ISSN 1807-1929
Revista Brasileira de Engenharia Agrícola e Ambiental
Brazilian Journal of Agricultural and Environmental Engineering
v.27, n.3, p.202-208, 2023
Campina Grande, PB – http://www.agriambi.com.br – http://www.scielo.br/rbeaa

DOI: http://dx.doi.org/10.1590/1807-1929/agriambi.v27n3p202-208

Pedotransfer functions for estimating the van Genuchten model parameters in the Cerrado biome[1]

Funções de pedotransferência para estimar parâmetros do modelo de van Genuchten no bioma Cerrado

Mariana F. Veloso[2*], Lineu N. Rodrigues[3], Elpídio I. Fernandes Filho[4], Carolina F. Veloso[5] & Bruna N. Rezende[6]

[1] Research developed at Universidade Federal de Viçosa, Departamento de Engenharia Agrícola, Viçosa, MG, Brazil
[2] Universidade Federal de Viçosa/Departamento de Engenharia Agrícola, Viçosa, MG, Brazil
[3] Empresa Brasileira de Pesquisa Agropecuária/Embrapa Cerrados, Planaltina, DF, Brazil
[4] Universidade Federal de Viçosa/Departamento de Solos, Viçosa, MG, Brazil
[5] Instituto Federal do Norte de Minas Gerais/Campus Montes Claros, Montes Claros, MG, Brazil
[6] Universidade de São Paulo/Departamento de Engenharia de Biossistemas/Campus Piracicaba, Piracicaba, SP, Brazil

HIGHLIGHTS:

Machine learning algorithms were superior to stepwise regression in estimating water content saturated and residual parameters.
The high variability of the fit parameters α and n produced a low precision of the PTFs developed for such parameters.
The variables sand, clay, microporosity, and microporosity were the most important variables for the development of PTFs.

ABSTRACT: The Cerrado biome has presented challenges in reconciling its agricultural expansion with water availability. In this sense, water resources planning and management are fundamental for the economic, social, and environmental development of the Cerrado biome, which has been hampered by the lack of data, especially those referring to irrigation strategies, such as, for example, the water retention curve. The water retention curve is essential to understand water dynamics in the soil; however, obtaining it can be laborious, opening an opportunity for Pedotransfer Functions (PTFs). The current study aimed to develop and evaluate PTFs to estimate the fit parameters of the van Genuchten model for the Cerrado biome. Multiple Linear Regression (MLR) and four machine learning (ML) algorithms were used to develop the PTFs. The ML algorithms were the Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR), and K Nearest Neighbors (KNN). Two combinations of soil data were evaluated, and the predictor variables used in each set were different. Using the RF and SVR models, the best estimates were obtained concerning the parameter θ_s (saturated water content). As for θ_r (residual water content), the models showed a moderate predictive capacity. For the other parameters, the models did not perform satisfactorily for α and n (fit parameters).

Key words: machine learning, multiple linear regression, irrigation

RESUMO: O bioma Cerrado tem apresentado desafios em conciliar sua expansão agrícola com a disponibilidade hídrica. Nesse sentido, o planejamento e o manejo de recursos hídricos são fundamentais para o desenvolvimento econômico, social e ambiental do bioma Cerrado, que tem sido prejudicado pela carência de dados, especialmente aqueles referentes às estratégias de irrigação, como, por exemplo, as curvas de retenção de água, essencial para compreender a dinâmica de água no solo. Contudo, como sua obtenção pode ser trabalhosa, seus parâmetros podem ser estimados indiretamente via Funções de Pedotransferência (FPTs). O objetivo do presente estudo foi desenvolver e avaliar FPTs para estimar parâmetros do modelo de van Genuchten, usado para descrever o processo de retenção de água, para o bioma Cerrado. Para o desenvolvimento das FPTs, foram usados os métodos de Regressão Linear Múltipla (RLM) e quatro algoritmos de aprendizado de máquina: Multivariate Adaptative Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR) e K Nearest Neightbors (KNN). Dados de solo de duas localidades foram utilizados, sendo que as variáveis preditoras selecionadas em cada conjunto foram diferentes. As melhores estimativas foram obtidas para o parâmetro θ_s (umidade de saturação), com destaque para os modelos RF e SVR. Já para a θ_r (umidade residual), os modelos apresentaram uma capacidade preditiva moderada. Para os demais parâmetros, os modelos não apresentaram um desempenho satisfatório para α e n (parâmetros de ajuste).

Palavras-chave: aprendizado de máquina, regressão linear múltipla, irrigação

• Ref. 263368 – Received 23 Apr, 2022
* Corresponding author - E-mail: mariana.f.veloso@ufv.br
• Accepted 05 Oct, 2022 • Published 19 Oct, 2022
Editors: Rener Luciano de Souza Ferraz & Carlos Alberto Vieira de Azevedo

Introduction

The Cerrado biome has presented challenges in reconciling its agricultural expansion with water availability, especially in regions already experiencing conflicts in water availability (Ferreira et al., 2021). Thus, integrated water resource planning in the Cerrado biome is necessary to establish strategies aiming to increase water use efficiency by different users.

A water retention curve (WRC) is a mathematical representation of the drying process of a particular soil. It is a nonlinear, empirical relationship between the suction (or tension) exerted by soil on the surrounding moisture and the soil water content (Campbell, 1974). The WRC shape depends on soil physic-hydric properties. It is considered fundamental information for understanding water dynamics in the soil, and water balance calculations require efficient irrigation management. This curve resembles an inverted smoothed S where the upper and lower bounds correspond to saturated water content and residual water content, respectively. Several mathematical models were developed as an attempt to adequately represent the general shape of the curve, such as the models by Campbell (1974), van Genuchten (1980), Hutson & Cass (1987), Durner (1994), Fredlund & Xing (1994), Kosugi (1994), Seki (2007), and others.

Parameters of the WRC are estimated from local information on the water content observed values for each tension applied soil water. However, when dealing with large areas, such as the Cerrado biome, direct estimation of the retention curve parameters is not feasible at an appropriate scale. Such determination requires time and laborious routines, making using Pedotransfer Functions (PTFs) appealing as an indirect way of obtaining retention curves. PTFs allow the estimation of WRC parameters from physical-hydraulic attributes that are easy to measure and at low costs, such as sand, silt, clay, organic matter, and bulk density, among others (Vereecken et al., 2010).

This study aimed to develop pedotransfer functions to estimate van Genuchten model parameters for the Cerrado biome using multiple linear regression models and machine learning algorithms.

Material and Methods

The data used in this study were obtained from the Research Group on Water Resources of Embrapa Cerrados and the Hybras dataset (Ottoni et al., 2018). Initially, soil samples with data from WRCs and sand, silt, and clay, bulk density, particle size, total porosity, macroporosity, and microporosity were selected. The georeferenced samples (Figure 1) within the territorial limit of the Cerrado biome (Brazil), with a buffer of up to 100 km, were selected, totaling 188 samples (Figure 1). As for the non-georeferenced samples, 384 samples were selected, of which 216 belong to the states of Goiás (GO), 103 to Tocantins (TO), and 65 to Mato Grosso do Sul (MS), totaling at first 572 WRCs selected for the Cerrado region.

The methods for determining soil properties were reported in the consulted databases. For the obtention of soil water content, a tension table was used for tensions between

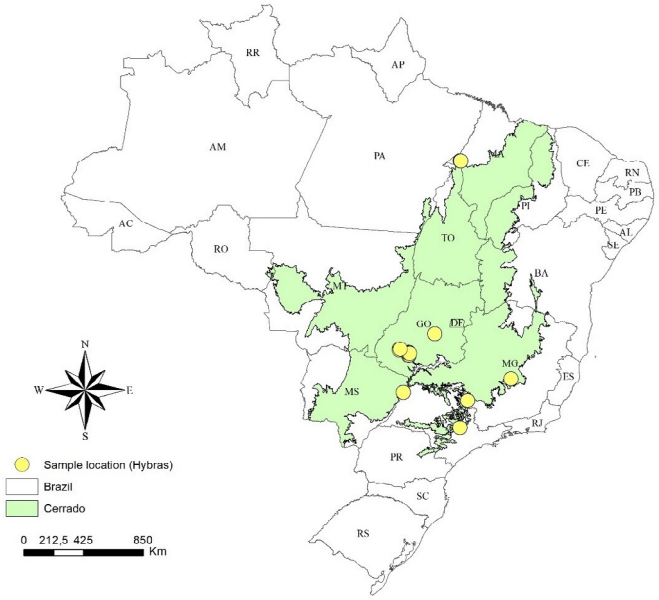


Figure 1. Location of georeferenced soil samples in the Cerrado biome

saturation and tension of 6 kPa, and a pressure chamber for higher tensions (EMBRAPA, 2017). For the determination of the granulometric contents, the pipette method was used. To obtain the bulk density, the volumetric ring method, and for the particle density, the volumetric flask and pycnometer methods (EMBRAPA, 2017). Total porosity (Pt) was based on bulk density (Bd) and particle density (Dp) (Pt = 1-Bd/Dp). The macroporosity was calculated by the difference between total porosity and microporosity (EMBRAPA, 2017).

The model used to represent the water retention curves was the van Genuchten (1980) equation. The model components, that is, the response variable ($\theta_{(\psi)}$), the predictor (ψ) and the set of model parameters are defined in Eq. 1.

$$\theta_{(\psi)} = \theta_r + \left[\frac{\theta_s - \theta_r}{1+(\alpha\psi)^n}\right] \quad (1)$$

where:

$\theta_{(\psi)}$ - soil water content, $m^3\ m^{-3}$;
θ_r - residual water content, $m^3\ m^{-3}$;
θ_s - saturated water content, $m^3\ m^{-3}$;
ψ - matrix potential, kPa;
α - scale factor, m;
n - shape factor, dimensionless; and,
m - a function of the shape factor (n), m = 1 – (1/n), dimensionless.

The SWRC Fit software (Seki, 2007) was used to obtain the estimated parameters of the van Genuchten (1980) model, namely, θ_s, θ_r, α, and n. It is worth mentioning that a minimum number of observed responses corresponding to each tension value (ψ, $\theta_{(\psi)}$) is necessary for model fitting. This number must be equal to or higher than the number of parameters. For instance, at least four points are required for those WRC fittings.

PTFs were developed only for the method that presented the overall best performance in estimating the WRC parameters for the Cerrado biome. The training subset was organized

considering two sets of predictors, namely, A1: sand, silt, clay, bulk density, particle density, total porosity, macroporosity, and microporosity; and A2: sand, silt, clay, bulk density, and particle density. Two different sets of predictors were used to visualize how the PTFs development models would behave.

In the development of the PTFs, five methods were evaluated: Multiple Linear Regression (MLR), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR), and K Nearest Neighbors (KNN), the last four algorithms being of machine learning. The PTFs were all developed in an R environment (R Core Team, 2019).

For the MLR, the Stepwise method was used to select the 'best' set of predictors. This method consists of adding more significant predictor variables or removing less significant ones during each model construction stage (Eq. 2) (Olubi, 2021). The PTF was developed using only the predictor set A1 in this case.

$$Y_i = \beta_{i,0} + \beta_{i,1} \cdot X_1 + ... + \beta_{i,n} \cdot X_n \quad (2)$$

where:

Y_i - variable to be estimated (parameters of the van Genuchten (1980) model: θ_s, θ_r, α, and n);

β_0 - intercept of multiple linear regression;

$\beta_1 ... \beta_n$ - angular coefficients linked to soil predictor variables; and,

$X_1 ... X_n$ - soil predictor variables (sets A1 and A2).

In addition, the Shapiro-Wilk test was used to verify the normality of the data, and those variables that showed a tendency to non-normality were transformed using the decimal logarithm function.

For the MARS model, the R package called earth was used. MARS is an algorithm that automatically models nonlinearity and interactions between variables where the training sets were divided into linear segments fitted into polynomial curves (splines) with different numbers of interactions and joined by knots (Hastie et al., 2009). For this, models with different numbers of interactions and nodes were developed, and the model with the lowest RMSE value was selected.

The RF is a model that combines regression trees, providing the average prediction of all trees (Liaw & Wiener, 2022). RF uses bootstrap sampling, i.e., it randomly draws a sample (with replacement), keeping the original size of the data in each tree. The R package randomForest (Liaw & Wiener, 2022) was used for this. A bootstrap sampling was performed for each generated tree, with the number of variables selected at each split of the tree controlled by the 'mtry hyperparameter', and the model with the lowest RMSE value was selected.

The SVR is an algorithm based on the hyperplane fit that separates the points in an n-dimensional space, where n is the number of predictor variables (Zhong et al., 2019). The R package e1071 (Meyer et al., 2019) was used for this. The model uses the kernel function to optimize the obtaining of the hyperparameters C (cost) and γ (gamma), responsible for the adjustment tolerance of the models. The radial kernel function was used for this study, and the model with the lowest RMSE value was selected.

Finally, KNN is a model that estimates the variable as a function of the average distance of its nearest neighbors in the data set, and for this, a measure of distance is used. In this case was adopted the Euclidean distance (Kohli et al., 2021). This means that models were developed with different numbers of nearest neighbors controlled by the hyperparameter k, and the one with the lowest RMSE value was selected. The R package kknn (Schliep & Hechenbichler, 2013) was used for this.

The repeated holdout method (Tanner et al., 2019) was used to validate the generated models. The database was divided into two independent subsets, namely, 'training' and 'test', consisting of 70 and 30% of the data, respectively.

The hyperparameters of each machine learning model were adjusted by the k-folds cross-validation method with repetitions (k = 10, n = 3). Finally, the performance of the tested methods was evaluated using the test set, allowing the evaluation of the generalization capacity of the PTFs.

In general, the predictive capacity of PTF is evaluated from indices that measure the errors between predicted and observed data (Nasta et al., 2021). To evaluate the performance of the PTFs developed for the WRC parameters, the following statistical indexes were used: the coefficient of determination (R^2), the mean error (ME), and the root mean squared error (RMSE) were used. These indexes are commonly used in the evaluation of PTFs (Nguyen et al., 2017; Nasta et al., 2021). The R^2 expresses the degree of agreement between the observed values and those predicted by the PTFs (Eq. 3); corresponds to the squared Pearson correlation (r), assuming values between 0 and 1. The ME (Eq. 4) is an overall measure that indicates if the model tends to overestimate (ME > 0) or underestimate (ME < 0), based on an average of model residuals the response variable, and the RMSE indicates the average absolute error magnitude without account to the sign (positive or negative) of the model residuals (Eq. 5).

$$R^2 = \frac{\sum\left(\hat{y}_j - \overline{y}_j\right)^2}{\sum\left(y_j - \overline{y}_j\right)^2} \quad (3)$$

$$ME = \frac{1}{n}\sum_{j=1}^{N} y_j - \hat{y}_j \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2} \quad (5)$$

where:

y_j and $\hat{y}_j$ - estimated and observed values of the response variable, respectively;

$\overline{y}_j$ - mean of the y_j values;

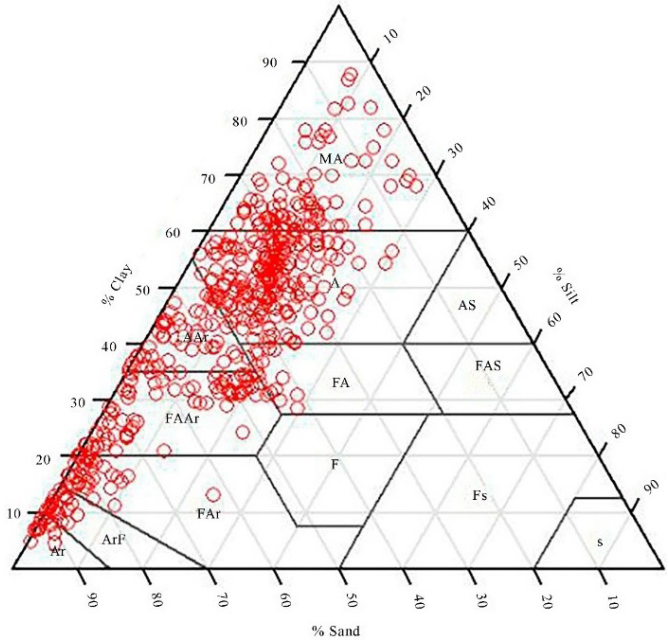
N - number of samples;

$\Sigma(\hat{y}_j - \overline{y}_j)^2$ - variance explained by the model; and,

$\Sigma(y_j - \overline{y}_j)^2$ - total variance.

Results and Discussion

The database used in this study presented textured soils with high levels of sand and clay, with most soils classified as clayey (A - clayey) (Figure 2). The soils of the Cerrado biome are mostly considered Oxisols, soils with high clay content and well structured, and may have high hydraulic conductivity. These characteristics confer one of the main differences between tropical soils and soils from temperate climates and, consequently, the reason for the inaccuracy of FPTs developed in these regions when applied to tropical soils.



MA - Very clayey; A - Clay; AS - Silty clay; AAr - Sandy clay; FA - Loamy loam; FAS - Silty clay loam; FAAr - Sandy clay loam; F - Loamy; FS - Silty loam; FAr - Sand loam; S - Silt; ArF - Loamy sand; Ar - Sand

Figure 2. Texture triangle of soil samples from the Cerrado biome used to develop the pedotransfer functions for estimating soil water retention curve parameters

Table 1 shows the descriptive statistics of the parameters of the van Genuchten (1980) equation. The parameter θ_s presented the lowest variability compared to the other parameters of the WRC, resulting in low values of both standard deviation and CV. Conversely, θ_r, α, and n presented the highest variability expressed by high values for both standard deviation and CV values. The CV of the n parameter also showed a large difference between the training (76.23%) and the test set (45.15%) in the test set. The α parameter presented the greatest variability, resulting in a large difference between its average values obtained from the training and test sets, around 280%.

The high variability of the α parameter is observed in several studies. Vereecken et al. (2010) and Gupta et al. (2022) emphasize this high variability of α as inherent to the parameter and state that retention curves with high values of α indicate soils with considerable sand contents, as verified in the database used in this study for the Cerrado biome (Table 1).

Regarding the results of the Shapiro-Wilk test, all variables showed a tendency toward non-normality in the data distribution, so these variables were submitted to transformation using the decimal logarithm function to construct the multiple linear regression. The normality trend was confirmed for all variables after the transformation.

The general ability of the methods for estimating the parameters of the WRC model was evaluated from the average value of the performance criteria (Table 2). The RF model showed the best performance for the θ_s and θr parameters. For α and n fit parameters, the R^2 values were approximately zero for all models and predictor sets, with the MLR as the best model.

ME values for θ_r were relatively low, varying between -0.009 and 0.004 $m^3\ m^{-3}$ for sets A1 and A2. As for θ_s, the MARS

Table 1. Descriptive statistics of the independent variables used to predict the parameters of the van Genuchten (1980) equation

Subset	Variable	Mean	Minimum	Maximum	SD	CV (%)
Training (N = 420)	Sand (%)	43.68	4.01	94.68	0.11	53.61
	Silt (%)	11.89	0.021	20.01	6.21	59.29
	Clay (%)	44.43	4.79	88.00	16.79	39.58
	Bulk density (g cm^{-3})	1.34	0.82	1.76	0.15	10.93
	Particle density (g cm^{-3})	2.67	2.31	2.98	0.13	4.51
	Total porosity (%)	47.71	37.64	59.56	4.47	9.37
	Microporosity (%)	38.33	16.26	50.43	4.85	12.65
	Macroporosity (%)	9.44	0.29	30.54	5.86	62.06
	θ_s ($m^3\ m^{-3}$)	0.49	0.39	0.57	0.06	1.09
	θ_r ($m^3\ m^{-3}$)	0.19	0.11	0.36	0.11	53.66
	Fit parameter α (m^{-1})	12.28	0.06	109.42	34.14	278.09
	Fit parameter n	1.65	0.12	15.52	0.74	76.23
Test (N = 178)	Sand (%)	46.10	4.00	93.02	20.18	42.84
	Silt (%)	10.78	0.021	27.0	6.21	59.29
	Clay (%)	43.11	4.79	82.76	16.79	39.58
	Bulk density (g cm^{-3})	1.35	0.82	1.61	0.15	10.93
	Particle density (g cm^{-3})	2.67	2.93	2.98	0.13	4.51
	Total porosity (%)	47.85	38.58	59.63	4.45	9.30
	Microporosity (%)	37.98	16.26	50.43	4.97	13.01
	Macroporosity (%)	9.97	0.29	30.54	6.47	64.91
	θ_s ($m^3\ m^{-3}$)	0.51	0.38	0.58	0.06	11.86
	θ_r ($m^3\ m^{-3}$)	0.19	0.11	0.38	0.09	50.99
	Fit parameter α (m^{-1})	3.08	0.081	73.23	9.01	292.06
	Fit parameter n	1.65	0.11	4.53	0.74	45.15

θ_s - Saturated soil water content; θ_r - Residual soil water content; SD - Standard deviation; CV - Coefficient of variation; N - Number of samples

Table 2. Mean values of the summary statistics used to evaluate the performance of the method tested for PTFs development to estimate the parameters of the van Genuchten (1980) model using the predictor sets A1 and A2

WRC parameter	Summary statistics	Methods				
		MLR	MARS	RF	SVR	KNN
A1 subset of predictors						
θ_s	R^2	0.62	0.65	0.76	0.75	0.74
	RMSE	0.029	0.034	0.029	0.029	0.029
	ME	-0.001	0.002	0.001	0.002	0.001
θ_r	R^2	0.21	0.21	0.48	0.29	0.38
	RMSE	0.097	0.075	0.049	0.869	0.078
	ME	0.002	-0.001	0.004	-0.001	-0.003
α	R^2	0.15	0.001	0.001	0.003	0.001
	RMSE	0.616	10.179	10.184	10.181	10.179
	ME	0.067	3.053	3.059	3.047	3.051
n	R^2	0.11	0.029	0.107	0.043	0.078
	RMSE	0.154	1.632	1.631	1.631	1.632
	ME	-0.005	1.441	1.445	1.428	1.439
A2 subset of predictors						
θ_s	R^2	0.62	0.68	0.71	0.72	0.64
	RMSE	0.029	0.032	0.031	0.029	0.034
	ME	-0.001	6.291	0.001	5.061	-0.001
θ_r	R^2	0.15	0.23	0.42	0.38	0.36
	RMSE	0.096	0.081	0.071	0.079	0.079
	ME	0.001	-0.001	0.004	-0.009	-0.001
α	R^2	-	0.001	0.02	0.001	0.001
	RMSE	-	10.179	0.998	10.177	10.179
	ME	-	0.961	0.969	0.754	1.035
n	R^2	-	0.023	0.09	0.034	0.032
	RMSE	-	1.645	1.642	1.641	1.639
	ME	-	1.443	1.448	1.431	1.442

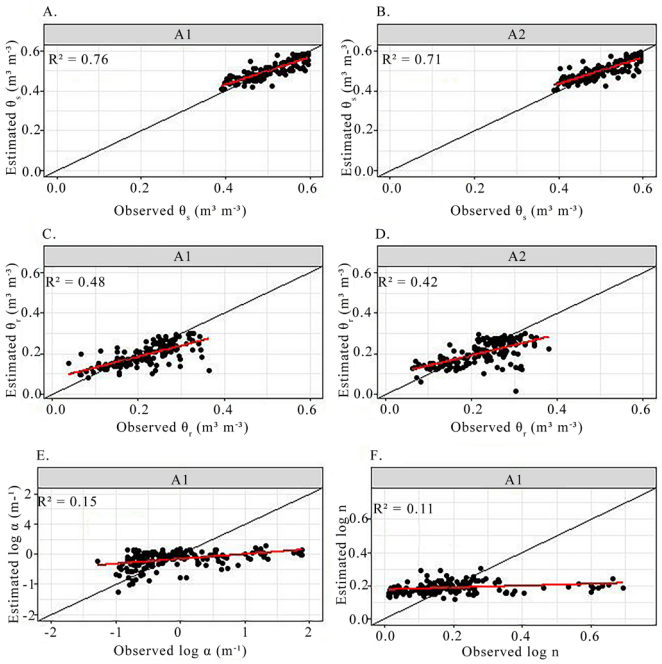
θs - Saturated water content; θr - Residual water content; α, n - Fitting parameters; MLR - Multiple Linear Regression; MARS - Multivariate Adaptive Splines; RF - Random Forest; SVR - Support Vector Regression; KNN - K Nearest Neighbors; A1 and A2 - Predictors sets A1 and A2, respectively; R^2 - Coefficient of determination; RMSE - Root Mean Square Error; ME - Mean Error; and WRC - Water Retention Curve

and SVR models were overestimated, with mean residuals around 5.0 and 6.0 $m^3 m^{-3}$. The ME value for the α parameter indicates that the model overestimates the unknown true values on average, with mean residuals around 3.0 m^{-1} for set A1 and around 1.0 m^{-1} for set A2. For n, the ME values also indicated overestimation, but the means residuals were lower, corresponding to the half part of ME for α (1.5 m^{-1}), except for the MLR.

Regarding the RMSE values, set A1 presented a variation close to zero for θ_s and θ_r, and higher values for α, with a mean equal to 10.18 m^{-1} for A1 and A2 predictor sets, except for the MLR. These high values of RMSE and ME, as well as the low values of R^2 for the parameters α and n, can be attributed to the high variability of the soil and, consequently, the model fitting difficulty, as mentioned by Vereecken et al. (2010).

It is observed that, in general, machine learning models present a better performance when compared to multiple linear regression. Araya & Ghezzehei (2019) highlighted the potential of machine learning algorithms due to the nonlinearity between the physical-hydraulic properties of the soil, allowing these more robust methods to perform better than the models considered more straightforward.

Figure 3 shows the estimates obtained by the best model for each parameter (θ_s, θ_r, α, n). Note that the fitting lines (red line) are far from the 1:1 line for parameters α and n, indicating that the adjustments were poor. As for the soil water content,



θs - Saturated water content; θr - Residual water content; α, n - Fitting parameters; A1 - Set predictor A1; A2 - Set predictor A2

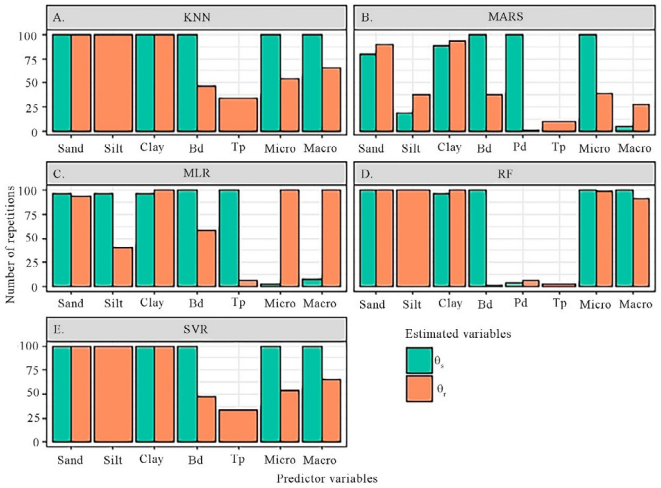
Figure 3. Estimated parameters of the van Genuchten (1980) equation obtained by the best-performing models (A, B, C, D) RF, and (E, F) MLR compared to the parameters of the van Genuchten (1980) equation estimated from the original data of tension and volumetric soil moisture in each selected location in the Cerrado biome

the behavior is more acceptable, emphasizing the saturated water content, which presented the best fit of all the evaluated parameters.

Barros et al. (2013) developed PTFs to estimate van Genuchten parameters in northeastern Brazil, finding R^2 values equal to 0.12 and 0.21 for parameters α and n, respectively, using soil texture data, bulk density, and organic matter. Other authors, such as Bai et al. (2022) and Baumann et al. (2022), also found difficulties in estimating the α and n parameters considering different types of soils and other regions of the world.

When analyzing the results of the A1 predictors set, PTFs performance improved for estimating θ_s and θ_r. Fatichi et al. (2020) highlight the importance of soil structural properties for assessing soil water content. The classification of the importance of the variables allows for visualizing the predictor variables that contributed the most in each round (repetition) in the development of the PTFs (Figure 4). The number of repetitions performed is found on the Y-axis, a hundred in this case, and the predictor variables on X-axis. It is observed that the structural variables, macroporosity, and microporosity, were crucial in all models evaluated (the bars are close to 100, indicating that they were considered in almost all repetitions). For example, analyzing the KNN model, Figure 4A, concerning the relevance of macroporosity, it is observed that for θ_s, the bar reaches the value of 100. On the other hand, the variable θ_r appears only in 60 repetitions.

Regarding the MARS model (Figure 4B), the independent variable total porosity is not very important as a predictor of



Bd - Bulk density; Pd - Particle density; Tp - Total porosity; Micro - Microporosity; Macro - Macroporosity; θs - Saturated water content; θr - Residual water content; MLR - Multiple Linear Regression; MARS - Multivariate Adaptive Splines; RF - Random Forest; SVR - Support Vector Regression; KNN - K Nearest Neighbors

Figure 4. Classification of the importance of predictor variables in the estimation of saturation (θ_s) and residual (θ_r) soil water content parameters using the A1 predictor set for the tested methods: (A) KNN, (B) MARS, (C) MLR, (D) RF, and (E) SVR

θ_s and θ_r, occurring in less than 25 of the 100 repetitions. For RF (Figure 4D), sand, clay, macroporosity, and microporosity were considered essential for θ_s and θ_r, occurring in the 100 repetitions. In MLR, Figure 4C, macroporosity was deemed important only for θ_r (less than 25 repetitions). Finally, the SVR, in Figure 4E, macroporosity occurred in all repetitions for θ_s and only 60 of the 100 repetitions for θ_r. It is worth mentioning that the macroporosity and microporosity variables are correlated so that the models can select one or the other to explain the variables θ_s and θ_r.

In addition to the macroporosity, microporosity sand and clay were selected in the PTFs development, considered necessary in all models evaluated. This behavior can be explained by the low variability of sand and clay contents in most of the soil samples that presented high levels of sand and clay in this study. The silt, in turn, because of its higher variability, was selected as a predictor in residual water content estimation by using the KNN, RF, MLR, or SVR models.

On the other hand, the total porosity was not selected for estimating the response variable θr due to its great correlation with this response. It is worth remembering that the total porosity was not used in the θ_s estimation; therefore, it is not considered in its analysis. A similar explanation can be given to the behavior of the Bd and Pd variables. It is observed that they had a higher occurrence in the saturated water content estimation than in the residual due to the density correlation in the θs calculation.

Although the MLR did not present the best performance among the evaluated models, it allows PTF to be represented as an equation and, consequently, can be directly applied, which differs from the machine learning algorithms used in this study. Thus, the PTFs obtained for the parameters θ_s, θ_r, α, and n using the predictor set A1 are presented in Table 3.

Table 3. Coefficients of pedotransfer functions obtained by stepwise multiple linear regression

Independent variable	Coefficient	θ_s	θ_r	α	n
Intercept	β_0	-0.5486392	-8.83933	-0.974911	0.74627149
log(Sand) (%)	β_1	0	2.7462026	0	0
log(Silt) (%)	β_2	0	0	0.3520109	-0.0031476
log(Clay) (%)	β_3	0	1.310042	0	0
log(Bd)	β_4	-0.7762026	0	0	0
log(Pd)	β_5	0.800532	0	0	0
log(Micro)	β_6	0	0	0	-0.33812542
log(Macro)	β_7	0	1.029454	0.5497387	0

θs - Saturated water content; θr - Residual water content; α, n - Fitting parameters; Bd - Bulk density; Pd - Particle density; Micro - microporosity; Macro - macroporosity; β_0 - intercept; β_1 ... β_7 - angular coefficients linked to soil predictor variables

Conclusions

1. Machine learning algorithms performed better when compared to multiple linear regression to estimate the parameters θ_s and θ_r. The variables sand and clay, as well as the incorporation of macroporosity and microporosity in the predictor, set A1, improved the performance of the machine learning algorithms in the estimation of saturation and residual water contents.

2. In general, the PTFs developed for the parameters α and n of the van Genuchten equation presented low performances in all models and predictor sets, tending to parameter overestimation. The models MLR were superior to machine learning algorithms to estimate α and n. For the parameter θ_r, the predictive capacity was moderate using machine learning algorithms and low using multiple linear regression. As for the parameter θ_s, the results were better using machine learning algorithms, and the equations obtained by multiple linear regression offer moderate predictive capacity.

Acknowledgements

The study was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Finance Code 001). To the Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) for providing the soil data used in the present study.

Literature Cited

Araya, S. N.; Ghezzehei, T. A. Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. Water Resources Research, v.55, p.5715-5737, 2019. https://doi.org/10.1029/2018WR024357

Bai, X.; Shao, M.; Jia, X.; Zhao, C. Prediction of the van Genuchten model soil hydraulic parameters for the 5-m soil profile in China's Loess Plateau. Catena, v.210, p.e105889, 2022. https://doi.org/10.1016/j.catena.2021.105889

Barros, A. H. C.; van Lier, Q. de J.; Maia, A. de H. N.; Scarpare, F. V. Pedotransfer functions to estimate water retention parameters of soils in northeastern Brazil. Revista Brasileira de Ciência do Solo, v.37, p.379-391, 2013. https://doi.org/10.1590/S0100-06832013000200009

Baumann, P.; Lee, J.; Behrens, T.; Biswas, A.; Six,J.; McLachlan, G.; Rossel, R. A. V. Modelling soil water retention and water-holding capacity with visible–near-infrared spectra and machine learning. European Journal of Soil Science, v.73, p.e13220, 2022. https://doi.org/10.1111/ejss.13220

Campbell, G. S. A simple method for determining unsaturated conductivity from moisture retention data. Soil Science, v.117, p.311-314, 1974. https://doi.org/10.1097/00010694-197406000-00001

Durner, W. Hydraulic conductivity estimation for soils with heterogeneous pore structure. Water Resources Research, v.32, p.211-223, 1994. https://doi.org/10.1029/93WR02676

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária. Manual de métodos de análise de solo. Brasília: EMBRAPA, 2017. 557p.

Fatichi, S.; Or, D.; Walko, R.; Vereecken, H.; Young, M. H.; Ghezzehei, T. A.; Hengl, T.; Kollet, S.; Agam, N.; Avissar, R. Soil structure is an important omission in Earth System Models. Nature Communications, v.11, p.1-11, 2020. https://doi.org/10.1038/s41467-020-14411-z

Ferreira, F. L. V.; Rodrigues, L. N.; Silva, D. D. da. Influence of changes in land use and land cover and rainfall on the streamflow regime of a watershed located in the transitioning region of the Brazilian Biomes Atlantic Forest and Cerrado. Environmental Monitoring and Assessment, v.193, p.1-16, 2021. https://doi.org/10.1007/s10661-020-08782-5

Fredlund, D. G.; Xing, A. Equations for the soilwater characteristic curve. Canadian Geotechnical Journal, v.31, p.521-532, 1994. https://doi.org/10.1139/t94-061

Gupta, S.; Papritz, A.; Lehmann, P.; Hengl, T., Bonetti, S.; Or, D. Global mapping of soil water characteristics parameters - fusing curated data with machine learning and environmental covariates. Remote Sensing, v.14, p.1-22, 2022. https://doi.org/10.3390/rs14081947

Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: Data mining, inference, and prediction. 2.ed. New York: Springer, 2009. 745p.

Hutson, J. L.; Cass, A. A retentivity function for use in soil-water simulation models. European Journal of Soil Science, v.38, p.105-113, 1987. https://doi.org/10.1111/j.1365-2389.1987.tb02128.x

Kohli, S.; Godwin, G. T.; Urolagin, S. Sales prediction using linear and KNN regression. Advances in Machine Learning and Computational Intelligence, v.1563, p.321-329, 2021. https://doi.org/10.1007/978-981-15-5243-4_29

Kosugi, K. Three-parameter lognormal distribution model for soil water retention. Water Resources Research, v.30, p.891-901, 1994. https://doi.org/10.1029/93WR02931

Liaw, A.; Wiener, M. Breiman and Cutler's random forests for classification and regression. R package version 4.7-1.1, 2022.

Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F.; Chang, C. C.; Lin, C. C. Package 'e1071'. R package version 1.7-11, 2022.

Nasta, P.; Szabó, B.; Romano, N. Evaluation of pedotransfer functions for predicting soil hydraulic properties: A voyage from regional to fiel scales across Europe. Journal of Hydrology: Regional Studies, v.37, p.1-20, 2021. https://doi.org/10.1016/j.ejrh.2021.100903

Nguyen, P. M.; Haghverdi, A.; De Pue, J.; Botula, Y.-D.; Le, K. V.; Waegeman, W.; Cornelis, W. M. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. Biosystems Engineering, v.153, p.12-27, 2017. https://doi.org/10.1016/j.biosystemseng.2016.10.013

Olubi, O. E.; Oniya, E. O.; Owolabi, T. O. Development of predictive model for radon-222 estimation inthe atmosphere using stepwise regression and grid search based-random forest regression. Journal of the Nigerian Society of Physical Sciences, v.3, p.132-139, 2021. https://doi.org/10.46481/jnsps.2021.177

Ottoni, M. V.; Ottoni Filho, T. B.; Shaap, M. G.; Lopes-Assad, M. L. R. C.; Rotunno Filho, O. C. Hydrophysical database for brazilian soils (HYBRAS) and pedotransfer functions for water retention. Vadose Zone Journal, v.18, p.1-17, 2018. https://doi.org/10.2136/vzj2017.05.0095

R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2019. Available on: < https://www.r-project.org/ >. Accessed on: Out. 2020.

Schliep, K.; Hechenbichler, K. Package 'kknn'. R package version 1.2-2, 2013.

Seki, K. SWRC Fit - A nonlinear fitting program with a water retention curve for soils having unimodal and bimodal pore structure. Hydrology and Earth System Sciences, v.4, p.407-437, 2007. https://doi.org/10.5194/hessd-4-407-2007

Tanner, E. M.; Bornehag, C. G.; Gennings, C. Repeated holdout validation for weighted quantile sum regression. MethodsX, v.6, p.2855-2860, 2019. https://doi.org/10.1016/j.mex.2019.11.008

van Genuchten, M. T. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Science Society of America Journal, v.44, p.892- 898, 1980. https://doi.org/10.2136/sssaj1980.03615995004400050002x

Vereecken, H.; Weynants, M.; Javaux, M.; Pachepsky, Y.; Schaap, M. G.; van Genuchten, M. T. Using pedotransfer functions to estimate the van Genuchten-Mualem soil hydraulic properties: A review. Vadose Zone Journal, v.9, p.1-26, 2010. https://doi.org/10.2136/vzj2010.0045

Zhong, H.; Wang, J.; Jia, H.; Mu, Y.; Lv, S. Vector field based support vector regression for building energy consumption prediction. Applied Energy, v.242, p.403-414, 2019. https://doi.org/10.1016/j.apenergy.2019.03.078