova, “Solvıng of classıfıcatıon problem ın spatıal analysıs applyıng the technology of gradıent boostıng catboost”*Folia Geographica*, vol. 62(1), pp. 112-126. 2020

[11] J. K. Friedman “Stochastic Gradient Boosting” 1999 [Online] Available at: https://statweb.stanford.edu/~jhf/ftp/stobst.pdf

[12] CatBoost “Overview of CatBoost” 2020 [Online] Available at: https://catboost.ai/docs/

[13] J. Friedman,T. Hastie andR. Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software*, vol. 33(1), pp. 1–22. 2010